

---


Electronic Theses and Dissertations, 2020-

---

2022

## Atmospheric Retrieval: Bayesian Methods, Machine Learning, and Application to Exoplanets

Michael Himes  
*University of Central Florida*

 Part of the [Astrophysics and Astronomy Commons](#)

Find similar works at: <https://stars.library.ucf.edu/etd2020>

University of Central Florida Libraries <http://library.ucf.edu>

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations, 2020- by an authorized administrator of STARS. For more information, please contact [STARS@ucf.edu](mailto:STARS@ucf.edu).

---

### STARS Citation

Himes, Michael, "Atmospheric Retrieval: Bayesian Methods, Machine Learning, and Application to Exoplanets" (2022). *Electronic Theses and Dissertations, 2020-*. 1218.

<https://stars.library.ucf.edu/etd2020/1218>



ATMOSPHERIC RETRIEVAL: BAYESIAN METHODS, MACHINE LEARNING, AND  
APPLICATION TO EXOPLANETS

by

MICHAEL D. HIMES

B.S. Physics, University of Central Florida, 2016

A dissertation submitted in partial fulfilment of the requirements  
for the degree of Doctor of Philosophy in Physics, Planetary Sciences Track  
in the Department of Physics  
in the College of Sciences  
at the University of Central Florida  
Orlando, Florida

Summer Term  
2022

Major Professor: Joseph Harrington

© 2022 Michael D. Himes

## ABSTRACT

Atmospheric retrieval is the inverse modeling method where atmospheric properties are constrained based on measured spectra. Due to the low signal-to-noise ratios of exoplanet observations, exoplanetary retrieval codes pair a radiative transfer (RT) simulator with a Bayesian statistical framework in order to characterize the distribution of atmospheric parameters that could explain the observations (the posterior distribution). This requires on the order of  $10^6$  RT model evaluations, which requires hours to days of compute time depending on model complexity. In this work, I investigate atmospheric retrieval methods and apply them to observations of hot Jupiters. Chapter 2 presents a set of RT and retrieval tests to validate the Bayesian Atmospheric Radiative Transfer (BART) retrieval code and applies BART to the emission spectrum of HD 189733 b. Chapter 3 investigates the dayside atmosphere of WASP-12b and resolves a tension in the literature over its composition. Chapter 4 introduces a machine learning direct retrieval framework which spawns virtual machines, generates spectra, trains neural networks, and performs atmospheric retrievals using trained neural networks. Chapter 5 builds on this and presents a machine learning indirect retrieval method, where the retrieval is performed using a neural network surrogate model for RT within a Bayesian framework, and compares it with BART. Chapter 6 utilizes the neural network surrogate modeling approach for thermochemical equilibrium chemistry models and compares it with other equilibrium estimation methods. Appendices address retrieval errors induced by choice of wavenumber gridding for opacity-sampling RT schemes, neural network model selection, the effects of data set size on neural network training, and the accuracy of Bayesian frameworks used for atmospheric retrieval.

## ACKNOWLEDGMENTS

I first thank my advisor, Dr. Joseph Harrington, for his constant support throughout my time in his lab. I am not sure if I would have entered graduate school if you had not offered me a position, but I *am* sure that I would not be the scientist I am today without your guidance. Your conscientiousness will always serve as a model regardless of the direction my career takes.

I also wish to thank the members of my committee, Dr. Atılım Güneş Baydin, Dr. Theodora Karalidi, and Dr. Yanga Fernández. I have learned so much from you over the past years, and I very much appreciate the time and feedback you have offered me during my time in graduate school.

Thank you to the UCF Planetary Group, especially Amy (and Joseph), Anicia, Keanna, Leslie (even if you're not in planetary), and Mary. You all have made this arduous journey more enjoyable, and I look forward to when our paths will cross again in the future.

Last but certainly not least, I thank my family and friends that have supported me, not only during graduate school, but throughout my life. I could not have written this dissertation without all of you. Though I cannot thank you all by name, I would like to give special recognition to a few that have been particularly influential. Mom, Dad, Abuela, Abuelo, Grandma, thank you for always being there for me. Most (if not all) of my success is due to the examples you set for me and the qualities you've instilled in me. Chris, Jr, Anthony, Matthew, I am thankful that our bond has been more than just blood. I treasure the many great times we have had together, and I look forward to many more. Cameron and Daniel, though not through blood, you are family all the same. In the half-a-lifetime we have known each other, we have helped each other to grow so much, and for that I am grateful. Finally, Mary, my best friend, wife, and the love of my life, thank you for being my partner through these past 3+ years. Your unending support has made all the difference, and I am excited to see what the future holds for us.

# TABLE OF CONTENTS

LIST OF FIGURES . . . . .	xii
LIST OF TABLES . . . . .	xvi
CHAPTER 1: INTRODUCTION . . . . .	1
1.1 Atmospheric Retrieval . . . . .	3
1.2 Machine Learning . . . . .	5
1.2.1 Dissertation Overview . . . . .	7
1.3 List of References . . . . .	8
CHAPTER 2: AN OPEN-SOURCE BAYESIAN ATMOSPHERIC RADIATIVE TRANSFER (BART) CODE. I. DESIGN, TESTS, AND APPLICATION TO EXOPLANET HD 189733b . . . . .	14
2.1 Tests and BARTTEST . . . . .	15
2.1.1 Analytic RT Tests . . . . .	16
2.1.1.1 f04broadening: Line Broadening . . . . .	21
2.1.1.2 f05abundance: Varying Abundance . . . . .	24
2.1.1.3 f08isothermal: Isothermal Atmosphere . . . . .	26

2.1.2	Comparison RT Test . . . . .	26
2.1.2.1	Barstow et al. (2020) Forward Models . . . . .	29
2.1.3	Synthetic Retrieval Tests . . . . .	31
2.1.3.1	Barstow et al. (2020) Synthetic Retrievals . . . . .	33
2.1.4	Real-Data Retrieval Test . . . . .	42
2.2	Application to HD 189733 b . . . . .	43
2.3	List of References . . . . .	52
 CHAPTER 3: ON THE DAYSIDE ATMOSPHERE OF WASP-12b . . . . .		57
3.1	Abstract . . . . .	58
3.2	Introduction . . . . .	58
3.3	BART . . . . .	63
3.4	Model Configurations . . . . .	65
3.5	Results and Discussion . . . . .	68
3.6	Conclusions . . . . .	80
3.7	Acknowledgments . . . . .	82
3.8	List of References . . . . .	82

CHAPTER 4: FROM BIOHINTS TO CONFIRMED EVIDENCE OF LIFE: POSSIBLE METABOLISMS WITHIN EXTRATERRESTRIAL ENVIRONMENTAL SUB- STRATES . . . . .	87
4.1 Abstract . . . . .	88
4.2 Background . . . . .	89
4.3 Methods . . . . .	92
4.3.1 Tools, Compute and Software Environment . . . . .	92
4.3.2 Generation of Planetary Spectra via NASA Goddard’s Planetary Spectrum Generator . . . . .	94
4.3.2.1 System Parameters . . . . .	94
4.3.2.2 Planetary Parameters . . . . .	96
4.3.3 Data Set INARA DS1 . . . . .	98
4.3.4 Machine Learning . . . . .	99
4.3.4.1 Model evaluation & experiments . . . . .	100
4.3.4.2 Producing predictive distributions over abundances by using the Monte Carlo dropout approximation . . . . .	101
4.4 Results . . . . .	102
4.4.1 INARA Performance . . . . .	102
4.4.2 Links to Code Repositories . . . . .	109



4.5	Discussion . . . . .	110
4.6	Future Work . . . . .	112
4.7	Conclusions . . . . .	113
4.8	Acknowledgemnets . . . . .	114
4.9	List of References . . . . .	114
CHAPTER 5: ACCURATE MACHINE LEARNING ATMOSPHERIC RETRIEVAL VIA		
A NEURAL NETWORK SURROGATE MODEL FOR RADIATIVE TRANS-		
	FER . . . . .	117
5.1	Abstract . . . . .	118
5.2	Introduction . . . . .	119
5.3	Methods . . . . .	123
5.3.1	Model Training . . . . .	123
5.3.2	Retrieval . . . . .	126
5.3.3	Software . . . . .	127
5.3.3.1	MARGE . . . . .	128
5.3.3.2	HOMER . . . . .	129
5.4	Results and Discussion . . . . .	131
5.4.1	Limitations . . . . .	136

5.4.2	Compute Cost	138
5.5	Conclusions	139
5.6	Acknowledgements	141
5.7	List of References	142

CHAPTER 6: TOWARDS 3D RETRIEVAL OF EXOPLANET ATMOSPHERES: ASSESS-  
 ING THERMOCHEMICAL EQUILIBRIUM ESTIMATION METHODS . . . 151

6.1	Abstract	152
6.2	Introduction	152
6.3	Methods	154
6.3.1	Equilibrium via Gibbs Energy Minimization	155
6.3.2	Equilibrium via Analytical Formulae	155
6.3.3	Equilibrium via Interpolation	156
6.3.4	Equilibrium via NN-based Surrogate Model	157
6.3.5	Performance Assessment	159
6.4	Results & Discussion	159
6.5	Conclusions	179
6.6	Acknowledgements	180

6.7	List of References . . . . .	180
APPENDIX A: RETRIEVAL ERRORS DUE TO WAVENUMBER SAMPLING GRID MIS-		
	MATCH . . . . .	183
A.1	List of References . . . . .	190
APPENDIX B: DETERMINING MODEL ARCHITECTURE . . . . .		
		191
B.1	List of References . . . . .	197
APPENDIX C: DATA SET SIZE CONSIDERATIONS . . . . .		
		198
APPENDIX D: ASSESSING BAYESIAN FRAMEWORKS FOR EXOPLANET ATMO-		
	SPHERIC RETRIEVAL . . . . .	201
D.1	Introduction . . . . .	202
D.2	Methods . . . . .	203
	D.2.1 Software . . . . .	206
D.3	Results and Discussion . . . . .	206
	D.3.1 Eggbox Problem . . . . .	207
	D.3.2 Log Gamma Problems . . . . .	209
	D.3.3 HD 189733 b . . . . .	212
D.4	Conclusions . . . . .	217

D.5 Acknowledgements . . . . . 218

D.6 List of References . . . . . 218

## LIST OF FIGURES

2.1	Comparison of <code>transit</code> line shape to <code>miniRT</code> 's Faddeeva function . . . . .	22
2.2	Spectra produced by <code>transit</code> for varying abundances . . . . .	24
2.3	Comparison between <code>transit</code> and the corresponding theoretical Planck function . . . . .	27
2.4	$T(p)$ profiles used in synthetic tests . . . . .	28
2.5	Comparison of spectra between <code>transit</code> and Barstow et al. (2020) . . . . .	30
2.6	Best-fit spectra for synthetic retrieval tests . . . . .	34
2.7	Comparison between input and retrieved $T(p)$ profiles, normalized contribution functions, and best-fit molecular abundances for synthetic emission retrieval tests . . . . .	35
2.8	Comparison between input and retrieved $T(p)$ profiles, normalized contribution functions, and best-fit molecular abundances for synthetic transmission retrieval tests . . . . .	36
2.9	Summary of retrieval results for the Model 0 case of Barstow et al. (2020) . . . . .	39
2.10	Summary of retrieval results for the Model 1 case of Barstow et al. (2020) . . . . .	40
2.11	Retrieval results for HD189733 b . . . . .	46
2.12	Comparison of best-fit retrieved log abundances for HD 189733 b . . . . .	48

3.1	BART results for Model 12 . . . . .	75
3.2	BART results for Model 13 . . . . .	76
3.3	Comparison of the retrieved H <sub>2</sub> O marginalized posteriors for the models using the Line et al. (2014) ‘null’ data set . . . . .	78
4.1	Schematic overview of the interaction between biology, chemistry, and physics in determining climate . . . . .	90
4.2	Overview of INARA’s cloud implementation . . . . .	93
4.3	Structure of the instantiated virtual machines . . . . .	93
4.4	Evaluated ML architectures . . . . .	103
4.5	Available and used data in ML phases (left) and model architecture (right) . .	103
4.6	INARA prediction performance for the logarithm of the normalized abundance of H <sub>2</sub> O, CO <sub>2</sub> , O <sub>2</sub> , N <sub>2</sub> and CH <sub>4</sub> across 1000 test planets . . . . .	104
4.7	Detailed results of INARA in predicting one random planet’s values for H <sub>2</sub> O and CH <sub>4</sub> . . . . .	105
4.8	Detailed results of INARA in predicting one random planet’s values for H <sub>2</sub> O, CO <sub>2</sub> , O <sub>2</sub> , N <sub>2</sub> , and CH <sub>4</sub> . . . . .	106
4.9	Same as Figure 4.8, but zoomed in around each distribution. . . . .	107
4.10	Same as Figure 4.8, but for all 12 molecules considered in the atmospheric model. . . . .	108

5.1	Schematic diagram of our inverse modeling method . . . . .	122
5.2	Four comparisons of planetary emission spectra predicted by MARGE and calculated by BART . . . . .	133
5.3	Comparisons between HOMER and BART posteriors . . . . .	137
6.1	Performance comparison between the various models and TEA for H <sub>2</sub> O. . . .	166
6.2	As in Figure 6.1, but for CO. . . . .	167
6.3	As in Figure 6.1, but for CH <sub>4</sub> . . . . .	168
6.4	As in Figure 6.1, but for CO <sub>2</sub> . . . . .	169
6.5	As in Figure 6.1, but for HCN. . . . .	170
6.6	As in Figure 6.1, but for C <sub>2</sub> H <sub>2</sub> . . . . .	171
6.7	As in Figure 6.1, but for C <sub>2</sub> H <sub>4</sub> . . . . .	172
6.8	As in Figure 6.1, but for NH <sub>3</sub> . . . . .	173
6.9	As in Figure 6.1, but for N <sub>2</sub> . . . . .	174
6.10	As in Figure 6.1, but for H <sub>2</sub> . . . . .	175
6.11	As in Figure 6.1, but for H. . . . .	176
6.12	As in Figure 6.1, but for He. . . . .	177
6.13	As in Figure 6.1, but using the grid of Cubillos et al. (2019) to illustrate NN inaccuracies at phase space extrema. . . . .	178

A.1	Example of the effect of wavenumber griddings when band integrating spectra	186
A.2	Retrieved 1D marginalized posterior distributions for each molecule considered in the synthetic retrieval . . . . .	187
A.3	Comparison of retrieved posteriors for the three indicated Barstow et al. (2020) model 0 spectra at two different wavenumber grid resolutions . . . . .	188
A.4	Comparison of retrieved posteriors for HD 189733 b at two different wavenumber grid resolutions . . . . .	189
B.1	Example range test . . . . .	194
D.1	Comparison of the thermal profile 1D marginalized posteriors for HD 189733 b between the 7 Bayesian frameworks listed. . . . .	214
D.2	Comparison of the gas abundance 1D marginalized posteriors for HD 189733 b between the 7 Bayesian frameworks listed. . . . .	215



## LIST OF TABLES

2.1	Summary of Tests . . . . .	17
2.2	Test Line Lists and Fictitious Test Gases . . . . .	20
2.3	Varying Abundance Test . . . . .	25
2.4	BARTTEST Retrievals: Posterior Accuracy . . . . .	37
2.5	BARTTEST Retrievals, Barstow et al. (2020) Model 0 Cases: Credible Regions	41
2.6	BARTTEST Retrievals, Barstow et al. (2020) Model 1 Cases: Credible Regions	41
2.7	Comparison of Fitted Log Abundances for HD 189733 b . . . . .	47
3.1	Summary of Retrieval Models . . . . .	64
3.2	Summary of Retrieved Credible Region Uncertainties . . . . .	69
3.3	Retrieved Molecular Log Abundances . . . . .	70
4.1	Upper limits on random uniform distribution draw for each molecule . . . . .	98
4.2	Structure of the data vector and its contents . . . . .	98
4.3	Reported error for five retrieved molecules . . . . .	105
4.4	Comparison of atmospheric retrieval methods . . . . .	110
5.1	Forward Model Parameter Space . . . . .	124

5.2	Model Evaluation: High-resolution Spectra . . . . .	131
5.3	Model Evaluation: Band-integrated Spectra . . . . .	132
5.4	Retrieved Credible Regions . . . . .	135
5.5	Credible Region Accuracy . . . . .	135
5.6	Bhattacharyya Coefficients . . . . .	136
6.1	TEA Model Grid Parameters . . . . .	156
6.2	Model Performance Comparison . . . . .	163
6.3	Linear Interpolation Model Performance Comparison . . . . .	164
6.4	IDW Model Performance Comparison . . . . .	165
B.1	Model Grid Search, 20 Epochs . . . . .	195
C.1	Model Comparison . . . . .	200
D.1	Bayesian Frameworks Considered . . . . .	205
D.2	Performance Comparison: Eggbox Problem . . . . .	208
D.3	Performance Comparison: 2D Log Gamma Problem . . . . .	210
D.4	Performance Comparison: 10D Log Gamma Problem . . . . .	211
D.5	Performance Comparison: HD 189733 b . . . . .	216

## CHAPTER 1: INTRODUCTION

Over the past few decades, the number of confirmed exoplanets has ballooned to over 5000 (NASA Exoplanet Archive) thanks to planet-hunting instruments and missions like the High-Accuracy Radial velocity Planetary Searcher, Kepler, and the Transiting Exoplanet Survey Satellite (Mayor et al. 2003, Borucki et al. 2010, Ricker et al. 2015). Yet, the atmospheres of most of these exoplanets remain a mystery. The vast majority of exoplanets are unable to be resolved separately from the host star, preventing a direct measurement of their emission spectra. Around 900 exoplanets have been detected via the radial velocity (RV) method, where the presence of an exoplanet is determined from the Doppler shift it induces on the stellar emission lines as it orbits the star. However, RV does not offer meaningful constraints on the atmospheric properties of the exoplanet, as it is only sensitive to the planet's mass and period. While some exoplanets have been detected via other methods like microlensing and timing variations, over 75% of confirmed exoplanets transit their host star, enabling not only their discovery, but also the potential for characterizing their atmosphere. Even so, of the nearly 4000 transiting exoplanets, only a small fraction have been observed over enough wavelength bands to provide meaningful constraints on their atmospheres.

As an exoplanet transits in front of its host star with respect to the observer ("primary transit"), the measured flux from the host star is reduced based on the relative size of the planet to the star. Some of the starlight is filtered through planet's atmosphere, where it is absorbed by the molecules present. This causes variations in the measured reduction in flux with respect to wavelength, which translates to variations in the apparent radius of the planet with respect to wavelength. These

variations provide the data necessary to infer some of the atmospheric properties via the inverse modeling technique known as atmospheric retrieval (see review by Madhusudhan 2018). However, due to the starlight's path through the limb of the atmosphere, the atmosphere appears optically thick at all but the lowest pressures, which provides an incomplete view of how the temperature and molecular abundances vary with altitude (de Wit & Seager 2013).

A subset of these transiting exoplanets can be characterized in greater detail by current instrumentation. As a transiting planet orbits the star, some appear to pass behind the star from the perspective of the observer ("secondary eclipse"). This results in a minute reduction in the total measured flux of the system, which corresponds to the emission of the planet. Thus, by computing the difference between the flux out of and during eclipse, the planet's emission spectrum can be inferred.

More detailed techniques to retrieve exoplanet atmospheres have been demonstrated in the past decade. Two-dimensional maps of the limb or dayside of an exoplanet can be inferred based on transit or eclipse measurements during ingress and egress (e.g., Majeau et al. 2012). Additionally, by observing the planet throughout its entire orbit, the resulting phase curve data can be used to infer longitudinal variations in the atmospheric properties (e.g., de Wit et al. 2012). However, at present, this is only possible for the hottest and largest exoplanets, which emit more radiation than smaller, cooler objects.

## 1.1 Atmospheric Retrieval

The radiative transfer (RT) equation describes how radiation is scattered, absorbed, and emitted as it interacts with some medium, like a planet’s atmosphere (Goody & Yung 1995). Using the RT equation under certain assumptions, a given set of atmospheric conditions can be used to deterministically calculate the corresponding spectrum. In theory, an infinite resolution spectrum would correspond exactly to some underlying atmospheric model parameters. In practice, the finite resolution of instruments, especially the low spectral resolutions of current exoplanet observations, coupled with the spectrum’s uncertainties induce uncertainties in the retrieved model parameters, resulting in a distribution of possibilities for each parameter. Consequently, this degeneracy problem requires exoplanet retrieval algorithms to combine a RT code with a Bayesian statistics framework in order to characterize that *posterior distribution* (Madhusudhan 2018).

Exoplanet retrieval codes typically employ either a Markov chain Monte Carlo (MCMC; e.g., ter Braak 2006, ter Braak & Vrugt 2008, Vrugt & ter Braak 2011, Foreman-Mackey et al. 2013) or nested sampling (NS; e.g., Feroz & Hobson 2008, Feroz et al. 2009, Handley et al. 2015, Speagle 2020, Buchner 2021) algorithm for the Bayesian framework and usually require on the order of hours to days of compute time, depending on the complexity of the RT forward model. MCMC and NS approximate the posterior through different means. MCMC randomly explores the phase space, probabilistically accepting steps according to the calculated  $\chi^2$  between the model and the observed data; by repeating this for many iterations (typically on the order of  $10^6$  for exoplanet retrievals), the distribution of accepted steps converges to the posterior. On the other hand, NS seeds the space with  $N$  *live points* by making random draws from the prior and calculating the corresponding likelihoods (related to  $\chi^2$ , as in MCMC). Then, it iteratively draws a new point from the prior, calculates the likelihood, and, if it is greater than the lowest likelihood among the  $N$  live points, then the new point replaces that lowest-likelihood point. In this way, the set of

live points will slowly approach the maximum likelihood, at which point the problem has been solved. Then, the set of discarded points and their associated weights can be transformed into the posterior distribution. While NS typically requires around an order of magnitude fewer likelihood evaluations than MCMC, NS algorithms can introduce significant overhead for certain problems, such as those with multi-modal posteriors.

Despite many exoplanet-focused RT and retrieval codes in the literature (e.g., Madhusudhan & Seager 2009, 2010, Madhusudhan et al. 2011, Benneke & Seager 2012, Line et al. 2013, Barstow et al. 2017, Oreshenko et al. 2017, Wakeford et al. 2017, Al-Refaie et al. 2021), detailed validation of these codes is lacking. Barstow et al. (2020) introduced the first set of synthetic exoplanet retrieval comparisons between three retrieval codes. They demonstrated that different RT/retrieval codes could both produce and retrieve numerically similar spectra. While they publicly released the forward spectra produced by their codes, they do not offer their code, the inputs (e.g., detailed atmospheric models), nor the numerical retrieval results, which could be helpful to diagnose differences between codes. As future RT and retrieval codes are developed to include more complex physics, it is imperative that a standardized set of RT and retrieval tests exists in order to ensure they are performing the correct calculations, analogous to the tests of Held & Suarez (1994) for general circulation models. In Chapter 2, I present such a test suite and apply it to the Bayesian Atmospheric Radiative Transfer (BART) retrieval code (Harrington et al. 2022, Cubillos et al. 2022, Blečić et al. 2022).

## 1.2 Machine Learning

Machine learning (ML) describes a class of computational algorithms that determines relationships within a given data set. Within ML, the subfield known as deep learning (Goodfellow et al. 2016) utilizes *neural networks*, a type of ML model that consists of a hierarchy of layers. Each layer, which consists of *nodes*, performs a weighted transformation of its inputs. Except for the final output layer, the output of one layer serves as input for the next layer. Through this sequence of transformations, the neural network is able to approximate the task at hand.

Neural networks begin with no knowledge of the task at hand; they must be trained on a data set to learn the task, that is, the weights between each connected node must be tuned in order to solve the problem. Training a neural network on a data set with inputs that correspond to some outputs (*supervised learning*) involves an iterative approach, where the network predicts outputs for some inputs, the predicted outputs are compared to the known true outputs via some error metric (the *loss*), and the network's weights are adjusted based on the magnitude of that error (*backpropagation*; Rumelhart et al. 1986). Through these adjustments, the neural network becomes progressively better at its task, until further training reduces the model's generality (overfitting the training data).

Common applications of neural networks include image classification (e.g., Krizhevsky et al. 2012, Simonyan & Zisserman 2015, Szegedy et al. 2015, He et al. 2016, Huang et al. 2017), speech recognition (e.g., Chorowski et al. 2014, Amodei et al. 2016, Chan et al. 2016, Xiong et al. 2016), and language translation (e.g., Cho et al. 2014, Bahdanau et al. 2015, Ranzato et al. 2016, Sennrich et al. 2016, Wu et al. 2016). More relevant to exoplanet retrievals, neural networks can perform complex regression tasks, such as emulating a code that simulates some complex process. After an upfront compute cost to train the neural network, the resulting emulator offers a significant reduction in compute time compared to the simulator code (Cranmer et al. 2019, Munk et al. 2019, Kasim et al. 2021).

Recently, a number of groups have applied ML methods to the problem of atmospheric retrieval. Márquez-Neila et al. (2018) used a random forest of regression trees to retrieve on WASP-12b's transmission spectrum and found results generally consistent with traditional Bayesian approaches in the literature. Zingales & Waldmann (2018) presented a deep convolutional generative adversarial network that could retrieve atmospheric parameters via inpainting (filling in missing values); they also found results consistent with a traditional Bayesian retrieval. Waldmann & Griffith (2019) used a convolutional neural network to map Saturn's features. We introduced `plan-net` in Cobb et al. (2019), which used an ensemble of neural networks to retrieve parameter distributions. However, while these and other applications can achieve results similar to traditional Bayesian retrievals, their output distributions do not exactly match and can significantly differ. They sacrifice the accuracy of a Bayesian framework for the speed of ML. Hayes et al. (2020) stands out from these other approaches, as they use a  $k$ -means clustering to inform the phase space for the Bayesian inference, though their approach does not offer as drastic of a reduction in compute cost as other ML approaches. In Chapter 5, I present a new approach to ML retrieval which utilizes a neural network as a surrogate model for RT, thereby circumventing the challenges associated with ML retrieval while achieving closer agreement with traditional Bayesian methods.



### *1.2.1 Dissertation Overview*

In this work, I validated the open-source BART retrieval code (Harrington et al. 2022, Cubillos et al. 2022, Blečić et al. 2022), analyzed the emission spectra of hot Jupiters using BART, investigated machine learning applications to reduce the compute time of retrievals, and compared the efficiency of Bayesian posterior sampling algorithms in the context of atmospheric retrieval. In Chapter 2, I introduce a suite of tests for RT and retrieval codes called BARTTEST, use it to validate BART, and apply BART to the emission spectrum of HD 189733 b. In Chapter 3, I investigate a discrepancy in the literature over the atmospheric composition of WASP-12b and present a detailed analysis of its emission spectra. In Chapter 4, I present an analysis pipeline to generate terrestrial exoplanetary spectra and train neural networks with dropout to directly retrieve their atmospheric properties. In Chapter 5, I present a neural network surrogate modeling approach to atmospheric retrieval which maintains the Bayesian framework and compare its performance and accuracy to the classical approach. Chapter 6 uses the aforementioned surrogate modeling approach to approximate thermochemical equilibrium of hot Jupiter atmospheres and compares its performance and accuracy to other equilibrium estimation methods. Appendices address errors in synthetic retrievals (Appendix A), performing a model grid search (Appendix B), data set size considerations when training neural network surrogate models (Appendix C), and the accuracy of Bayesian frameworks used for atmospheric retrieval (Appendix D).

I wish to make a note that all of the work presented here is collaborative in nature, as is typical across the hard sciences. While the first page of each chapter provides a list of coauthors who contributed to that work, Chapter 2 and Appendix A deserve additional clarification. That chapter and appendix contain the text that I led in the associated publication. This work's Section 2.1 and Section 2.2 correspond to Sections 6 and 7 from Harrington et al. (2022), while this work's Appendix A corresponds to Appendix D from Harrington et al. (2022). They are reproduced exactly and hence do not follow the typical paper format found in the other chapters of this work.

### 1.3 List of References

Al-Refaie, A. F., Changeat, Q., Waldmann, I. P., & Tinetti, G. 2021, *ApJ*, 917, 37

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., & Zhu, Z. 2016, in *Proceedings of Machine Learning Research*, Vol. 48, *Proceedings of The 33rd International Conference on Machine Learning*, ed. M. F. Balcan & K. Q. Weinberger (New York, New York, USA: PMLR), 173–182

Bahdanau, D., Cho, K., & Bengio, Y. 2015, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, ed. Y. Bengio & Y. LeCun

- Barstow, J. K., Aigrain, S., Irwin, P. G. J., & Sing, D. K. 2017, *ApJ*, 834, 50
- Barstow, J. K., Changeat, Q., Garland, R., Line, M. R., Rocchetto, M., & Waldmann, I. P. 2020, *MNRAS*, 493, 4884
- Benneke, B. & Seager, S. 2012, *ApJ*, 753, 100
- Blecic, J., Harrington, J., Cubillos, P. E., Bowman, M. O., Rojo, P. M., Stemm, M., Challener, R. C., Himes, M. D., Foster, A. J., Dobbs-Dixon, I., Foster, A. S. D., Lust, N. B., Blumenthal, S. D., Bruce, D., & Loredó, T. J. 2022, *The Planetary Science Journal*, 3, 82
- Borucki, W. J., Koch, D., Basri, G., Batalha, N., Brown, T., Caldwell, D., Caldwell, J., Christensen-Dalsgaard, J., Cochran, W. D., DeVore, E., Dunham, E. W., Dupree, A. K., Gautier, T. N., Geary, J. C., Gilliland, R., Gould, A., Howell, S. B., Jenkins, J. M., Kondo, Y., Latham, D. W., Marcy, G. W., Meibom, S., Kjeldsen, H., Lissauer, J. J., Monet, D. G., Morrison, D., Sasselov, D., Tarter, J., Boss, A., Brownlee, D., Owen, T., Buzasi, D., Charbonneau, D., Doyle, L., Fortney, J., Ford, E. B., Holman, M. J., Seager, S., Steffen, J. H., Welsh, W. F., Rowe, J., Anderson, H., Buchhave, L., Ciardi, D., Walkowicz, L., Sherry, W., Horch, E., Isaacson, H., Everett, M. E., Fischer, D., Torres, G., Johnson, J. A., Endl, M., MacQueen, P., Bryson, S. T., Dotson, J., Haas, M., Kolodziejczak, J., Van Cleve, J., Chandrasekaran, H., Twicken, J. D., Quintana, E. V., Clarke, B. D., Allen, C., Li, J., Wu, H., Tenenbaum, P., Verner, E., Bruhweiler, F., Barnes, J., & Prsa, A. 2010, *Science*, 327, 977
- Buchner, J. 2021, *The Journal of Open Source Software*, 6, 3001
- Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. 2016, in *ICASSP*
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. 2014, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar: Association for Computational Linguistics), 1724–1734

- Chorowski, J., Bahdanau, D., Cho, K., & Bengio, Y. 2014, in NIPS 2014 Workshop on Deep Learning, December 2014
- Cobb, A. D., Himes, M. D., Soboczenski, F., Zorzan, S., O'Beirne, M. D., Güneş Baydin, A., Gal, Y., Domagal-Goldman, S. D., Arney, G. N., Angerhausen, D., & 2018 NASA FDL Astrobiology Team, I. 2019, *AJ*, 158, 33
- Cranmer, K., Brehmer, J., & Louppe, G. 2019, arXiv preprint arXiv:1911.01429
- Cubillos, P. E., Harrington, J., Blečić, J., Himes, M. D., Rojo, P. M., Loredó, T. J., Lust, N. B., Challener, R. C., Foster, A. J., Stemm, M. M., Foster, A. S. D., & Blumenthal, S. D. 2022, *The Planetary Science Journal*, 3, 81
- de Wit, J., Gillon, M., Demory, B. O., & Seager, S. 2012, *A&A*, 548, A128
- de Wit, J. & Seager, S. 2013, *Science*, 342, 1473
- Feroz, F. & Hobson, M. P. 2008, *MNRAS*, 384, 449
- Feroz, F., Hobson, M. P., & Bridges, M. 2009, *MNRAS*, 398, 1601
- Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, *PASP*, 125, 306
- Goodfellow, I., Bengio, Y., & Courville, A. 2016, *Deep Learning* (MIT Press)
- Goody, R. & Yung, Y. 1995, *Atmospheric Radiation: Theoretical Basis* (OUP USA)
- Handley, W. J., Hobson, M. P., & Lasenby, A. N. 2015, *MNRAS*, 450, L61
- Harrington, J., Himes, M. D., Cubillos, P. E., Blečić, J., Rojo, P. M., Challener, R. C., Lust, N. B., Bowman, M. O., Blumenthal, S. D., Dobbs-Dixon, I., Foster, A. S. D., Foster, A. J., Green, M. R., Loredó, T. J., McIntyre, K. J., Stemm, M. M., & Wright, D. C. 2022, *The Planetary Science Journal*, 3, 80

- Hayes, J. J. C., Kerins, E., Awiphan, S., McDonald, I., Morgan, J. S., Chuanraksasat, P., Komonjinda, S., Sanguansak, N., Kittara, P., & SPEARNet Collaboration. 2020, *MNRAS*, 494, 4492
- He, K., Zhang, X., Ren, S., & Sun, J. 2016, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778
- Held, I. M. & Suarez, M. J. 1994, *Bulletin of the American Meteorological Society*, 75, 1825
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. 2017, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269
- Kasim, M. F., Watson-Parris, D., Deaconu, L., Oliver, S., Hatfield, P., Froula, D. H., Gregori, G., Jarvis, M., Khatiwala, S., Korenaga, J., Topp-Mugglestone, J., Viezzer, E., & Vinko, S. M. 2021, *Machine Learning: Science and Technology*, 3, 015013
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, in *Advances in neural information processing systems*, 1097–1105
- Line, M. R., Wolf, A. S., Zhang, X., Knutson, H., Kammer, J. A., Ellison, E., Deroo, P., Crisp, D., & Yung, Y. L. 2013, *ApJ*, 775, 137
- Madhusudhan, N. 2018, in *Handbook of Exoplanets*, ed. H. J. Deeg & J. A. Belmonte (Springer International Publishing AG), 104
- Madhusudhan, N., Harrington, J., Stevenson, K. B., Nymeyer, S., Campo, C. J., Wheatley, P. J., Deming, D., Blečić, J., Hardy, R. A., Lust, N. B., Anderson, D. R., Collier-Cameron, A., Britt, C. B. T., Bowman, W. C., Hebb, L., Hellier, C., Maxted, P. F. L., Pollacco, D., & West, R. G. 2011, *Nature*, 469, 64
- Madhusudhan, N. & Seager, S. 2009, *ApJ*, 707, 24
- . 2010, *ApJ*, 725, 261

- Majeau, C., Agol, E., & Cowan, N. B. 2012, *ApJ*, 747, L20
- Márquez-Neila, P., Fisher, C., Sznitman, R., & Heng, K. 2018, *Nature Astronomy*, 2, 719
- Mayor, M., Pepe, F., Queloz, D., Bouchy, F., Rupprecht, G., Lo Curto, G., Avila, G., Benz, W., Bertaux, J. L., Bonfils, X., Dall, T., Dekker, H., Delabre, B., Eckert, W., Fleury, M., Gilliotte, A., Gojak, D., Guzman, J. C., Kohler, D., Lizon, J. L., Longinotti, A., Lovis, C., Megevand, D., Pasquini, L., Reyes, J., Sivan, J. P., Sosnowska, D., Soto, R., Udry, S., van Kesteren, A., Weber, L., & Weilenmann, U. 2003, *The Messenger*, 114, 20
- Munk, A., cibior, A., Baydin, A. G., Stewart, A., Fernlund, G., Poursartip, A., & Wood, F. 2019, arXiv preprint arXiv:1910.11950
- Oreshenko, M., Lavie, B., Grimm, S. L., Tsai, S.-M., Malik, M., Demory, B.-O., Mordasini, C., Alibert, Y., Benz, W., Quanz, S. P., Trotta, R., & Heng, K. 2017, *ApJ*, 847, L3
- Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. 2016, in 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, ed. Y. Bengio & Y. LeCun
- Ricker, G. R., Winn, J. N., Vanderspek, R., Latham, D. W., Bakos, G. Á., Bean, J. L., Berta-Thompson, Z. K., Brown, T. M., Buchhave, L., Butler, N. R., Butler, R. P., Chaplin, W. J., Charbonneau, D., Christensen-Dalsgaard, J., Clampin, M., Deming, D., Doty, J., De Lee, N., Dressing, C., Dunham, E. W., Endl, M., Fressin, F., Ge, J., Henning, T., Holman, M. J., Howard, A. W., Ida, S., Jenkins, J. M., Jernigan, G., Johnson, J. A., Kaltenegger, L., Kawai, N., Kjeldsen, H., Laughlin, G., Levine, A. M., Lin, D., Lissauer, J. J., MacQueen, P., Marcy, G., McCullough, P. R., Morton, T. D., Narita, N., Paegert, M., Palle, E., Pepe, F., Pepper, J., Quirrenbach, A., Rinehart, S. A., Sasselov, D., Sato, B., Seager, S., Sozzetti, A., Stassun, K. G., Sullivan, P., Szentgyorgyi, A., Torres, G., Udry, S., & Villaseñor, J. 2015, *Journal of Astronomical Telescopes, Instruments, and Systems*, 1, 014003

- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, *Nature*, 323, 533
- Sennrich, R., Haddow, B., & Birch, A. 2016, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Berlin, Germany: Association for Computational Linguistics), 1715–1725
- Simonyan, K. & Zisserman, A. 2015, in *International Conference on Learning Representations*
- Speagle, J. S. 2020, *MNRAS*, 493, 3132
- Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. 2015, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9
- ter Braak, C. 2006, *Statistics and Computing*, 16, 239
- ter Braak, C. J. F. & Vrugt, J. A. 2008, *Statistics and Computing*, 18, 435
- Vrugt, J. A. & ter Braak, C. J. F. 2011, *Hydrology and Earth System Sciences*, 15, 3701
- Wakeford, H. R., Sing, D. K., Kataria, T., Deming, D., Nikolov, N., Lopez, E. D., Tremblin, P., Amundsen, D. S., Lewis, N. K., Mandell, A. M., Fortney, J. J., Knutson, H., Benneke, B., & Evans, T. M. 2017, *Science*, 356, 628
- Waldmann, I. P. & Griffith, C. A. 2019, *Nature Astronomy*, 3, 620
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. 2016, *arXiv preprint arXiv:1609.08144*
- Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., & Zweig, G. 2016, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP
- Zingales, T. & Waldmann, I. P. 2018, *AJ*, 156, 268

**CHAPTER 2: AN OPEN-SOURCE BAYESIAN ATMOSPHERIC  
RADIATIVE TRANSFER (BART) CODE. I. DESIGN, TESTS, AND  
APPLICATION TO EXOPLANET HD 189733b**

**Joseph Harrington<sup>1,2</sup>, Michael D. Himes<sup>1</sup>, Patricio E. Cubillos<sup>1,3</sup>, Jasmina Blecic<sup>1,4,5</sup>, Patricio M. Rojo<sup>6</sup>, Ryan C. Challener<sup>1,7</sup>, Nate B. Lust<sup>1,8</sup>, M. Oliver Bowman<sup>1</sup>, Sarah D. Blumenthal<sup>1,9</sup>, Ian Dobbs-Dixon<sup>4,5,10</sup>, Andrew S. D. Foster<sup>1,11</sup>, Austin J. Foster<sup>1</sup>, M. R. Green<sup>1</sup>, Thomas J. Loredó<sup>11</sup>, Kathleen J. McIntyre<sup>1</sup>, Madison M. Stemm<sup>1</sup>, David C. Wright<sup>1</sup>**

<sup>1</sup> *Planetary Sciences Group, Department of Physics, University of Central Florida, Orlando, FL 32816-2385, USA*

<sup>2</sup> *Florida Space Institute, University of Central Florida, Orlando, FL 32826-0650, USA*

<sup>3</sup> *Space Research Institute, Austrian Academy of Sciences, Graz, Austria*

<sup>4</sup> *Department of Physics, New York University Abu Dhabi, PO Box 129188, Abu Dhabi, UAE*

<sup>5</sup> *Center for Astro, Particle and Planetary Physics (CAP3), New York University Abu Dhabi, PO Box 129188, Abu Dhabi, UAE*

<sup>6</sup> *Departamento de Astronomia, Universidad de Chile, Camino El Observatorio, 1515 Las Condes, Santiago, Chile*

<sup>7</sup> *Department of Astronomy, University of Michigan, 1085 S. University Avenue, Ann Arbor, MI 48109, USA*

<sup>8</sup> *Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA*

<sup>9</sup> *Department of Physics, University of Oxford, Oxford OX1 3PU, UK*

<sup>10</sup> *Center for Space Science, NYUAD Institute, New York University Abu Dhabi, PO Box 129188, Abu Dhabi, UAE*

<sup>11</sup> *Center for Astrophysics and Planetary Science, Space Sciences Building, Cornell University, Ithaca, NY 14853-6801, USA*

Published in *The Planetary Science Journal* as part of  
Harrington, J., M. D. Himes, P. E. Cubillos, J. Blecic, P. Rojo, R. C. Challener, N. B. Lust, M. O.  
Bowman, S. D. Blumenthal, I. Dobbs-Dixon, A. S. D. Foster, A. J. Foster, M. R. Green, T. J.  
Loredó, K. J. McIntyre, M. M. Stemm, and D. C. Wright. 2022, PSJ, 3, 80

<https://doi.org/10.3847/PSJ/ac3513>



## 2.1 Tests and BARTTEST

At the level required for most retrievals, calculating RT is sufficiently complicated that one cannot verify the correctness of efficient codes by inspection. Codes calculating it are, therefore, subject to numerous types of errors (“bugs”), including subtle changes of values that produce plots that look correct, but are wrong. We have thus developed BARTTEST, an independent package of quantitative and qualitative tests for both RT and retrieval codes. This section presents an initial set of tests, using `transit` and BART as the first test subjects.

BARTTEST has four kinds of tests: analytic RT, comparison RT, synthetic retrieval, and real-data retrieval. The analytic and synthetic tests quantitatively assess correctness against known results. Several tests developed to catch file-reading and data-combining bugs also appear in the analytic group. The comparison and real-data retrieval tests compare complex, real-world calculations among multiple codes. The reliability of these tests depends on the strength of the consensus result. The analytic RT tests can be useful in diagnosing differences between results if a comparison test does not match the consensus.

Given the number of variables and setup parameters, this text focuses on general description and important points. The `transit` and BART configuration files for each test appear in the BARTTEST package as a reference for users to configure their own RT and retrieval codes to run these tests. The version of BARTTEST described here appears in the compendium. **These detailed configurations and their meanings in our codes are the official versions of these tests,** not the high-level descriptions in this paper. Others performing these tests should thus configure their codes to mimic the configurations, including the specific line lists, wavelengths computed, layer boundaries, thermal profile, included species, etc. For retrievals, it is important to use the same observations and uncertainties, even if better results appear in the literature in the future.

We encourage RT- and retrieval-code authors to validate their codes with BARTTEST’s analytic tests, to contribute their results to build the consensus for comparison tests, and to add new tests, especially to handle calculations not yet found in `transit` or BART, such as those involving hazes and clouds. For example, in Section 2.1.2.1, we compare `transit` to the hot-Jupiter cases of Barstow et al. (2020). Results and new tests may be submitted via pull requests at the code’s development repository on GitHub.

Table 2.1 summarizes the tests. Subsequent subsections expand on some of them and define terminology in the table.

### *2.1.1 Analytic RT Tests*

Our simple RT code, `miniRT`, performs BARTTEST’s analytic calculations for eclipse geometry (Equation (14)–(18) of BART2). Written in Python, its goal is verifiability by inspection, not efficiency. It follows how a human thinks about RT. Additional BARTTEST routines accept the output of the code being tested in human-readable form, compare these to the output from `miniRT`, and make comparison plots.

Many tests use a set of fictional gases with line lists constructed to facilitate the tests (see Table 2.2). Otherwise, these gases behave like the common molecules given in the table. To make mathematical confirmation straightforward, there is no continuum opacity in most tests, and there are just a few lines in each list.

Table 2.1: Summary of Tests

Name	Purpose	Atmospheric Composition	Notes
<b>Radiative-Transfer Analytic Tests</b>			
f01oneline	Location, width, shape, and strength of a single, known line.	One layer of LG1 at $\sim 40$ mbar, rest CG1 (see Table 2.2).	Calculating the line shape in transmission is nontrivial, so that case is not tested.
f02fewline	Combination of multiple, separate lines from one molecule.	Same as f01oneline, but with LG2.	
f03multiline	Combination of lines from multiple molecules and line lists.	Three layers each have a different gas with three lines (LG2, LG3, LG4). Others have a gas with no lines (CG1).	
f04broadening	Broadening line shape in opacity, isolating the broadening calculation from the radiation integral.	One layer of LG1, rest CG3.	This test produces an opacity table sampled every $0.005 \text{ cm}^{-1}$ over a short range, in addition to an intensity spectrum. All other tests produce only intensity spectra.
f05abundance	(Near) linear relationship between abundance and line depth at low optical depth.	Based on f01oneline, but abundance varies in 10 steps uniformly from $10^{-4}$ – $10^{-3}$ , with CG1 filling in.	LG1 trades off against CG1 to keep line broadening constant. Line is optically thin for near-linear absorption increase.
f06blending	Line blending from different molecules in the same layer.	The $\sim 9$ mbar layer has 1% LG1, 9% LG2, and 90% CG1; others have 85% CG1 and 15% CG2.	Two lines in LG1 and LG2 are $\sim 0.04 \text{ cm}^{-1}$ apart at $\sim 2.29 \mu\text{m}$ . The wavenumber sampling interval is $0.005 \text{ cm}^{-1}$ .
f07multicia	Multiple sources, similar to f03multiline.	CIA Uniform 85% $\text{H}_2$ , 15% He. $10^{-98}\%$ LG1, if code requires a line list. CIA line lists.	One of the following: No CIAs, $\text{H}_2$ -He CIAs, both $\text{H}_2$ - $\text{H}_2$ and $\text{H}_2$ -He CIAs (some codes may also require LG1).

f08isothermal	Background emission and emission-absorption cancellation for the isothermal case.	Uniform composition of 60% H <sub>2</sub> , 10% each CO, CO <sub>2</sub> , CH <sub>4</sub> , & H <sub>2</sub> O, using their full line lists, but no CIAs. Constant background and atmosphere temperatures.	Full line lists for all species. Result is a Planck spectrum. Uses just Sharp & Burrows (2007) gases, as some codes only have these, but users may configure many more. Not offered in transmission, as result is not a Planck spectrum.
---------------	---	---	--

---

**Radiative-Transfer Comparison Tests**

---

c01hjcclariso	Forward model of cloudless HD 189733 b-like planet.	Isothermal $T(p)$ profile in emission & transmission. Mean temperature $\sim 1100$ K. Full line lists for CH <sub>4</sub> , CO, CO <sub>2</sub> , H <sub>2</sub> O, NH <sub>3</sub> , and H <sub>2</sub> . H <sub>2</sub> -H <sub>2</sub> & H <sub>2</sub> -He CIAs.	Test of all code features on realistic cases. Validated by comparison to others.
c02hjcclarnoinv	Forward model of cloudless HD 189733 b-like planet.	Same as c01hjcclariso, except noninverted $T(p)$ profile in emission & transmission.	Same.
c03hjcclarinv	Forward model of cloudless HD 189733 b-like planet.	Same as c01hjcclariso, except inverted $T(p)$ profile in emission & transmission.	Same.
c04hjcclariso-BarstowEtal	Forward model of cloudless HD 189733 b-like planet.	Follows models of Barstow et al. (2020). Includes some CO-only models (1.0 $R_{\odot}$ , 1.0 $M_J$ , 1.0 $R_{J,\text{mean}}^*$ , 0.85 H <sub>2</sub> :0.15 He. 10 ppmv CO at 1500 K; 100 ppmv CO at 1000 K and 1500 K) and Model 0 (0.781 $R_{\odot}$ , 1.162 $M_J$ , 1.138 $R_{J,\text{mean}}^*$ , 0.85 H <sub>2</sub> :0.15 He, 1500 K, 300 ppmv H <sub>2</sub> O, 350 ppmv CO).	Same.

c05hjcloudiso- BarstowEtal	Forward model of cloudy HD 189733 b-like planet.	Follows Model 1 of Barstow et al. (2020) ( $0.781 R_{\odot}$ , $1.162 M_J$ , $1.138 R_{J,\text{mean}}^*$ , $0.85 \text{ H}_2:0.15 \text{ He}$ , $1500 \text{ K}$ , $300 \text{ ppmv H}_2\text{O}$ , $350 \text{ ppmv CO}$ , clouddeck at 10 mbar).	Same.
-------------------------------	---	--	-------

Name	Purpose	Data	Notes
<b>Retrieval Synthetic Tests</b>			
s01hjcleariso	Can we retrieve what we put in?	Model from c01hjcleariso.	
s02hjclearnoinv	Same.	Model from c02hjclearnoinv.	
s03hjclearinv	Same.	Model from c03hjclearinv.	
s04hjcleariso- BarstowEtal	Same.	Model 0 from c04hjclearisoBarstowEtal.	
s05hjcloudiso- BarstowEtal	Same.	Model 1 from c05hjcloudisoBarstowEtal.	
<b>Retrieval Real-Data Test</b>			
r01hd189733b	Reality check	Photometry: <i>Spitzer</i> IRAC chan- nels 1-4, IRS $16 \mu\text{m}$ , MIPS $24 \mu\text{m}$ . Spectra: <i>Spitzer</i> IRS, <i>HST</i> NIC- MOS G206 grism. <sup>1</sup>	Multiple reductions of these data exist. Tests must use the same eclipse depths and uncertainties as BARTTEST to be accurate com- parisons.

<sup>1</sup> IRAC is the InfraRed Array Camera. IRS is the InfraRed Spectrograph. MIPS is the Multi-band Imaging Photometer for Spitzer. *HST* is the *Hubble Space Telescope*. NICMOS is the Near Infrared Camera and Multi-Object Spectrograph.

\* Barstow et al. (2020) reports the planetary radii in terms of the volumetric mean radius of Jupiter ( $69,911 \text{ km}$ ), rather than the IAU-defined value of  $R_J$  ( $71,492 \text{ km}$ , Prša et al. 2016). For codes that use the IAU value, the radii must be converted accordingly.

Table 2.2: Test Line Lists and Fictitious Test Gases

Name <sup>1</sup>	Like <sup>2</sup>	# of lines	$\lambda$
CG1	H <sub>2</sub>	0	
CG2	He	0	
CG3	N <sub>2</sub>	0	
LG1	H <sub>2</sub> O	1	2.28919
LG2	CH <sub>4</sub>	3	2.28921, 2.15, 3.20
LG3	CO	3	2.38, 2.50, 2.54
LG4	CO <sub>2</sub>	3	2.86, 3.02, 3.78

<sup>1</sup> CG = Clear Gas, LG = Line Gas.

<sup>2</sup> Properties (mass, isotopes, etc.) same as this molecule, except line list.

All tests have an emission (eclipse geometry) version. BARTTEST includes transmission (transit geometry) cases only where it makes sense. For example, in `f01oneline`, the emission case is simply the product of the line strength, a Voigt curve, density, and layer thickness, which verifies by inspection. The transmission case entails calculating the slanted path length through the single layer of LG1 of rays at multiple altitudes, calculating the optical depth and transmission as above, multiplying by  $2\pi r$ , where  $r$  is layer distance from the planet center, and integrating over altitude. Further, the slant path may hit the layer twice or only partially. This loses the inspection-level simplicity of the emission case and combines these calculations with the Voigt function, making the test less diagnostic. Broadening is the same calculation for eclipse and transit geometries, so well-written codes will have one routine for it and will not require two tests. The transmission case for test `f05abundance` would require assessing a linear change in  $\tau = 1$  altitude, where  $\tau$  is optical depth, which has the issues outlined above and also requires many tightly spaced layers.

Next, we expand on selected tests.

### 2.1.1.1 *f04broadening: Line Broadening*

At least theoretically, spectral lines broaden into Voigt profiles,  $V$ . These are the convolution of Gaussian ( $G$ ) and Lorentzian ( $L$ ) functions. The Gaussian derives from Doppler broadening due to the Maxwell-Boltzmann distribution of velocities in a gas. The Lorentzian derives from Heisenberg uncertainty in the transition energy due to short state lifetimes, especially at high pressures and temperatures. As a function of wavenumber,  $\nu$ :

$$G(\nu; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\nu^2/2\sigma^2}, \quad (2.1)$$

$$L(\nu; \gamma) = \frac{\gamma}{\pi(\nu^2 + \gamma^2)}, \quad (2.2)$$

where  $\sigma$  and  $\gamma$  are the Gaussian width and Lorentzian half width, respectively (see BART2 for a more detailed description).

The Voigt profile can be constructed from the Faddeeva function:

$$w(z) = \frac{i}{\pi} \int_{-\infty}^{\infty} \frac{e^{-\eta^2}}{z - \eta} d\eta \quad (2.3)$$

$$z = \frac{\nu + i\gamma}{\sigma\sqrt{2}} \quad (2.4)$$

$$\eta = \frac{\nu'}{\sigma\sqrt{2}} \quad (2.5)$$

$$w(z) = \frac{i}{\pi} \int_{-\infty}^{\infty} \frac{e^{-\frac{\nu'^2}{2\sigma^2}}}{\left(\frac{\nu+i\gamma}{\sigma\sqrt{2}} - \frac{\nu'}{\sigma\sqrt{2}}\right) \sigma\sqrt{2}} d\nu' \quad (2.6)$$

$$= \frac{i}{\pi} \int_{-\infty}^{\infty} \frac{e^{-\frac{\nu'^2}{2\sigma^2}} (\nu - i\gamma - \nu')}{(\nu - \nu')^2 + \gamma^2} d\nu' \quad (2.7)$$

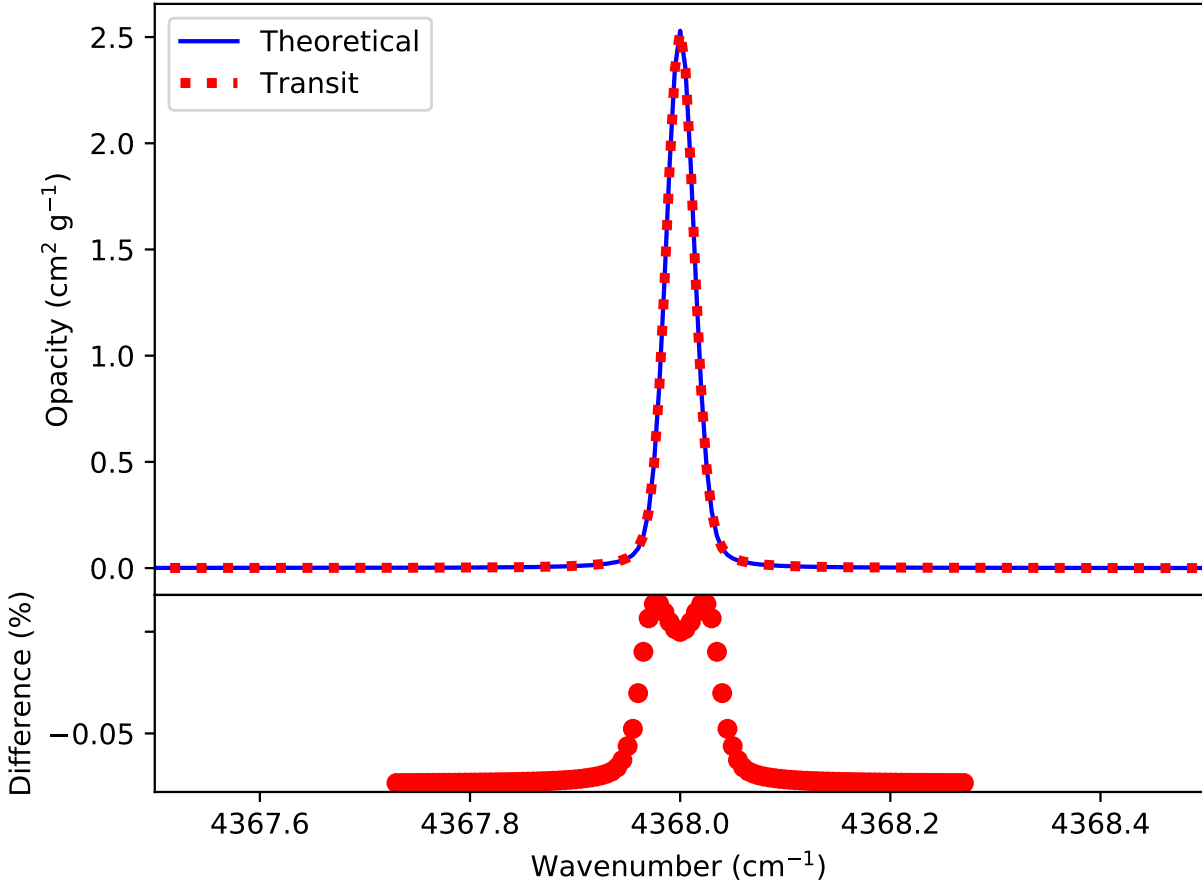


Figure 2.1: Comparison of `transit` line shape to `miniRT`'s Faddeeva function for that case (test `f04broadening`). Here and throughout, the axes (horizontal, in this case) may state a constant offset, for clarity.

Taking the real part of the Faddeeva function, one obtains the Voigt function:

$$\frac{\text{Re}[w(z)]}{\sigma\sqrt{2\pi}} = \frac{\gamma}{\sigma\sqrt{2\pi^3}} \int_{-\infty}^{\infty} \frac{e^{-\frac{\nu'^2}{2\sigma^2}}}{(\nu - \nu')^2 + \gamma^2} d\nu' \quad (2.8)$$

$$= \int_{-\infty}^{\infty} G(\nu'; \sigma) L(\nu - \nu'; \gamma) d\nu' \quad (2.9)$$

$$= V(\nu; \sigma, \gamma) \quad (2.10)$$



SciPy (Virtanen et al. 2020) offers a Python binding to a fast C implementation of  $w(z)$ , `scipy.special.wofz()`. BARTTEST uses this binding to compare to `transit`, which uses an approximation of this function, described by Pierluissi (1977).

Before building an opacity table, `transit` creates a pre-calculated table of Voigt profiles for a range of  $\sigma$  and  $\gamma$  values. When broadening each molecular line, `transit` uses the profile with the closest parameters. This approach avoids needing to calculate a Voigt profile for each line, which becomes computationally expensive for extensive line lists (e.g., ExoMol). For each molecule, `transit` computes the opacity table at the pressures and wavenumbers specified in the atmospheric configuration file and over a grid of temperatures.

When calculating an emission or transmission spectrum for a planet, it interpolates in the opacity table (if specified) to the temperature of each atmospheric layer. These approximations sharply reduce `transit`'s run time. The `f04broadening` test assesses their combined effect on accuracy. This test uses the default temperature sampling interval (100 K) and a grid of Voigt profiles over 60  $\sigma$  and 60  $\gamma$  values. The atmospheric layer containing the line-producing species has a temperature of 1442.58 K and a pressure of 0.33516 bar. With a difference of >42 K to the closest temperature in the opacity grid, this considers a case with (almost) the largest possible interpolation error. `Transit` differs from BARTTEST by <0.1% (Figure 2.1).

At extremely high spectral resolution or at long wavelengths, one must configure `transit`'s pre-calculated Voigt table to ensure sufficient accuracy, which can be assessed by running a modified version of this test (set the resolution/wavelength range, move the fake line to that location).

### 2.1.1.2 *f05abundance: Varying Abundance*

This tripwire test relies on the property that, in the optically thin regime (optical depth  $\tau \ll 1$ ), the fraction of the interior (blackbody) radiation absorbed,  $A$ , scales nearly linearly with the optical depth, since the linear term dominates in its Taylor-series expansion,

$$A = 1 - \frac{F}{F_0} = 1 - e^{-\tau} \approx 1 - \sum_{n=0}^{\infty} \frac{(-\tau)^n}{n!} \approx \tau - \frac{\tau^2}{2} + \frac{\tau^3}{6} - \dots, \quad (2.11)$$

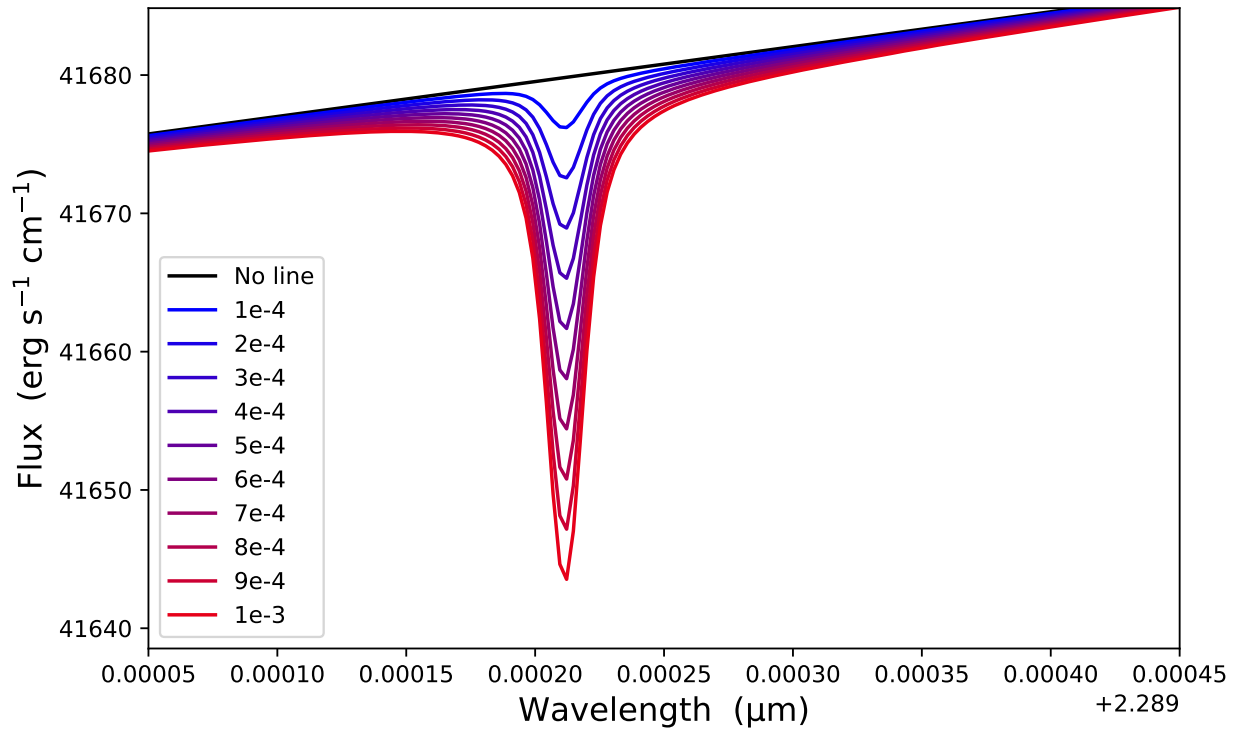


Figure 2.2: Spectra produced by `transit` for the eleven cases in the `f05abundance` test: the reference spectrum with no line, and the ten spectra with 0.01% to 0.1% LG1 in steps of 0.01%.

where  $F$  is transmitted flux and  $F_0$  is interior flux.  $\tau$  is linearly dependent on the density, which is proportional to the abundance. For two spectra of a gas with lower and higher abundance, each frequency channel should closely obey

$$f = \frac{F_0 - F_h}{F_0 - F_l}, \quad (2.12)$$

where  $F_0$  is the spectral flux without any lines,  $F_l$  is the flux in the low-abundance spectrum,  $F_h$  is the flux in the high-abundance spectrum, and  $f$  is the abundance ratio.

We test this using a non-inverted atmospheric model uniformly composed of 0.00% - 0.10% LG1, with the remainder CG1 (any CG species will do). The abundance of LG1 varies in steps of 0.01%. The 0.00 abundance case is  $F_0$ , the planetary interior blackbody without the LG1 spectral line. Taking  $F_l$  as the 0.01% abundance spectrum and starting  $F_h$  with the 0.02% case,  $f$  takes on the integers 2–10. Figure 2.2 shows the results of `transit`. Table 2.3 shows the output of `BARTTEST`.

Table 2.3: Varying Abundance Test

Abundance	$f^1$ vs. 0.01% case
0.02%	1.999784
0.03%	2.999535
0.04%	3.999005
0.05%	4.998531
0.06%	5.997734
0.07%	6.996982
0.08%	7.995986
0.09%	8.994910
0.1%	9.993651

<sup>1</sup> Factor difference in line depth

### 2.1.1.3 *f08isothermal: Isothermal Atmosphere*

This tripwire test recognizes that, without scattering, line emission and absorption are equal in any optically thick, isothermal gas mixture. The mixture thus emits as a blackbody. The atmosphere for `f08isothermal` has many molecules uniformly present in all layers and a full line list for all of the species present. This should produce a blackbody emission spectrum peaking at the wavenumber corresponding to the atmospheric model's temperature, according to Wien's law. Deviations may arise from approximations or precision issues in the code. Figure 2.3 shows the theoretical Planck function plotted over `transit`'s result.

### 2.1.2 *Comparison RT Test*

To avoid aggregating working parts into an erroneous whole, one must validate the entire RT calculation. As the complexity is too great to verify reliably by inspection, `c01hjcclariso`, `c02hjclearnoinv`, and `c03hjcclarinv` are tests inspired by the HD 189733 system. The strength of these tests rests on the number of participating codes, so we invite the community to perform these calculations in their own codes and to submit the results for inclusion in `BARTTEST`.

It is important to identify the sources of any differences without assuming that the tested codes are correct, as there is no assurance that several codes do not all share the same bug. With such honest testing, as the group of tested codes grows, so the likelihood of a groupthink bug decreases. The comparison of many codes implementing a single model also shows the range of outputs due to modeling approaches and assumptions, even with identical inputs.

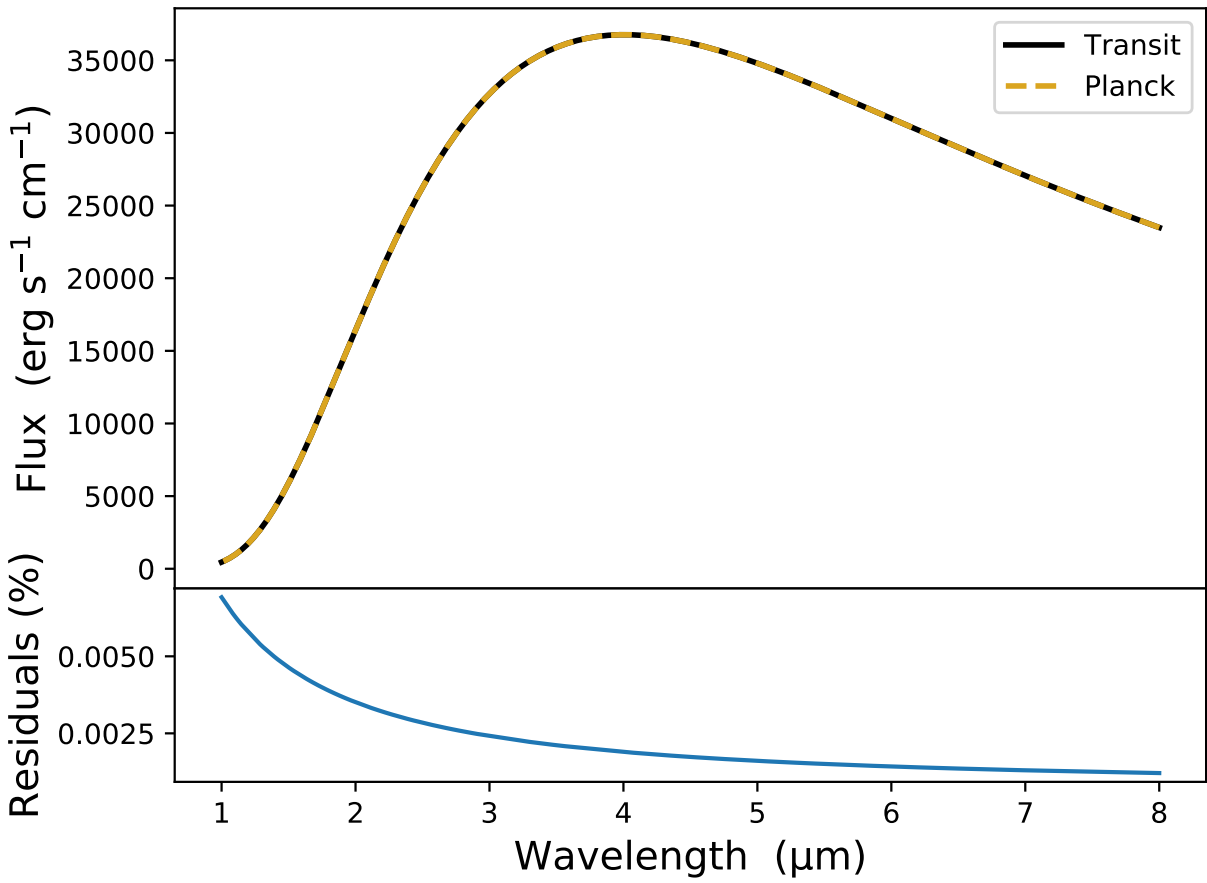


Figure 2.3: Transit's result for an isothermal atmosphere with theoretical Planck function overplotted (test f08isothermal).

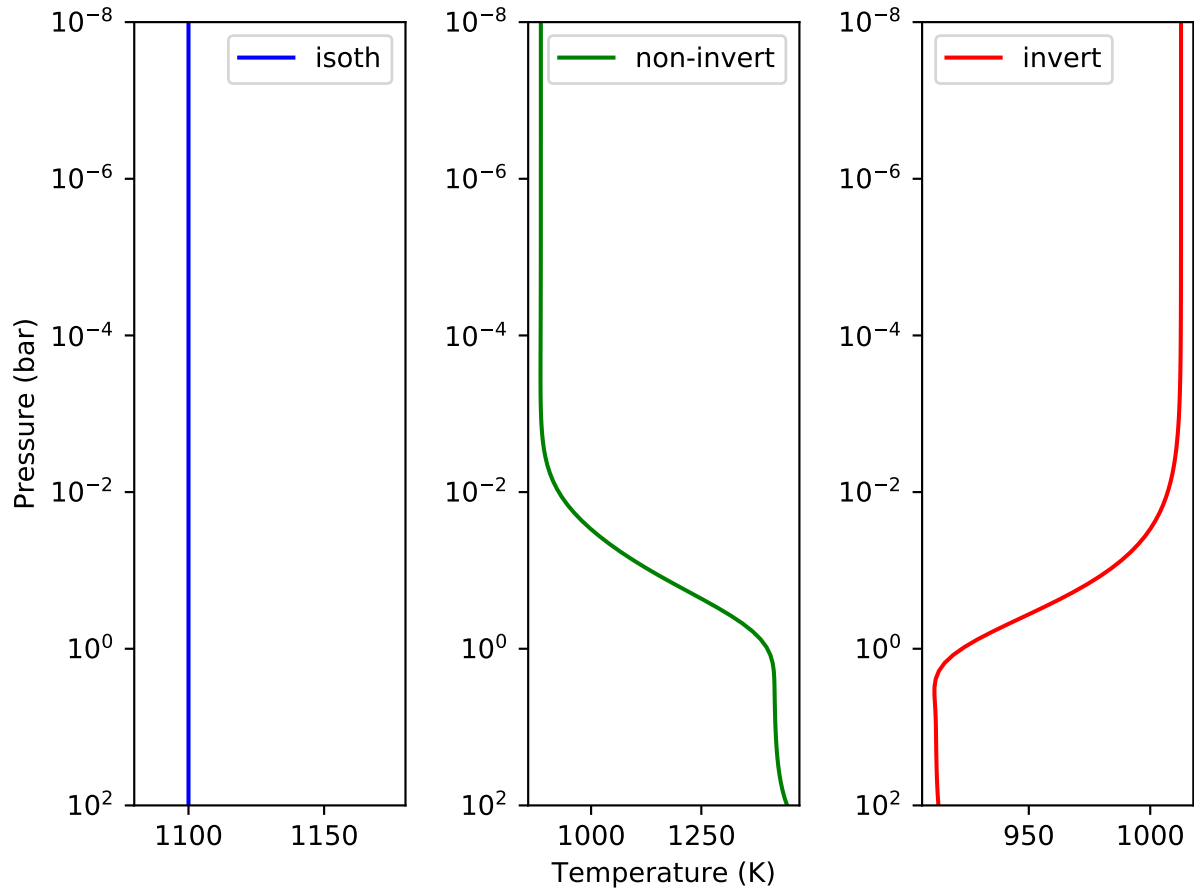


Figure 2.4: Isothermal, non-inverted, and inverted  $T(p)$  profiles used for the RT model tests `c01hjcLEARISO`, `c02hjcLEARNOINV`, `c03hjcLEARINV`, and their associated retrieval tests (`s01hjcLEARISO`, `s02hjcLEARNOINV`, and `s03hjcLEARINV`). The inverted and non-inverted profiles use the  $T(p)$  parameterization of Line et al. (2013). The profiles have effective temperatures generally around 1100 K, to resemble HD 189733 b.

Our model planet resembles HD 189733 b. Tests must adopt the stellar radius of 0.756 solar radii, stellar temperature of 5000 K, planetary radius of 1.138  $R_J$  (Torres et al. 2008), and planetary gravity of 2182.73  $\text{cm s}^{-2}$ , which corresponds to a mass of 1.14  $M_J$ . The reference pressure of 0.1 bar corresponds to this planetary radius. Tests must calculate spectra for the wavelength region between 1 and 11  $\mu\text{m}$ , as the most spectroscopically active species show features in this region. Tests should calculate both emission intensity spectra (secondary eclipse) and transmission modulation spectra (primary transit).

The common inputs are the isothermal, inverted, and non-inverted  $T(p)$  profiles given in Figure 2.4. Profiles are close to an effective temperature  $T_{\text{eff}} = 1100$  K, assuming zero albedo and uniform day-night distribution. They derive from the temperature-parameterization model of Line et al. (2013).

#### 2.1.2.1 *Barstow et al. (2020) Forward Models*

To compare BART with additional peer-reviewed codes, we emulate some of the setups described in Barstow et al. (2020), which were executed using the NEMESIS (Irwin et al. 2008, Lee et al. 2012), CHIMERA (Line et al. 2013), and TAU-REX (Waldmann et al. 2015b) codes. These codes utilize the correlated- $k$  method, whereas `transit` uses a line-by-line approach. (If the user-selected output resolution is low enough to miss some lines, it is properly called “line sampling” rather than “line-by-line”, but there is no difference in what the code does. As this is a user choice, we call it “line-by-line”, below.) Specifically, we emulate the setups for the cloud-free Model 0, cloudy Model 1, and three of the CO-only cases. We summarize these setups in Table 2.1 (see `c04hjc clearisoBarstowEtal` and `c05hjc cloudisoBarstowEtal`), but we direct

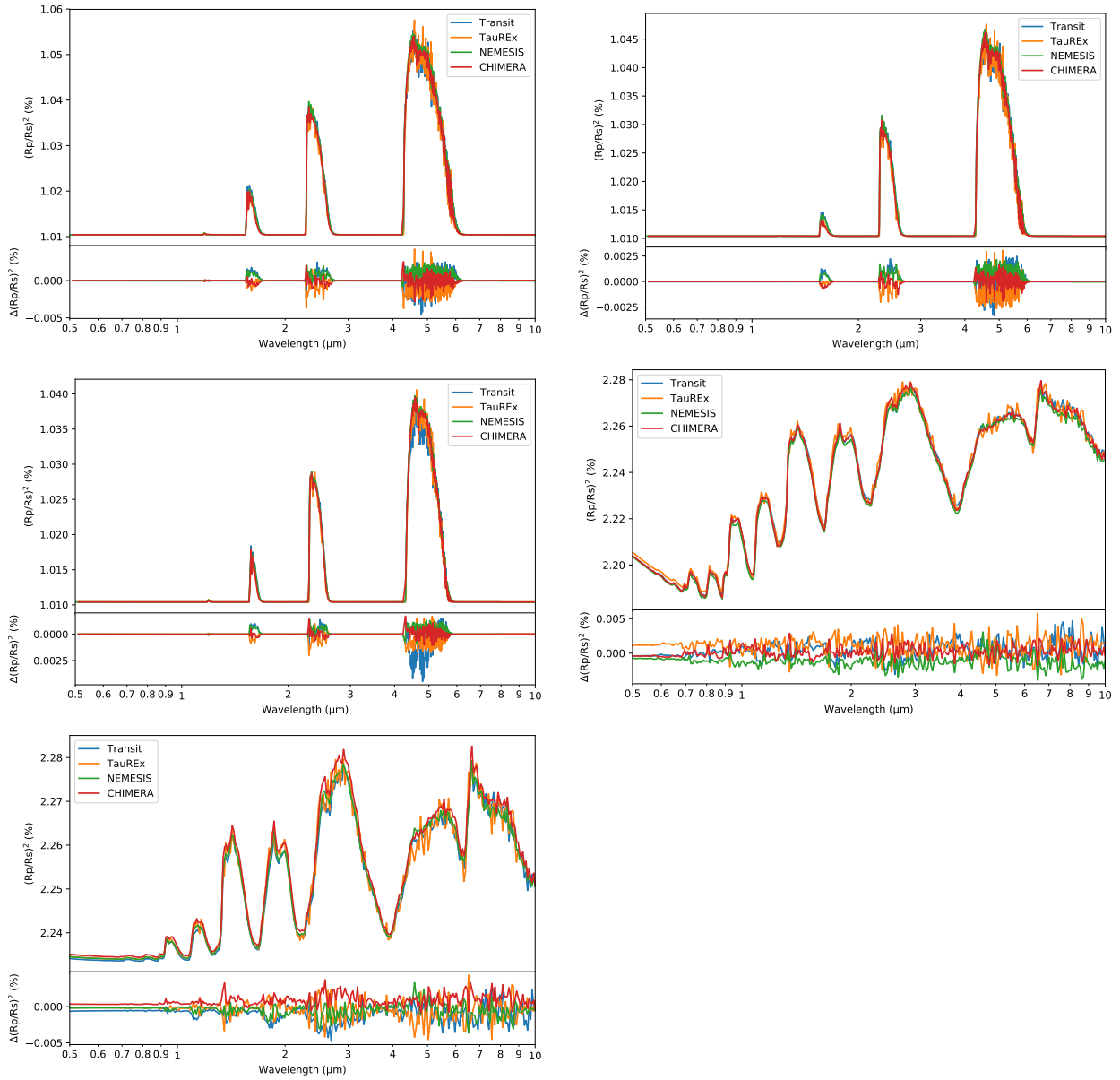


Figure 2.5: Forward spectra for the `c04hjclearisoBarstowEtal` and `c05hjcloudisoBarstowEtal` tests: CO at 1500 K for  $10^{-4}$  mixing ratio (top left) and  $10^{-5}$  mixing ratio (top right), CO at 1000 K for  $10^{-4}$  mixing ratio (middle left), the cloudfree HD 189733 b-inspired Model 0 (middle right) and cloudy Model 1 (bottom left). The plotted residuals are computed based on the average of the NEMESIS, CHIMERA, and TAU-REX spectra, as in Barstow et al. (2020).



readers to Barstow et al. (2020) for more detailed descriptions of the tests and to BARTTEST for the exact setups. We note that Barstow et al. (2020) report planetary radii in terms of Jupiter’s mean volumetric radius ( $R_{J,\text{mean}}$ , 69,911 km), rather than the IAU-defined value of  $R_J$  (71,492 km, Prša et al. 2016); not properly accounting for this will lead to a vertical offset in the transmission spectra.

Figure 2.5 shows comparisons between the spectra produced by `transit`, NEMESIS, CHIMERA, and TAU-REX. As in Barstow et al. (2020), we bin the CO spectra to steps of 0.01  $\mu\text{m}$  and compute the residuals with respect to the average of the NEMESIS, CHIMERA, and TAU-REX spectra; for the Model 0 and 1 cases, we bin according to CHIMERA’s reported wavelengths. In general, there is close agreement between `transit` and the other codes. The differences are on the order of the differences between the other codes, despite `transit`’s opacity-sampling approach.

### 2.1.3 Synthetic Retrieval Tests

To test BART retrievals, we used the synthetic planets of the `c01hjcleariso`, `c02hjclearnoinv`, `c03hjclearinv`, `c04hjclearisoBarstowEtal`, and `c05hjcloudisoBarstowEtal` tests. The tests are called `s01hjcleariso`, `s02hjclearnoinv`, `s03hjclearinv`, `s04hjclearisoBarstowEtal`, and `s05hjcloudisoBarstowEtal`. We consider both eclipse and transit geometry for each atmospheric model, except for the Barstow et al. (2020) cases, which are only in transmission.

To generate eclipse and transit depths for the  $s01 - s03$  cases, we use 47 channels spanning  $2 - 5 \mu\text{m}$  that have perfect transmission over their spectral ranges. We use relatively high-S/N synthetic data so that the retrieved credible regions will be relatively small, and thus more likely to expose small coding errors when compared to inputs. We set uncertainties such that each channel has  $S/N = 50$  for emission cases and 300 for transmission cases. For stellar emission, we use a K2 solar abundance Kurucz stellar model (Castelli & Kurucz 2003). For the  $s04$  and  $s05$  cases, we consider each of the simulated spectra by NEMESIS, CHIMERA, and TAU-REX with noise levels of 60 parts per million.

We include opacities for the species as described in Section 2.1.2. The  $s01 - s03$  retrievals each have five parameters for the  $T(p)$  profile (Line et al. 2013); five for the scaling factors of the log abundances of  $\text{H}_2\text{O}$ ,  $\text{CO}$ ,  $\text{CO}_2$ ,  $\text{CH}_4$ , and  $\text{NH}_3$ ; and, for the transmission cases, a parameter for the planetary radius at 0.1 bar. The  $s04$  and  $s05$  retrievals have free parameters for the isothermal temperature, the planetary radius at 10 bar, the log mixing ratios of  $\text{H}_2\text{O}$  and  $\text{CO}$ , and the pressure corresponding to an opaque cloudtop. All cases feature uniform priors on the model parameters; parameters that are the logarithm of the true parameter therefore have log-uniform priors.

BART's results for the  $s01 - s03$  retrievals are similar in some respects (Figures 2.6, 2.7, and 2.8). The retrieved thermal profiles and molecular abundances generally match the inputs in the regions of the atmospheres probed by these synthetic observations (Figures 2.7 and 2.8, middle columns). The lower atmospheres are generally poorly constrained, as the spectrum is minimally influenced by those pressure levels at the wavelengths of the synthetic data. The emission cases provide better constraints than the transmission cases on the  $T(p)$  profile and abundances, as expected (Griffith 2014, Heng & Kitzmann 2017, Madhusudhan 2018).

The isothermal emission case’s retrieved abundances and  $T(p)$  profile demonstrate the inability to detect molecular features from an isothermal atmosphere. In the region with sensitivity, the best-fit thermal profile is isothermal; the inversion seen in the explored thermal profiles corresponds to regions with negligible or no contribution to the spectrum. We allow for any  $T(p)$  profile rather than enforcing an isothermal condition because it would not be known *a priori* whether the atmosphere were isothermal. The 1D marginalized posteriors for the molecular abundances are poorly constrained and tend to favor a log mixing ratio  $< -4$ , with significant probability for log mixing ratios  $< -8$ , consistent with a lack of spectral features for an isothermal atmosphere.

Table 2.4 shows the SPEIS, ESS, and posterior accuracies for these retrievals. The large SPEIS (and small ESS) are due to a combination of factors. We choose the highest SPEIS value among all chains and all parameters as a conservative estimate; the non-inverted eclipse case has a SPEIS  $> 15000$ , with a median SPEIS of  $\sim 2730$ . Compared to the Barstow et al. (2020) cases and the HD 189733 b retrieval, these SPEIS values are significantly greater. This may be related to a numerical effect seen in synthetic retrievals tests (see Appendix A).

### 2.1.3.1 Barstow et al. (2020) Synthetic Retrievals

Figures 2.9 and 2.10 show the best-fit spectra and marginalized posteriors for the retrievals on the Barstow et al. (2020) synthetic spectra produced by NEMESIS, CHIMERA, and TAU-REX for Models 0 and 1, respectively, with an uncertainty of 60 ppm. Tables 2.5 and 2.6 summarize the retrieved credible regions for each of the retrievals. The true parameters are contained within the 95.45% credible regions, with most also being contained within the 68.27% regions.

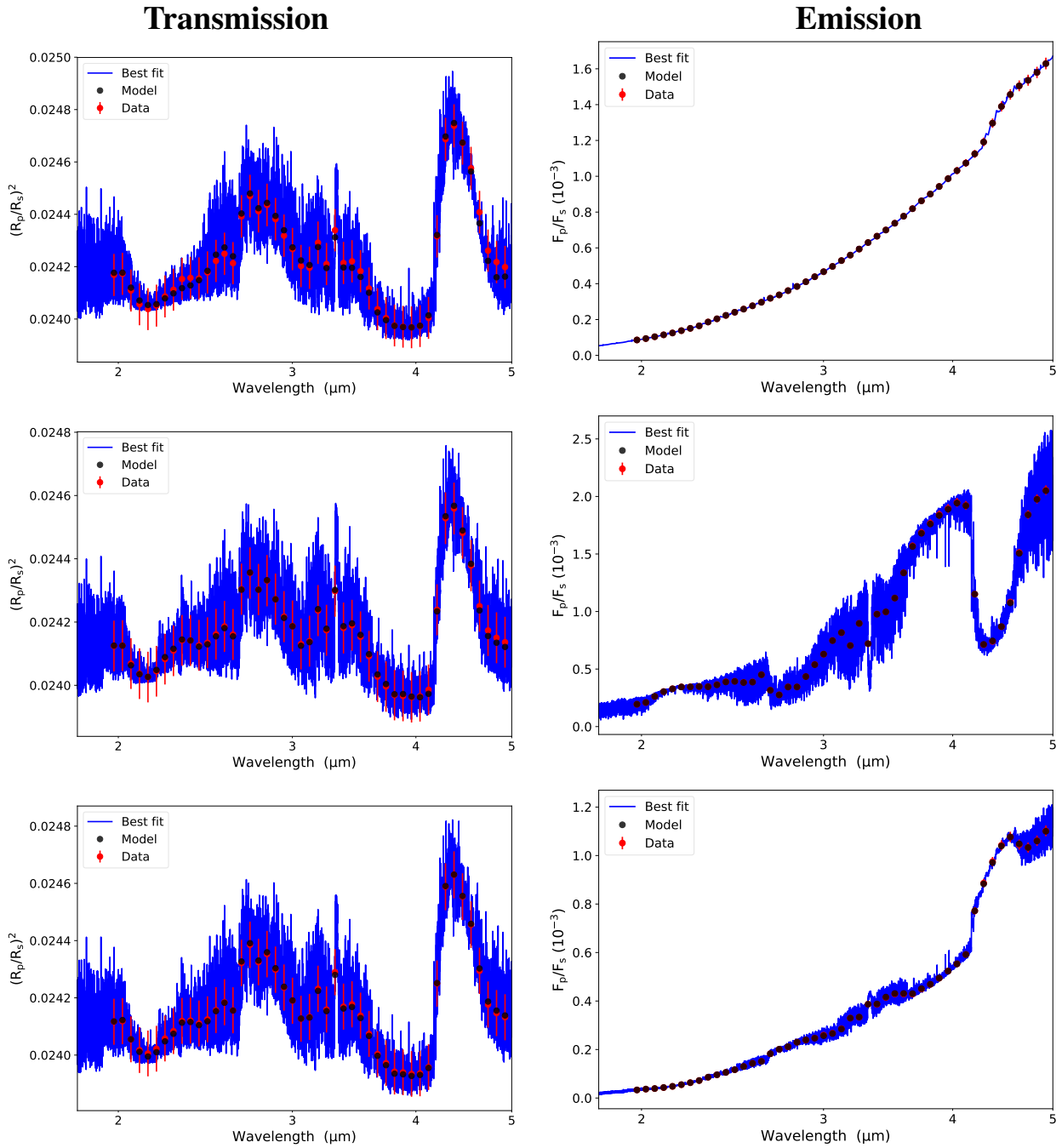


Figure 2.6: Best-fit spectra for the six synthetic retrieval tests of `s01hjcleariso`, `s02hjclearnoinv`, and `s03hjclearinv`: isothermal (top row), non-inverted (middle row), and inverted (bottom row) atmospheres in transit (left column) and eclipse (right column) geometries. The transmission data look similar for all three cases as they only differ in thermal profiles, to which transit geometry is generally insensitive.

# Emission

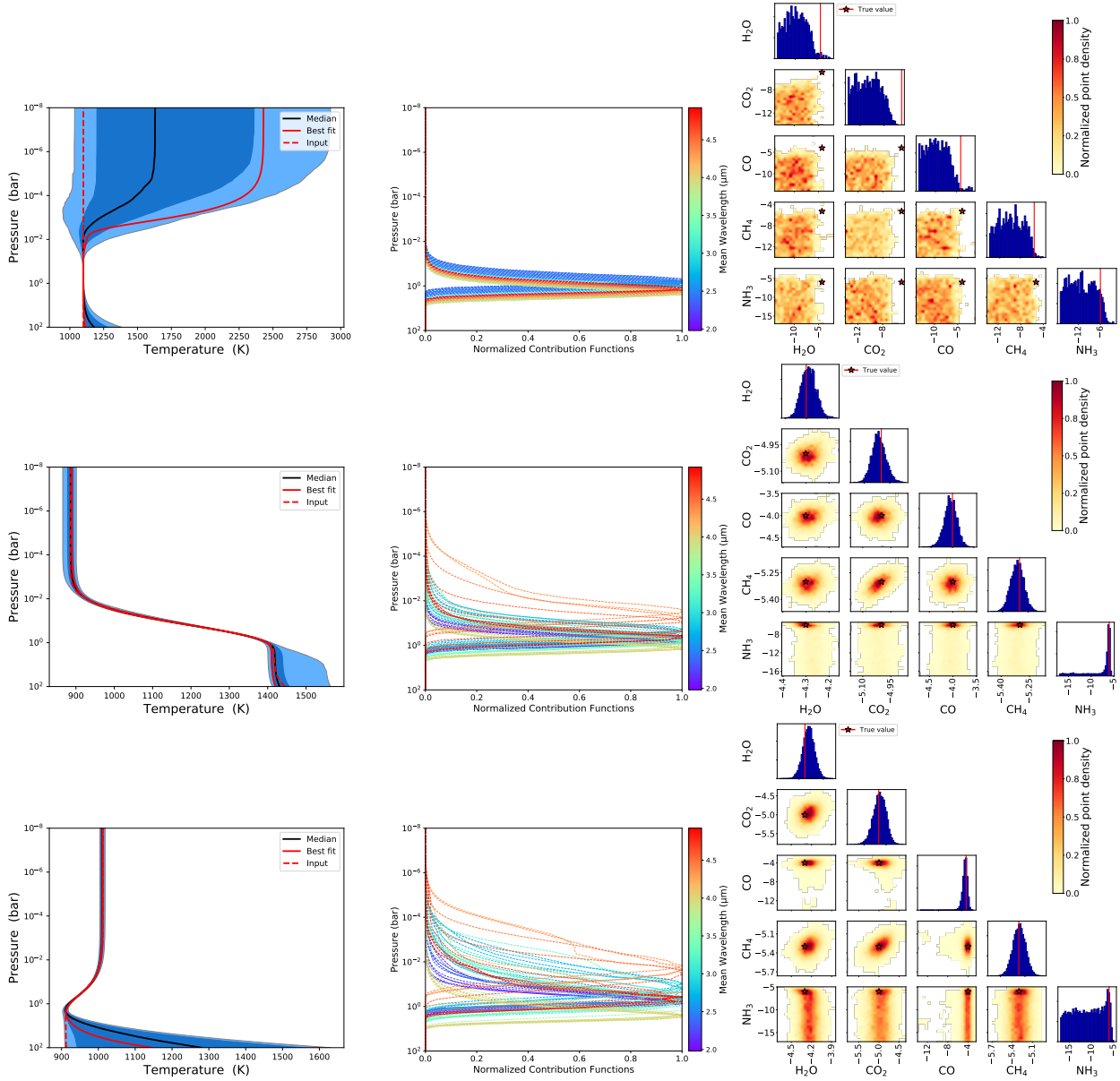


Figure 2.7: Comparison between input and retrieved  $T(p)$  profiles (left column), the normalized contribution functions (middle column; Blečić et al. 2022), and best-fit molecular abundances (right column) for the isothermal (top row), non-inverted (middle row), and inverted (bottom row) emission cases of the `s01hjc_cleariso`, `s02hjc_clearnoinv`, and `s03hjc_clearinv` tests. Dark- and light-blue shading in  $T(p)$  profile plots designate the 68.27% and 95.45% credible regions. These regions and the median derive from all the fits on a per-pressure-level basis. They do not follow the functional form of the individual profiles. Few, if any, individual  $T(p)$  profiles, including the best fit, stay confined to these regions, especially where the contribution functions indicate low sensitivity. BART accurately retrieves the  $T(p)$  profile and abundances wherever they contribute to the spectra.

## Transmission

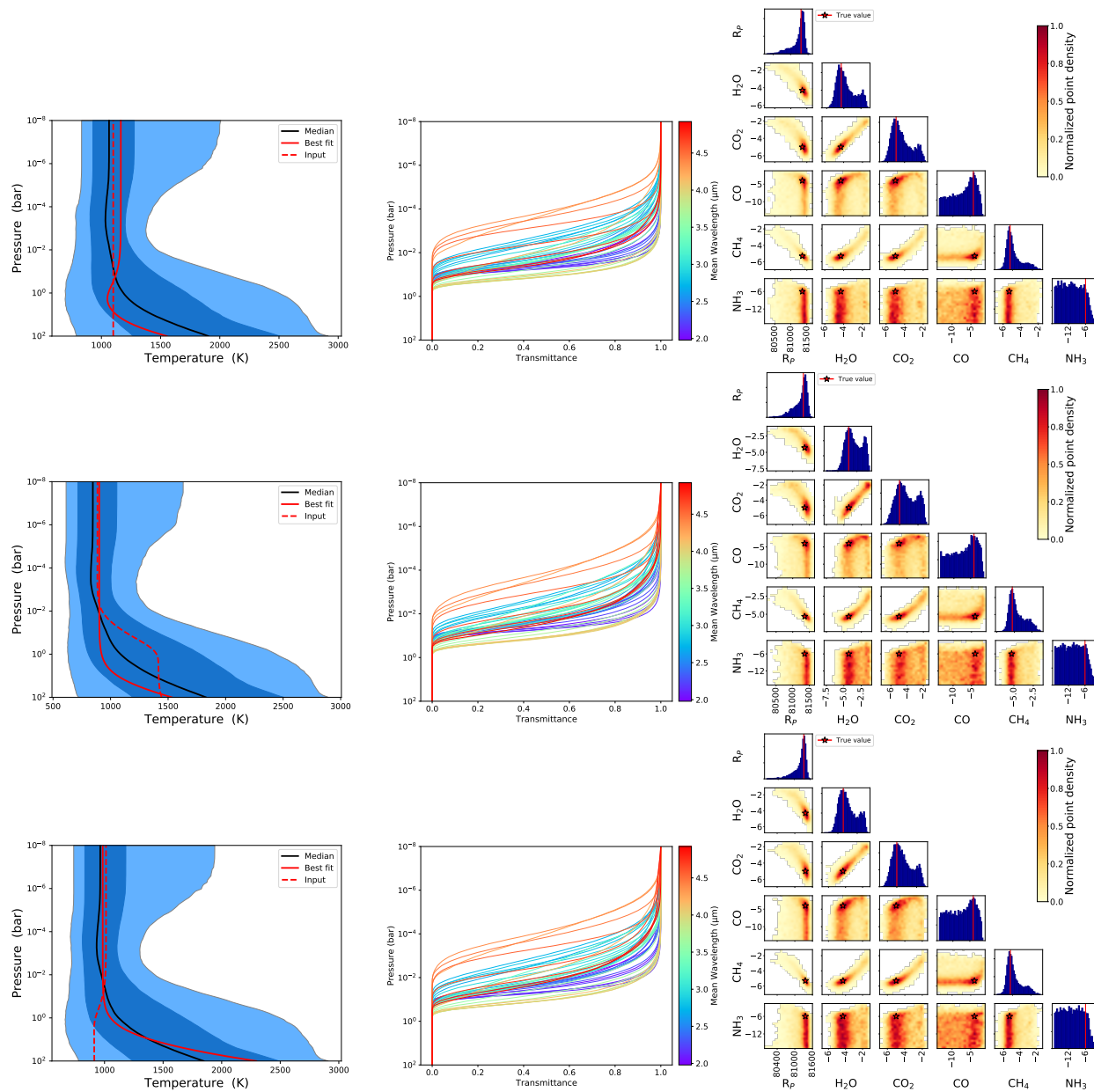


Figure 2.8: Same as Figure 2.7, but for transmission cases of the *s01hjc cleariso*, *s02hjc clearnoinv*, and *s03hjc clearinv* tests. (Recall that low-resolution transmission spectra are relatively insensitive to temperature structure.)

Table 2.4: BARTTEST Retrievals: Posterior Accuracy

Test	Geometry	SPEIS	ESS <sup>1</sup>	Credible Region Uncertainty		
				68.27% (“1 $\sigma$ ”) <sup>2</sup>	95.45% (“2 $\sigma$ ”) <sup>2</sup>	99.73% (“3 $\sigma$ ”) <sup>2</sup>
s01hjcleariso	Eclipse	5131	97	4.65%	2.08%	0.52%
	Transit	9204	162	3.62%	1.62%	0.40%
s02hjclearnoinv	Eclipse	15506	96	4.68%	2.09%	0.52%
	Transit	8129	184	3.40%	1.52%	0.38%
s03hjclearinv	Eclipse	14485	103	4.52%	2.02%	0.50%
	Transit	4553	329	2.55%	1.14%	0.28%
s04hjcleariso NEMESIS	Transit	54	1851	1.08%	0.48%	0.12%
s04hjcleariso CHIMERA	Transit	51	980	1.48%	0.66%	0.17%
s04hjcleariso TAU-REX	Transit	65	769	1.68%	0.75%	0.19%
s05hjcloudiso NEMESIS	Transit	511	978	1.49%	0.67%	0.17%
s05hjcloudiso CHIMERA	Transit	412	970	1.49%	0.67%	0.17%
s05hjcloudiso TAU-REX	Transit	812	862	1.58%	0.71%	0.18%
r01hd189733b	Eclipse	2084	959	1.50%	0.67%	0.17%

<sup>1</sup> Computed from the non-burned iterations for each case.

<sup>2</sup> Here and in the literature, these credible regions are labeled in analogy to the Gaussian, although they are not, generally, multiples of the posterior’s standard deviation.

Comparing the reported  $1\sigma$  credible regions for each retrieval shows minor differences. For most cases, BART finds narrower credible regions for temperature than the other codes. In the case of Model 0, the upper bounds of the  $1\sigma$  temperature regions are consistently just below the known 1500 K temperature, while the true radius falls at the lower bound of the  $2\sigma$  region of the BART retrieval on the TAU-REX forward model. For the Model 0 cases, BART favors high cloudtop pressures, consistent with the absence of clouds; slightly greater radii; and narrower credible regions for CO than the other codes. The CO retrieval on the NEMESIS spectrum falls in the  $2\sigma$  region. For H<sub>2</sub>O in the Model 0 cases, BART finds similar values, but with narrower credible regions compared to TAU-REX and CHIMERA. For the Model 1 cases, BART favors similar radii and lower cloudtop pressures. Compared to NEMESIS, BART finds narrower credible regions for CO and H<sub>2</sub>O when retrieving on TAU-REX and wider credible regions when retrieving on CHIMERA. Compared to TAU-REX, BART finds a similar amount of H<sub>2</sub>O but with greater uncertainty; BART also favors greater CO with similar or slightly smaller uncertainties. Compared to CHIMERA, BART favors greater CO and H<sub>2</sub>O when retrieving on NEMESIS, similar CO and less H<sub>2</sub>O when retrieving on TAU-REX, and generally finds narrower credible regions. For all cases except NEMESIS on CHIMERA, BART agrees with the other codes at  $1\sigma$  or less. The single exception agrees at just greater than  $1\sigma$ .

Together with the tests in Section 2.1.3, this demonstrates BART's ability to retrieve parameters accurately from synthetic data produced by various RT codes.



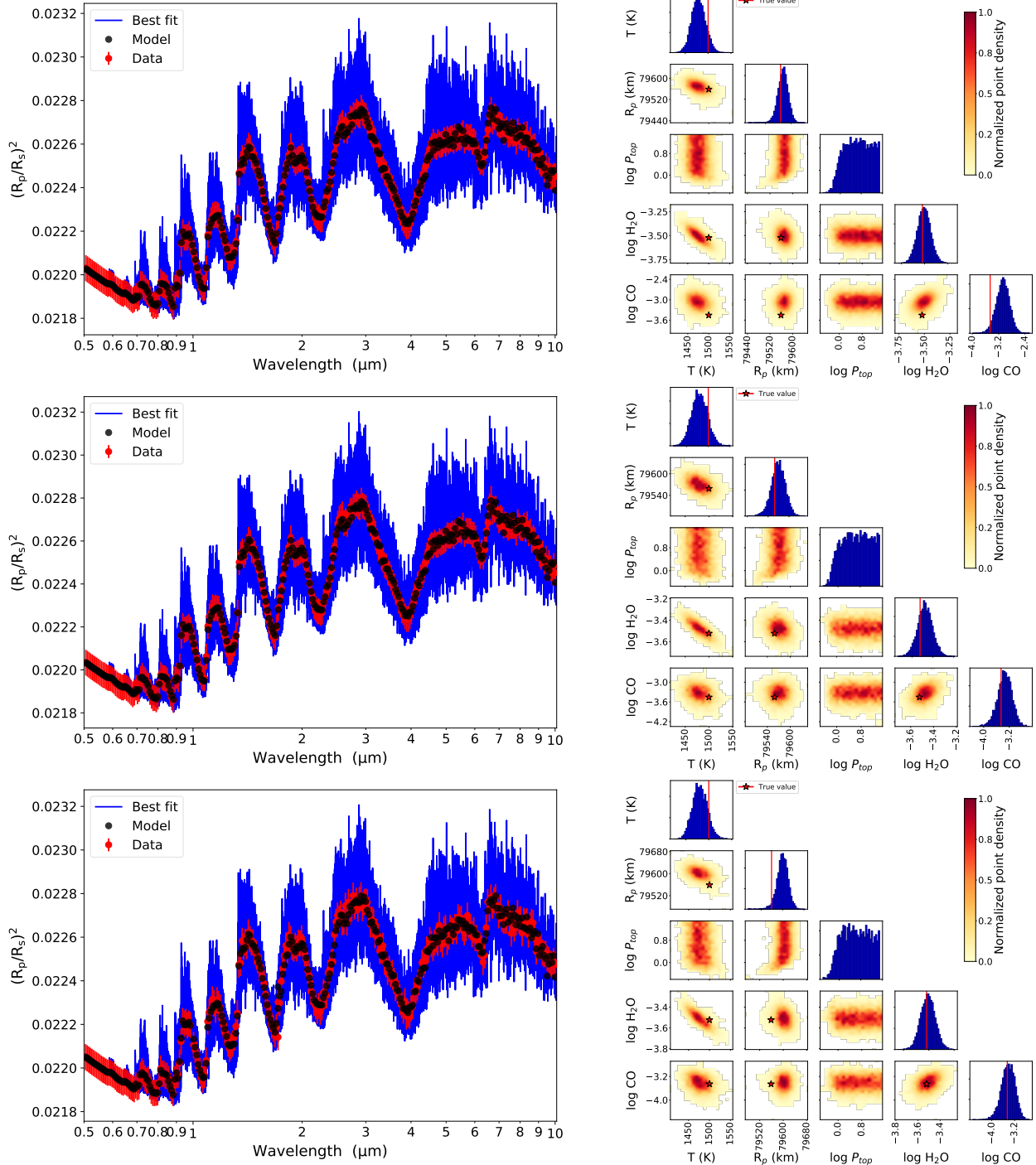


Figure 2.9: Summary of retrieval results for the s04hjc clearisoBarstowEtal test, which retrieve on the spectra of Barstow et al. (2020) Model 0 cases with 60 ppm uncertainties. Best-fit spectra (left column) and marginalized posteriors (right column) for the NEMESIS (top row), CHIMERA (middle row), and TAU-REX (bottom row) data sets.

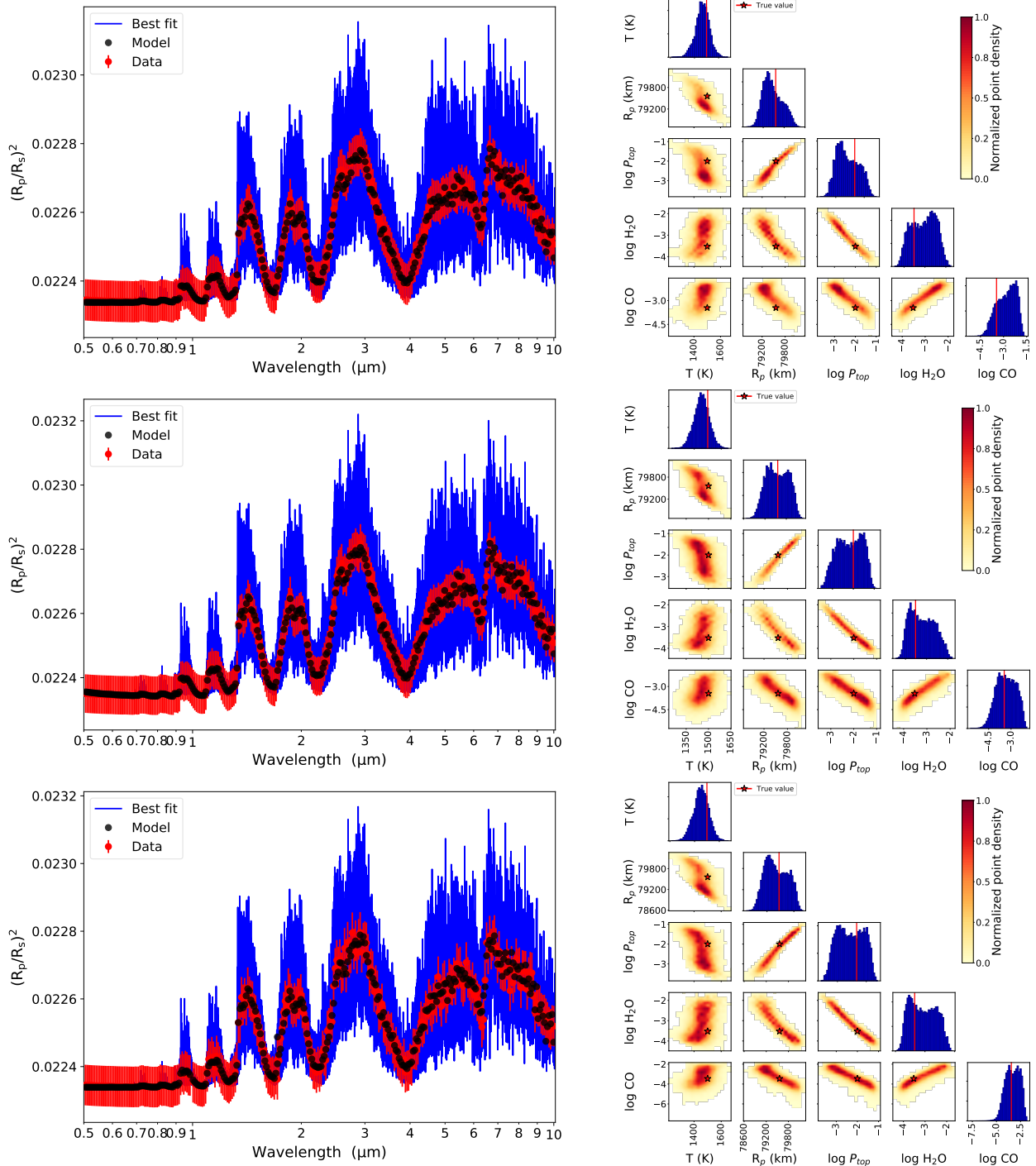


Figure 2.10: Summary of retrieval results for the s05hjccloudisoBarstowEtal test, which retrieve on the spectra of Barstow et al. (2020) Model 1 cases with 60 ppm uncertainties. Best-fit spectra (left column) and marginalized posteriors (right column) for the NEMESIS (top row), CHIMERA (middle row), and TAU-REX (bottom row) data sets.

Table 2.5: BARTTEST Retrievals, Barstow et al. (2020) Model 0 Cases: Credible Regions

Forward Model	T (K)	$R_p$ (km)	H <sub>2</sub> O (ppmv)	CO (ppmv)	log $P_{cloud}$
True	1500	79558.718	300	350	n/a
NEMESIS	[1455, 1492]	[79549, 79588]	[265, 369]	[548, 1372]	[0.17, 1.38]
	[1440, 1513]	[79527, 79608]	[223, 436]	[325, 2126]	[-0.12, 1.50]
	[1421, 1531]	[79477, 79626]	[190, 522]	[178, 3476]	[-0.40, 1.50]
CHIMERA	[1462, 1499]	[79551, 79590]	[286, 395]	[282, 783]	[0.24, 1.35]
	[1443, 1517]	[79529, 79610]	[242, 469]	[153, 1289]	[-0.15, 1.49]
	[1424, 1540]	[79502, 79628]	[204, 542]	[76, 2094]	[-0.38, 1.50]
TAU-REX	[1457, 1497]	[79581, 79621]	[260, 365]	[228, 670]	[0.04, 1.17]
	[1437, 1516]	[79557, 79643]	[221, 432]	[124, 1084]	[-0.18, 1.49]
	[1421, 1539]	[79471, 79660]	[180, 535]	[62, 1777]	[-0.46, 1.50]

For each data set, we report the 68.27%, 95.45%, and 99.73% credible regions, from top to bottom per model.

Table 2.6: BARTTEST Retrievals, Barstow et al. (2020) Model 1 Cases: Credible Regions

Forward Model	T (K)	$R_p$ (km)	H <sub>2</sub> O (ppmv)	CO (ppmv)	log $P_{cloud}$
True	1500	79558.718	300	350	-2.0
NEMESIS	[1416, 1524]	[79122, 79666]	[214, 5198]	[584, 10750]	[-3.13, -1.87]
	[1344, 1578]	[79019, 80024]	[82, 7905]	[110, 15241]	[-3.32, -1.30]
	[1270, 1630]	[78869, 80204]	[52, 14246]	[38, 22065]	[-3.55, -1.06]
CHIMERA	[1402, 1512]	[79221, 79947]	[104, 2418]	[96, 2678]	[-2.82, -1.36]
	[1342, 1564]	[79024, 80086]	[75, 7096]	[36, 6925]	[-3.22, -1.17]
	[1291, 1618]	[78865, 80195]	[52, 13029]	[11, 9897]	[-3.50, -1.03]
TAU-REX	[1399, 1520]	[79062, 79946]	[98, 568]	[100, 391]	[-3.21, -1.38]
	[1330, 1573]	[78929, 80088]	[69, 1068]	[24, 891]	[-3.43, -1.19]
	[1273, 1633]	[78774, 80206]	[47, 2006]	[3, 1448]	[-3.70, -1.02]

For each data set, we report the 68.27%, 95.45%, and 99.73% credible regions, respectively.

#### 2.1.4 *Real-Data Retrieval Test*

In a synthetic test, we may or may not know the answer as we work (e.g., in blind testing), but we know in principle what could have gone into the test. A test on real data is a full analysis, including concerns for unknown systematic errors, time variability of the star and planet, the 3D structure of the atmosphere, physics and chemistry not included in the model, errors and incompleteness in line lists, and even the unknown existence of background sources within the point-spread function of the target star. Given the state of exoplanet data today, we chose HD 189733 b, a system with a high planetary S/N, a relatively large number of observations in the literature, and little controversy over their interpretation. Those implementing this test for comparison to our result must configure their codes as we have, in one dimension, without clouds, using the same line lists, and using exactly the same observations. The value of any comparison test lies in the number of comparisons, so we reiterate our invitation to those willing to contribute results of their own application of this test.

In designing this test, we must mimic one of the two Bayesian models (discussed more in Section 2.2), by Line et al. (2014) and Waldmann et al. (2015a). Further, we wish the test to be accessible on a modest computer. We chose to emulate the analysis of Line et al. (2014) with CHIMERA, for two reasons. First, CHIMERA has been applied more broadly than Waldmann et al. (2015a)'s TAU-REX. Second, the TAU-REX analysis uses the Yurchenko & Tennyson (2014) CH<sub>4</sub> line list, which is so complete that it exceeds the storage capacity of many modest computers. Although `transit` can use this line list either directly or digested into a continuum opacity table and a separate list of the strongest lines (Cubillos 2017), not all codes can, and the TAU-REX test did not. Additionally, Hargreaves et al. (2020) showed that ExoMol's CH<sub>4</sub> list does not match laboratory studies.

This first test on real data is thus deliberately simple, as appropriate for a test suite, as it ignores clouds and the largest line lists. More-complete comparisons, with clouds, the ExoMol lists, numerous variations, and comparison to multiple other works, appear in BART2 and BART3. As this test evaluates whether algorithms consistently fit data not produced by, and unlikely to be fit perfectly by, their respective underlying RT codes, the metric of success is not  $\chi^2$  to the data, but similarity among retrievals.

Of course, as future observations and models improve, it may be desirable to add more complete tests, with more recent observations, line lists, physics, and code configurations. Given the length and complexity of the analysis, we present our test, including discussion of the literature for this planet, in its own section.

## 2.2 Application to HD 189733 b

Due to both its proximity to Earth and its high S/N, the atmosphere of HD 189733 b has been extensively studied since its discovery in 2005 (Bouchy et al. 2005). Being one of the most analyzed hot Jupiters to date, HD 189733 b is a prime candidate for a real-data retrieval test using published secondary-eclipse data. Here we discuss previous retrieval analyses of HD 189733 b (Madhusudhan & Seager 2009, Lee et al. 2012, Line et al. 2012, Moses et al. 2013, Line et al. 2014, Waldmann et al. 2015a) and BART’s retrieved atmospheric profiles and dayside emission spectra in the context of these prior analyses.

At an orbital semimajor axis of  $0.0312 \pm 0.00037$  AU and with an eccentricity of  $0.0041 \pm 0.0025$ , it takes  $2.21857312 \pm 0.00000076$  days for HD 189733 b to orbit its  $5052 \pm 16$  K, K-type host star (Triaud et al. 2009, Stassun et al. 2018). In Jovian units, its mass is  $1.130 \pm 0.025$  and its radius is  $1.178 \pm 0.023$  (Triaud et al. 2009), making its bulk density just over three-quarters of Jupiter's (Southworth 2010).

Previous studies on HD 189733 b use data from NICMOS (Swain et al. 2009), IRAC (Charbonneau et al. 2008, Knutson et al. 2009, Agol et al. 2010, Knutson et al. 2012), IRS (Deming et al. 2006, Grillmair et al. 2007), and MIPS (Charbonneau et al. 2008, Knutson et al. 2009).

When Madhusudhan & Seager (2009) published the first exoplanet retrieval via a parametric grid search, they derived that the atmospheric composition of HD 189733 b was high in CO and CO<sub>2</sub>, had a moderate abundance of H<sub>2</sub>O, and had minimal CH<sub>4</sub>. Further studies by Lee et al. (2012) and Line et al. (2012) confirm these abundances, though Moses et al. (2013) indicate that the comparatively large abundance of CO<sub>2</sub> is somewhat of an anomaly. Barstow et al. (2014) included clouds. These studies used optimal estimation rather than a Bayesian sampler. The upper limit on the CH<sub>4</sub> abundances presented by Line et al. (2012) is higher than previous results, such as Madhusudhan & Seager (2009) and Swain et al. (2009). Line et al. (2014) attribute the discrepancy to non-Gaussian posterior distributions under Gaussian-assuming optimal estimation.

We adopt the following data set: NICMOS data from Swain et al. (2009), with the 4 shortest-wavelength channels omitted; IRS data from Grillmair et al. (2008); IRAC 5.8  $\mu$ m, IRS 16  $\mu$ m photometric, and MIPS 24  $\mu$ m data of Charbonneau et al. (2008); and IRAC 8.0  $\mu$ m data of Agol et al. (2010). IRAC 3.6 and 4.5 data are adopted as  $0.1533 \pm 0.0029\%$  and  $0.1886 \pm 0.0071\%$ , consistent with Line et al. (2014).

While newer data (e.g., the secondary-eclipse measurements of Knutson et al. 2012) and analyses (e.g., the IRS re-analysis by Todorov et al. 2014) are available, this setup is meant as a real-data retrieval test to benchmark BART and future retrieval codes. This data set exactly matches that of Line et al. (2014), allowing a direct comparison of results. We note that Line et al. (2014) cite the IRS data as coming from Grillmair et al. (2007) but use data from Grillmair et al. (2008). The IRAC 5.8  $\mu\text{m}$  data are cited as coming from Agol et al. (2010), who did not publish 5.8  $\mu\text{m}$  data. Rather, they use the 5.8  $\mu\text{m}$  data of Charbonneau et al. (2008). Similarly, the IRAC 3.6 and 4.5  $\mu\text{m}$  data are cited as Knutson et al. (2012) but the data used do not appear to match any published data (M. Line, priv. comm.).

The retrieval model has nine free parameters: five for the  $T(p)$  profile (Line et al. 2013) and four scaling factors for the vertically constant log abundances of CO, CO<sub>2</sub>, CH<sub>4</sub>, and H<sub>2</sub>O. All priors are uniform, with the log parameters therefore having a log-uniform prior. The model atmosphere has 100 layers spanning  $10^{-8}$  – 100 bar, evenly spaced in log pressure. We include HITEMP opacities for CO, CO<sub>2</sub>, and H<sub>2</sub>O (Rothman et al. 2010, valid at the temperature of HD 189733 b), and HITRAN opacities for CH<sub>4</sub> (Rothman et al. 2013, measured at 296 K with some hot bands). We also include H<sub>2</sub>-H<sub>2</sub> and H<sub>2</sub>-He CIAs (Richard et al. 2012). For compatibility with the initial comparison studies, we do not include opacities for minor species whose abundances are not sought, although this is a good practice.

We also consider two additional models that are identical in setup except for line lists, to explore their effect on the retrieval results. To match the setup of Line et al. (2014), one model only differs for CH<sub>4</sub>, using the theoretically derived Spherical Top Database System (STDS, Wenger & Champion 1998) at wavelengths greater than 1.7  $\mu\text{m}$  and HITRAN 2008 (Rothman et al. 2009) at shorter wavelengths. The other model uses the latest HITEMP CH<sub>4</sub> (Hargreaves et al. 2020) and CO lists as well as ExoMol lists for H<sub>2</sub>O (Polyansky et al. 2018) and CO<sub>2</sub> (Yurchenko et al. 2020) processed via REPACK (Cubillos 2017). While both these fits are better, the setups are more

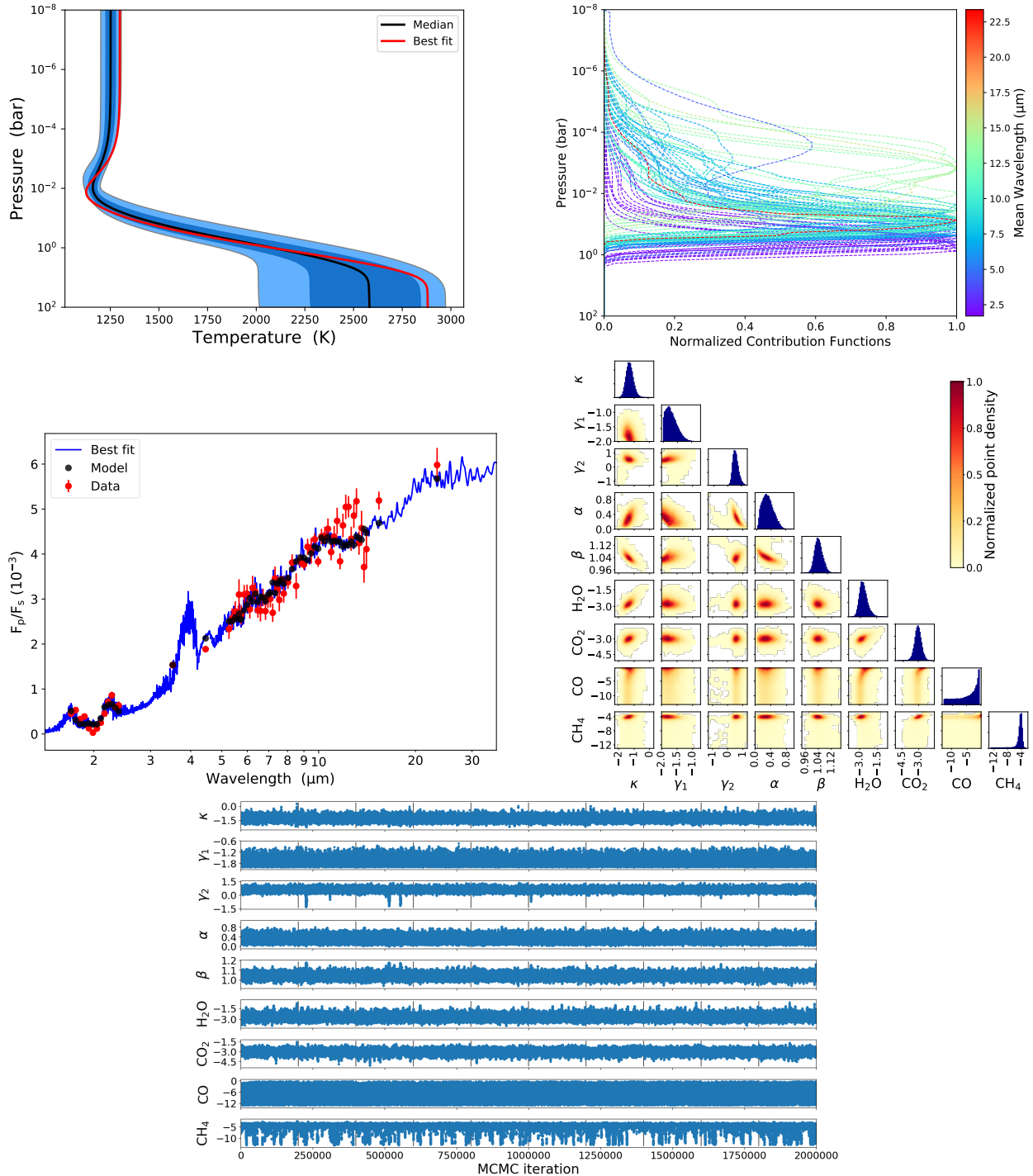


Figure 2.11: Retrieval results for HD189733 b (r01hd189733b test). **Top left:**  $T(p)$  profiles explored by the MCMC (see Figure 2.7). **Top right:** Normalized contribution functions of filters (see Figure 2.7). **Middle left:** Best-fit spectrum. **Middle right:** Pairwise correlations (marginalized over posterior) for all model parameters. Values are in base-10 logarithm, except for  $\alpha$  and  $\beta$ . **Bottom:** Trace plots of explored model parameters. The 10 chains are concatenated.



Table 2.7: Comparison of Fitted Log Abundances for HD 189733 b

Item	This work <sup>a</sup> , HH <sup>b</sup>	This work, STDS <sup>c</sup>	This work, EM <sup>d</sup>	Waldmann et al. 2015a Stage-1	Line et al. 2014 <sup>e</sup>	Lee et al. 2012 <sup>f</sup>	Line et al. 2012	MadhusudhanSwain & Seager 2009 <sup>g</sup>	et al. 2009 <sup>g</sup>
H <sub>2</sub> O	[-3.1, -2.4] [-3.4, -1.9] [-3.6, -1.4]	[-3.4, -2.6] [-3.8, -2.2] [-4.2, -1.6]	[-3.4, -2.7] [-3.7, -2.3] [-3.9, -1.8]	-3.9 ± 0.2	[-3.5, -2.9]	[-4.5, -2.0]	[-4.3, -3.5]	[-5.0, -3.0]	[-5.0, -4.0]
CH <sub>4</sub>	[-4.7, -3.7] [-10.5, -3.1] [-12.7, -3.1]	[-4.5, -3.9] [-4.8, -3.5] [-5.1, -3.1]	[-5.1, -4.5] [-5.2, -4.2] [-5.5, -3.7]	-6.7 ± 0.7	[-5.0, -4.6]	< -4.0	< -2.0	< -5.2	< -5.0
CO	[-6.6, -0.5] [-12.5, -0.5] [-12.9, -0.5]	[-1.8, -0.5] [-12.4, -0.5] [-12.7, -0.5]	[-4.8, -0.5] [-12.5, -0.5] [-12.9, -0.5]	-2.7 ± 1.4	[-4.6, -1.5]	n/a (best fit: -2.5)	[-2.4, -1.4]	[-4.0, -2.0]	[-4.0, -3.5]
CO <sub>2</sub>	[-3.3, -2.7] [-3.7, -2.3] [-4.0, -2.0]	[-2.6, -1.8] [-2.9, -1.4] [-3.2, -0.9]	[-2.9, -2.3] [-3.2, -2.0] [-3.5, -1.6]	-3.7 ± 0.5	[-2.9, -2.4]	[-3.8, -1.5]	[-2.8, -2.2]	~ -1.2	[-7.0, -6.0]
$\chi^2$	169.60	139.53	147.54	— <sup>h</sup>	149.82 <sup>i</sup>	—	—	—	—
Red. $\chi^{2j}$	2.98	2.45	2.59	—	2.63	—	—	—	—

<sup>a</sup> 68.27%, 95.45%, and 99.73% credible regions, stacked vertically.

<sup>b</sup> Main BARTTEST model, which uses HITRAN 2012 and HITEMP 2010 line lists.

<sup>c</sup> Same as HH, but using STDS CH<sub>4</sub> line list at wavelengths >1.7 μm and HITRAN 2008 at wavelengths <1.7 μm.

<sup>d</sup> Model using most recent ExoMol line lists for H<sub>2</sub>O and CO<sub>2</sub> and HITEMP line lists for CO and CH<sub>4</sub>.

<sup>e</sup> Reported 68.27% interval.

<sup>f</sup> Possible fit range for  $\Delta\chi^2/N < 1.0$  case, with n/a indicating no constraint.

<sup>g</sup> Reported range of model abundances.

<sup>h</sup> Comparing  $\chi^2$  between different models fitting different data does not tell which model is better, so we do not report the  $\chi^2$  of other studies besides Line et al. (2014).

<sup>i</sup> Calculated from their reported statistic of  $\chi^2/N_{\text{data}}$ .

<sup>j</sup> 57 degrees of freedom, though the independence of adjacent spectral channels is questionable if they sense the same molecular band.

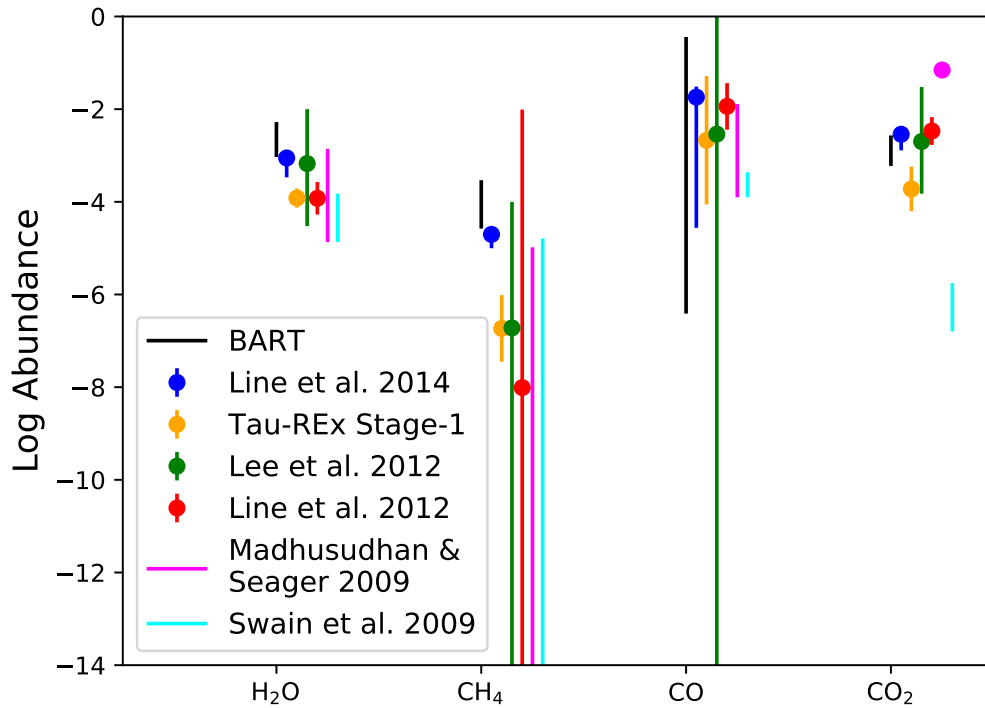


Figure 2.12: Comparison of best-fit retrieved log abundances for HD 189733 b: BART's retrieved 68% credible region, TAU-REX's Stage 1 result, CHIMERA's best fit and 68% interval, the best fit and possible fit range reported by Lee et al. (2012), and the range of values reported by Line et al. (2012), Madhusudhan & Seager (2009), and Swain et al. (2009).

complex, involving the relatively obscure STDS list and the large ExoMol database. Since none of these comes close to a perfect spectrum fit, the test in BARTTEST is the simplest of the three, using just HITRAN 2012 and HITEMP 2010. Those interested in reproducing the STDS and ExoMol runs can find the relevant lists, setup details, and plots in this article’s electronic compendium (see below).

Figure 2.11 shows BART’s retrieved results for HD 189733 b. Except for Line et al. (2014), each of the studies from the literature differs from BART in several fundamental ways, including data, line lists, and modeling approach. Figure 2.12 and Table 2.7 show the effect of these differences among the studies.

In the BARTTEST run, except for the IRAC channel 2 data point, the best-fit spectrum qualitatively agrees with that of Line et al. (2014), and the retrieved abundances agree at  $1\sigma$ , as shown in Table 2.7 and Figure 2.12. While Line et al. (2014) find no evidence of a thermal inversion, BART favors a slight inversion in all three models, with retrieved  $T(p)$  parameters closely agreeing. Except for CO<sub>2</sub>, the retrieved molecular-abundance credible regions for the three models agree at  $1\sigma$ . For CO<sub>2</sub>, the ExoMol model agrees with the BARTTEST and STDS models at  $1\sigma$ , while the BARTTEST and STDS models differ by less than  $2\sigma$ . When restricting BART to non-inverted thermal profiles,  $\chi^2$  increases significantly, favoring the inverted model by a maximum likelihood ratio of 30 – 90, depending on the forward-model gridding (0.1 vs. 1 cm<sup>-1</sup>, respectively).

We demonstrated  $1\sigma$  agreement between BART and the modern CHIMERA in Section 2.1.3.1, using synthetic cases. In those tests, the forward model used in the retrieval, or one very similar to it, generated the test data, so a (near-)perfect match exists in the retrieval phase space. Yet, despite similar spectra and mostly consistent posteriors,  $\chi^2$  values in Table 2.7 differ substantially. The reduced  $\chi^2$  values are greater than two, indicating model misspecification. Real planets will

always have physics that are not in any model, in this case including additional opacity sources, more sophisticated and varied clouds, some reflected stellar spectrum, and 3D temperature and compositional variation. There are also well known, uncorrected systematics in the NICMOS data (e.g., Gibson et al. 2011, Crouzet et al. 2012). The error from such misspecification must distribute somehow among the parameters, but model differences could distribute it differently.

There are some modest model differences. BART’s Bayesian sampler was DEMCzs vs. DEMC for CHIMERA. All Bayesian methods should converge to the same posterior, within the noise of random sampling. For example, we find similar results to CHIMERA’s PyMultiNest with our DEMCzs in Section 2.1.3.1, and these algorithms are less similar than DEMCzs and the DEMC of Line et al. (2014). So, we do not blame the samplers for the discrepancy among models. BART uses Kurucz stellar models, while Line et al. (2014) use PHOENIX, but Martins & Coelho (2007) found that Kurucz and PHOENIX models are comparable in the near infrared for stars with effective temperatures greater than 4250 K, like HD 189733. The pre-computed opacity grid of Line et al. 2014 used 20 temperatures ranging 500–3000 K and 20 pressures ranging  $20\text{--}10^{-6}$  bar, while BART used 25 temperatures spanning 600–3000 K and 100 pressures ranging  $100\text{--}10^{-8}$  bar. This may explain BART’s slight improvement in reduced  $\chi^2$  for the STDS case, but not likely the inversion, as the contribution plots indicate little sensitivity in the extended regions and the pressure gridding should be sufficient for interpolation in both cases. Priors on thermal-profile parameters differed (uniform vs. Gaussian), which could have kept Line et al. (2014) from finding a  $\chi^2$  minimum with an inverted profile.

The model differences above lead to a  $\chi^2$  improvement of 10.29 from Line et al. (2014) to our STDS case, which uses the same line lists, a maximum likelihood ratio of 172. One might expect even more improvement using the much-more-complete HITEMP and ExoMol line lists. Instead,  $\chi^2$  deteriorates to just 2.28 better than Line et al. (2014, maximum likelihood ratio of 3.1). Evidently, the need to distribute the misspecification error dominates the improvement in the modern line lists.

The next-closest study, and the only other Bayesian approach, used TAU-REX 2's Stage-1 MCMC (Waldmann et al. 2015a). With the release of version 3 (Al-Refaie et al. 2021), many new results from TAU-REX 2's Stage-1 MCMC are not anticipated, so we have not emulated it directly, but we can discuss it.

They omit the IRAC 5.8  $\mu\text{m}$  point, which is quite constraining, due to its smaller uncertainties than the IRS spectrum at those wavelengths. At 10 bar, their uncertainties on  $T(p)$  are small (their Figure 15), despite the data not probing to that depth. This may be due to conditioning  $T(p)$ . Our Figure 2.11 and similar plots elsewhere show the lack of contribution from that level and the resultant broadening of the  $T(p)$  credible region there.

They used the  $\text{CH}_4$  line list of Yurchenko & Tennyson (2014). Its higher limiting temperature yields many times the lines of the test's HITRAN list, and thus greater overall opacity. This tends to reduce the abundance of  $\text{CH}_4$ , and could also reduce other species' abundances, if  $\text{CH}_4$  opacity appeared where there had been none previously. This may explain their usually-lower retrieved abundances (Table 2.7 and Figure 2.12). BART's retrieved  $T(p)$  profile (all three cases) differs substantially, with a lower temperature in the upper atmosphere and a lower tropopause pressure ( $10^{-2}$  vs.  $10^{-1}$  bar) than TAU-REX's. Like Line et al. (2014), they also find no inversion.

In Table 2.7 and Figure 2.12, we also provide the fitted ranges for pre-Bayesian-retrieval abundances reported by Madhusudhan & Seager (2009), Swain et al. (2009), Lee et al. (2012), and Line et al. (2012). We find agreement within  $3\sigma$  for most molecular abundances. BART's results differ at  $3\sigma$  for  $\text{CO}_2$  and  $\text{H}_2\text{O}$  reported by Swain et al. (2009) and  $\text{CO}_2$  reported by Madhusudhan & Seager (2009). Like Lee et al. (2012), we similarly find that CO is poorly constrained. In the case of  $T(p)$  profiles, Lee et al. (2012) found the upper atmosphere to be isothermal ( $\sim 1100$  K) down to  $\sim 0.1$  bar. Line et al. (2012) find a variety of potential  $T(p)$  profiles, some of which are consistent with Lee et al. (2012), and some of which are consistent with the results of BART and Line et al. (2014). The  $T(p)$  profiles explored by Madhusudhan & Seager (2009) generally agree with these other analyses, with the exception of the upper atmosphere, which is cooler. Swain et al. (2009) did not publish a  $T(p)$  profile. These investigations used either a grid search or optimal estimation rather than a Bayesian method, and they used a different data set from that used here. Both of these differences could explain discrepancies in retrieved parameters.

### 2.3 List of References

- Agol, E., Cowan, N. B., Knutson, H. A., Deming, D., Steffen, J. H., Henry, G. W., & Charbonneau, D. 2010, ApJ, 721, 1861
- Al-Refaie, A. F., Changeat, Q., Waldmann, I. P., & Tinetti, G. 2021, ApJ, 917, 37
- Barstow, J. K., Aigrain, S., Irwin, P. G. J., Hackler, T., Fletcher, L. N., Lee, J. M., & Gibson, N. P. 2014, ApJ, 786, 154
- Barstow, J. K., Changeat, Q., Garland, R., Line, M. R., Rocchetto, M., & Waldmann, I. P. 2020, MNRAS, 493, 4884

- Blecic, J., Harrington, J., Cubillos, P. E., Bowman, M. O., Rojo, P. M., Stemm, M., Challener, R. C., Himes, M. D., Foster, A. J., Dobbs-Dixon, I., Foster, A. S. D., Lust, N. B., Blumenthal, S. D., Bruce, D., & Loredó, T. J. 2022, *The Planetary Science Journal*, 3, 82
- Bouchy, F., Udry, S., Mayor, M., Moutou, C., Pont, F., Iribarne, N., da Silva, R., Illovaisky, S., Queloz, D., Santos, N. C., Ségransan, D., & Zucker, S. 2005, *A&A*, 444, L15
- Castelli, F. & Kurucz, R. L. 2003, in *IAU Symposium*, Vol. 210, *Modelling of Stellar Atmospheres*, ed. N. Piskunov, W. W. Weiss, & D. F. Gray, A20
- Charbonneau, D., Knutson, H. A., Barman, T., Allen, L. E., Mayor, M., Megeath, S. T., Queloz, D., & Udry, S. 2008, *ApJ*, 686, 1341
- Crouzet, N., McCullough, P. R., Burke, C., & Long, D. 2012, *ApJ*, 761, 7
- Cubillos, P. E. 2017, *ApJ*, 850, 32
- Deming, D., Harrington, J., Seager, S., & Richardson, L. J. 2006, *ApJ*, 644, 560
- Gibson, N. P., Pont, F., & Aigrain, S. 2011, *Monthly Notices of the Royal Astronomical Society*, 411, 2199
- Griffith, C. A. 2014, *Philosophical Transactions of the Royal Society of London Series A*, 372, 20130086
- Grillmair, C. J., Burrows, A., Charbonneau, D., Armus, L., Stauffer, J., Meadows, V., van Cleve, J., von Braun, K., & Levine, D. 2008, *Nature*, 456, 767
- Grillmair, C. J., Charbonneau, D., Burrows, A., Armus, L., Stauffer, J., Meadows, V., Van Cleve, J., & Levine, D. 2007, *ApJ*, 658, L115
- Hargreaves, R. J., Gordon, I. E., Rey, M., Nikitin, A. V., Tyuterev, V. G., Kochanov, R. V., & Rothman, L. S. 2020, *ApJS*, 247, 55

- Heng, K. & Kitzmann, D. 2017, MNRAS, 470, 2972
- Irwin, P. G. J., Teanby, N. A., de Kok, R., Fletcher, L. N., Howett, C. J. A., Tsang, C. C. C., Wilson, C. F., Calcutt, S. B., Nixon, C. A., & Parrish, P. D. 2008, J. Quant. Spec. Radiat. Transf., 109, 1136
- Knutson, H. A., Charbonneau, D., Cowan, N. B., Fortney, J. J., Showman, A. P., Agol, E., Henry, G. W., Everett, M. E., & Allen, L. E. 2009, ApJ, 690, 822
- Knutson, H. A., Lewis, N., Fortney, J. J., Burrows, A., Showman, A. P., Cowan, N. B., Agol, E., Aigrain, S., Charbonneau, D., Deming, D., Désert, J.-M., Henry, G. W., Langton, J., & Laughlin, G. 2012, ApJ, 754, 22
- Lee, J.-M., Fletcher, L. N., & Irwin, P. G. J. 2012, MNRAS, 420, 170
- Line, M. R., Knutson, H., Wolf, A. S., & Yung, Y. L. 2014, ApJ, 783, 70
- Line, M. R., Wolf, A. S., Zhang, X., Knutson, H., Kammer, J. A., Ellison, E., Deroo, P., Crisp, D., & Yung, Y. L. 2013, ApJ, 775, 137
- Line, M. R., Zhang, X., Vasisht, G., Natraj, V., Chen, P., & Yung, Y. L. 2012, ApJ, 749, 93
- Madhusudhan, N. 2018, in Handbook of Exoplanets, ed. H. J. Deeg & J. A. Belmonte (Springer International Publishing AG), 104
- Madhusudhan, N. & Seager, S. 2009, ApJ, 707, 24
- Martins, L. P. & Coelho, P. 2007, Monthly Notices of the Royal Astronomical Society, 381, 1329
- Moses, J. I., Madhusudhan, N., Visscher, C., & Freedman, R. S. 2013, ApJ, 763, 25
- Pierluissi, J. 1977, J. Quant. Spec. Radiat. Transf., 18, 555



- Polyansky, O. L., Kyuberis, A. A., Zobov, N. F., Tennyson, J., Yurchenko, S. N., & Lodi, L. 2018, *Monthly Notices of the Royal Astronomical Society*, 480, 2597
- Prša, A., Harmanec, P., Torres, G., Mamajek, E., Asplund, M., Capitaine, N., Christensen-Dalsgaard, J., Depagne, É., Haberreiter, M., Hekker, S., Hilton, J., Kopp, G., Kostov, V., Kurtz, D. W., Laskar, J., Mason, B. D., Milone, E. F., Montgomery, M., Richards, M., Schmutz, W., Schou, J., & Stewart, S. G. 2016, *The Astronomical Journal*, 152, 41
- Richard, C., Gordon, I. E., Rothman, L. S., Abel, M., Frommhold, L., Gustafsson, M., Hartmann, J.-M., Hermans, C., Lafferty, W. J., Orton, G. S., Smith, K. M., & Tran, H. 2012, *J. Quant. Spec. Radiat. Transf.*, 113, 1276
- Rothman, L., Gordon, I., Barbe, A., Benner, D., Bernath, P., Birk, M., Boudon, V., Brown, L., Campargue, A., Champion, J.-P., Chance, K., Coudert, L., Dana, V., Devi, V., Fally, S., Flaud, J.-M., Gamache, R., Goldman, A., Jacquemart, D., Kleiner, I., Lacome, N., Lafferty, W., Mandin, J.-Y., Massie, S., Mikhailenko, S., Miller, C., Moazzen-Ahmadi, N., Naumenko, O., Nikitin, A., Orphal, J., Perevalov, V., Perrin, A., Predoi-Cross, A., Rinsland, C., Rotger, M., imekov, M., Smith, M., Sung, K., Tashkun, S., Tennyson, J., Toth, R., Vandaele, A., & Vander Auwera, J. 2009, *Journal of Quantitative Spectroscopy and Radiative Transfer*, 110, 533, HITRAN
- Rothman, L. S., Gordon, I. E., Babikov, Y., Barbe, A., Benner, D. C., Bernath, P. F., Birk, M., Bizzocchi, L., Boudon, V., Brown, L. R., et al. 2013, *J. Quant. Spec. Radiat. Transf.*, 130, 4
- Rothman, L. S., Gordon, I. E., Barber, R. J., Dothe, H., Gamache, R. R., Goldman, A., Perevalov, V. I., Tashkun, S. A., & Tennyson, J. 2010, *J. Quant. Spec. Radiat. Transf.*, 111, 2139
- Sharp, C. M. & Burrows, A. 2007, *ApJS*, 168, 140
- Southworth, J. 2010, *MNRAS*, 408, 1689
- Stassun, K. G., Corsaro, E., Pepper, J. A., & Gaudi, B. S. 2018, *AJ*, 155, 22

- Swain, M. R., Vasisht, G., Tinetti, G., Bouwman, J., Chen, P., Yung, Y., Deming, D., & Deroo, P. 2009, *ApJ*, 690, L114
- Todorov, K. O., Deming, D., Burrows, A., & Grillmair, C. J. 2014, *ApJ*, 796, 100
- Torres, G., Winn, J. N., & Holman, M. J. 2008, *ApJ*, 677, 1324
- Triaud, A. H. M. J., Queloz, D., Bouchy, F., Moutou, C., Collier Cameron, A., Claret, A., Barge, P., Benz, W., Deleuil, M., Guillot, T., Hébrard, G., Lecavelier Des Étangs, A., Lovis, C., Mayor, M., Pepe, F., & Udry, S. 2009, *A&A*, 506, 377
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., & SciPy 1.0 Contributors. 2020, *Nature Methods*, 17, 261
- Waldmann, I. P., Rocchetto, M., Tinetti, G., Barton, E. J., Yurchenko, S. N., & Tennyson, J. 2015a, *ApJ*, 813, 13
- Waldmann, I. P., Tinetti, G., Rocchetto, M., Barton, E. J., Yurchenko, S. N., & Tennyson, J. 2015b, *ApJ*, 802, 107
- Wenger, C. & Champion, J. 1998, *Journal of Quantitative Spectroscopy and Radiative Transfer*, 59, 471, *atmospheric Spectroscopy Applications* 96
- Yurchenko, S. N., Mellor, T. M., Freedman, R. S., & Tennyson, J. 2020, *Monthly Notices of the Royal Astronomical Society*, 496, 5282
- Yurchenko, S. N. & Tennyson, J. 2014, *MNRAS*, 440, 1649

## CHAPTER 3: ON THE DAYSIDE ATMOSPHERE OF WASP-12b

Michael D. Himes<sup>1</sup>, Joseph Harrington<sup>1</sup>

<sup>1</sup> *Planetary Sciences Group, Department of Physics, University of Central Florida, Orlando, FL 32816-2385*

Received 23 July 2019.

Accepted 9 October 2020.

Published in *The Astrophysical Journal* 26 May 2022.

Himes, M. D., Harrington, J. 2022, ApJ, 931, 86

<https://doi.org/10.3847/1538-4357/ac1e9f>

©AAS. Reproduced with permission.

### 3.1 Abstract

The atmospheric structure of WASP-12b has been hotly contested for years, with disagreements on the presence of a thermal inversion as well as the carbon-to-oxygen ratio,  $C/O$ , due to retrieved abundances of  $H_2O$ ,  $CO_2$ , and other included species such as  $HCN$  and  $C_2H_2$ . Previously, these difficult-to-diagnose discrepancies have been attributed to model differences; assumptions in these models were thought to drive retrievals toward different answers. Here, we show that some of these differences are independent of model assumptions and are instead due to subtle differences in the inputs, such as the eclipse depths and line-list databases. We replicate previously published retrievals and find that the retrieved results are data driven and are mostly unaffected by the addition of species such as  $HCN$  and  $C_2H_2$ . We also propose a new physically motivated model that takes into consideration the formation of  $H^-$  via the thermal dissociation of  $H_2O$  and  $H_2$  at the temperatures reached in the dayside atmosphere of WASP-12b, but the data's current resolution does not support its inclusion in the atmospheric model. This study raises the concern that other exoplanet retrievals may be similarly sensitive to slight changes in the input data.

### 3.2 Introduction

The thousands of exoplanets discovered to date span a wide range of properties and conditions, from small, rocky bodies to hot, Jupiter-like gas giants (Batalha 2014, Winn & Fabrycky 2015). Understanding the compositions of these planets provides real-world tests for atmospheric simulations and formation theories. Characterizing exoplanetary atmospheres requires an observed spectrum. Presently, most exoplanets can only be characterized via transit (when the planet moves in front of its host star, as seen from Earth) measurements. These observations capture starlight

that has filtered through the exoplanet’s atmosphere at the day–night terminator, imprinting information about its composition. For hot exoplanets, its secondary eclipse (when the planet moves behind the star, as seen from Earth) can be observed, which measures the planet’s thermal emission. This provides better data than transits to constrain the atmospheric properties (Deming & Seager 2017).

The inference of atmospheric conditions from observed spectra is known as atmospheric retrieval (Madhusudhan 2018). For exoplanets, atmospheric retrieval involves the proposal of atmospheric models from some prior distribution (e.g., uniform or Gaussian), computation of the theoretical observed spectra, and determination of how well the proposed models explain the observations. Unlike solar system observations which only require least-squares minimization, a Bayesian approach is better suited to estimating the uncertainties of exoplanet retrievals due to the high relative noise levels. The Bayesian sampler explores the parameter space and accepts/rejects new models with some probability based in part on the goodness of fit. The collection of accepted models forms the posterior distribution, which informs the range and relative likelihood of values for the model parameters.

WASP-12b stands out as one of the hottest exoplanets found to date. It orbits an F9V star with a temperature of  $6360 \pm_{140}^{130}$  K at  $0.02340 \pm_{0.00050}^{0.00056}$  AU every  $\sim 1.09$  days (Hebb et al. 2009, Collins et al. 2017). With a mass of  $1.47 \pm_{0.069}^{0.076} M_J$  and radius of  $1.90 \pm_{0.055}^{0.057} R_J$ , its density of  $0.0266 \pm_{0.014}^{0.015}$  g cm<sup>-3</sup> is less than a quarter of Jupiter’s density of 1.326 g/cm<sup>3</sup>. Due to its extreme equilibrium temperature ( $>2500$  K), it is expected to be in thermochemical equilibrium (Moses 2014).

WASP-12b has been the target of numerous observations and analyses since its discovery in 2008. Its secondary eclipse has been observed across the near- and mid-infrared by a variety of instruments, including the Hubble Space Telescope (HST) Wide Field Camera 3 (WFC3), Spitzer Space Telescope Infrared Array Camera (IRAC), Canada-France-Hawaii Telescope (CFHT) Wideband Imaging Camera (WIC), Michigan-Dartmouth-MIT (MDM) Observatory TIFKAM, and Apache Point Observatory (APO) Near-Infrared Camera (López-Morales et al. 2010, Campo et al. 2011, Croll et al. 2011, Zhao et al. 2012, Cowan et al. 2012, Crossfield et al. 2012, Swain et al. 2013, Föhning et al. 2013). Combinations of these data have been used for retrievals of the dayside  $T(p)$  pressure–temperature profile and molecular abundances by Madhusudhan et al. (2011), Line et al. (2014), Stevenson et al. (2014), and Oreshenko et al. (2017) to investigate the atmospheric properties of this highly irradiated hot Jupiter. Madhusudhan et al. (2011) uses less data than the others due to the limited data at the time of publication. Further, that retrieval occurred before the discovery of WASP-12’s binary M-dwarf companions, which reduces the measured eclipse depths (Bergfors et al. 2011, Bechter et al. 2014). As a result, our paper does not thoroughly compare those results to other investigations.

The results of the Line et al. (2014), Stevenson et al. (2014), and Oreshenko et al. (2017) retrievals are inconsistent in some respects. When considering CO, CO<sub>2</sub>, CH<sub>4</sub>, and H<sub>2</sub>O, Line et al. (2014) do not find evidence for a high C/O due to high abundances of CO<sub>2</sub> and H<sub>2</sub>O, while Stevenson et al. (2014) find a bimodal C/O, both of which have a high abundance of CO<sub>2</sub>. However, these analyses find an abundance of CO<sub>2</sub> that is greater than both CO and H<sub>2</sub>O, which has been shown to be highly improbable in the atmosphere of a planet like WASP-12b (Madhusudhan 2012, Moses et al. 2013, Heng & Lyons 2016). Line et al. (2014) comment that this is implausible and place an upper limit of  $10^{-5}$  on the CO<sub>2</sub> mixing ratio; retrievals under this limit drive the H<sub>2</sub>O mixing ratio over 100 parts per million, resulting in a more realistic CO<sub>2</sub> mixing ratio and maintaining a C/O near solar. Stevenson et al. (2014) also mention the implausibility of their retrieved CO<sub>2</sub> abundance

and propose the addition of HCN and C<sub>2</sub>H<sub>2</sub> into the retrieval model to solve this problem. These species have been shown to exist when C/O > 1 (Madhusudhan 2012, Moses et al. 2013) and have spectral features in Spitzer’s IRAC channel 2; this allows the Bayesian sampler to fit the eclipse depths in that channel using the added species. Consequently, they retrieve an abundance of CO<sub>2</sub> that is less than CO, which is a physically plausible result. This C-rich result is more probable than their O-rich result by a factor of 670. They exclude an isothermal model at 7 $\sigma$  significance.

Oreshenko et al. (2017) performed retrievals using the same data as Stevenson et al. (2014) and expands upon that work by including clouds in their model. They find that the cloud compositions are unconstrained, an expected result considering the degeneracy between cloud composition and gas mixing ratios at low spectral resolutions. When considering CO, CO<sub>2</sub>, CH<sub>4</sub>, and H<sub>2</sub>O, they replicate the results of Line et al. (2014) and Stevenson et al. (2014) of an unrealistically high CO<sub>2</sub> mixing ratio. In general, they find that their retrieval results are prior dominated. When assuming Gaussian priors for the C/H and O/H ratios of WASP-12b matching that of its host star (Teske et al. 2014), the resulting C/O is close to solar. However, when increasing the uncertainties on the values, C/O > 1 becomes the favored solution, consistent with Madhusudhan et al. (2011). Including HCN and C<sub>2</sub>H<sub>2</sub> in the model results in an HCN mixing ratio of 10<sup>-2</sup> – 10<sup>-1</sup>, which they comment is implausible for reasons discussed in Moses et al. (2013); this mixing ratio is over three orders of magnitude greater than that found in Stevenson et al. (2014). An interesting result of their retrieval models is the wide range of possible  $T(p)$  profiles based on the assumed prior, ranging from no to strong inversions. However, despite this wide range of possibilities, the resulting spectra are qualitatively similar.

A curious difference among the retrieved  $T(p)$  profiles of these three investigations is that Line et al. (2014) find the temperature of the atmosphere to be almost entirely above 3000 K with the potential for an inversion, while Stevenson et al. (2014) and Oreshenko et al. (2017) find an upper limit of 3000 K with no inversion. Note that when computing opacities using HITRAN databases, the available data to calculate partition functions has an upper limit of 3000 K (Laraia et al. 2011). The CHIMERA code used by Line et al. (2014) probes temperatures above this limit by assuming that cross sections at temperatures  $> 3000$  K are equal to those at 3000 K (M. Line, priv. comm.). In general, cross sections will differ between temperatures  $> 3000$  K and 3000 K, so this assumption is more likely to give misleading results the greater the deviation from 3000 K. However, extrapolation beyond 3000 K could be even more misleading as there is no guarantee that it will match the true cross sections. This assumption in CHIMERA could lead to a better model fit by allowing the code to explore background emission temperatures greater than 3000 K. Nonetheless, this underscores the need for higher-temperature data to more accurately characterize planets in this temperature regime.

The temperatures found in these retrievals bring attention to another implicit assumption put into these models. Retrieval models consider some set of molecules to fit an observed spectrum. Omitting a molecule that is present in the real object will therefore bias the results: Stevenson et al. (2014) showed that the omission of HCN and  $C_2H_2$  drives up the inferred  $CO_2$  abundance, while including the additional molecules allows for a more realistic fit. At the temperatures retrieved for WASP-12b, both  $H_2$  and  $H_2O$  thermally dissociate, forming H (Arcangeli et al. 2018, Kreidberg et al. 2018, Parmentier et al. 2018). Some H gain an electron from ionized metals, forming  $H^-$ . To date, WASP-12b retrieval models have omitted  $H^-$ , which provides an important continuum opacity source from bound-free and free-free transitions (John 1988), as more thoroughly discussed in Parmentier et al. (2018) and Arcangeli et al. (2018).



In this paper, we perform retrievals using the Bayesian Atmospheric Radiative Transfer code (BART, Harrington et al. 2022, Cubillos et al. 2022, Blečić et al. 2022) matching the setups of Line et al. (2014) and Stevenson et al. (2014) to investigate the discrepancies in their results, and we present a new, physically motivated model that includes additional species not considered in previous investigations. Section 3.3 describes the BART code, and Section 3.4 discusses the setup for each of the nine models. In Section 3.5, we discuss our results in the context of previous analyses of WASP-12b’s dayside atmosphere. Finally, we draw conclusions from our findings in Section 3.6.

### 3.3 BART

Our retrieval code, BART (Harrington et al. 2022, Cubillos et al. 2022, Blečić et al. 2022), pairs the Transit radiative-transfer code (Rojo 2006) with Multi-Core Markov chain Monte Carlo (MCCubed, Cubillos et al. 2017), a Bayesian framework. The user specifies a parameter space to be explored for some model parameters (e.g., the  $T(p)$  profile and molecular abundances). Other inputs include the observational data and its type (e.g., transit or eclipse depths) as well as the instrument filters associated with each data point. For each proposed atmospheric model, the theoretical spectrum is calculated at a high resolution, binned according to the filters, and compared to the observational data. The abundance profiles begin from a user-specified atmosphere (e.g., uniform profiles, or thermochemical equilibrium for a certain  $T(p)$  profile), and MCCubed scales the abundances of the molecules being fit. For credible regions estimated from the posterior, BART computes the steps per effective independent sample (SPEIS) and effective sample size (ESS) to estimate the uncertainty in a given credible region, as detailed in Harrington et al. (2022).

Table 3.1: Summary of Retrieval Models

Characteristic	1	2	3	4	5	6	7	8	9	10	11	12	13
CO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CO <sub>2</sub>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CH <sub>4</sub>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
H <sub>2</sub> O	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
HCN					✓		✓		✓		✓	✓	✓
C <sub>2</sub> H <sub>2</sub>					✓		✓		✓		✓	✓	✓
NH <sub>3</sub>												✓	✓
TiO												✓	✓
H <sup>-</sup>								✓	✓	✓	✓	✓	✓
e <sup>-</sup>								✓	✓	✓	✓	✓	✓
Cross sections > 3000 K	✓	✓	✓			✓		✓	✓	✓	✓	✓	✓
HST WFC3 G141	Sw13 <sup>a</sup>	Sw13	Sw13	St14 <sup>b</sup>	St14	St14	St14	Sw13	Sw13	St14	St14	Sw13	St14
APO ARC z <sup>c</sup>			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CFHT WIC J <sup>d</sup>				✓	✓	✓	✓			✓	✓		✓
CFHT WIC H <sup>d</sup>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CFHT WIC Ks <sup>d</sup>				✓	✓	✓	✓			✓	✓		✓
MDM Hiltner								✓	✓			✓	
TIFKAM Ks <sup>e</sup>	✓	✓	✓										
Subaru MOIRCS NB2315 <sup>f</sup>	✓	✓	✓					✓	✓			✓	
Spitzer IRAC ch1	Co12 <sup>g</sup>	Co12	Co12	St14	St14	St14	St14	Co12	Co12	St14	St14	Co12	St14
Spitzer IRAC ch2	Co12 (n) <sup>i</sup>	Co12 (e) <sup>j</sup>	Co12 (n)	St14	St14	St14	St14	Co12 (n)	Co12 (n)	St14	St14	Co12 (n)	St14
Spitzer IRAC ch3	Ca11 <sup>h</sup>	Ca11	Ca11	St14	St14	St14	St14	Ca11	Ca11	St14	St14	Ca11	St14
Spitzer IRAC ch4	Ca11	Ca11	Ca11	St14	St14	St14	St14	Ca11	Ca11	St14	St14	Ca11	St14

a: Swain et al. (2013); b: Stevenson et al. (2014); c: López-Morales et al. (2010), corrected by Crossfield et al. (2012); d: Croll et al. (2011), corrected by Crossfield et al. (2012); e: Zhao et al. (2012), corrected by Crossfield et al. (2012); f: Crossfield et al. (2012); g: Cowan et al. (2012), corrected by Crossfield et al. (2012); h: Campo et al. (2011), corrected by Crossfield et al. (2012); i: ‘null’ hypothesis; j: ‘ellipsoidal’ hypothesis

### 3.4 Model Configurations

To investigate the previously published retrievals of WASP-12b, we replicate their setups to the best of BART's ability, and we expand upon those setups to delve deeper into the nature of the discrepancies in results. Our nine retrieval models are that of

1. Line et al. (2014) *null* case,
2. Line et al. (2014) *ellipsoidal* case,
3. Line et al. (2014) *null* case plus the data from López-Morales et al. (2010) corrected by Crossfield et al. (2012),
4. Stevenson et al. (2014) case without HCN and C<sub>2</sub>H<sub>2</sub>,
5. Stevenson et al. (2014) case with HCN and C<sub>2</sub>H<sub>2</sub>,
6. Stevenson et al. (2014) without HCN and C<sub>2</sub>H<sub>2</sub> with the assumptions of CHIMERA about cross sections,
7. Stevenson et al. (2014) with HCN and C<sub>2</sub>H<sub>2</sub> mixing ratios fixed to their reported C-rich best-fit values,
8. Model 3, with H<sup>-</sup>,
9. Model 8, with HCN and C<sub>2</sub>H<sub>2</sub>,
10. Model 6, with H<sup>-</sup>,
11. Model 10, with HCN and C<sub>2</sub>H<sub>2</sub>,
12. Model 3, with NH<sub>3</sub>, HCN, C<sub>2</sub>H<sub>2</sub>, TiO, and H<sup>-</sup>, and

### 13. Model 6, with the molecules of Model 12.

At the time of writing, BART does not have a realistic cloud model, so we do not try to replicate the Oreshenko et al. (2017) result that finds cloud composition to be unconstrained. Table 3.1 summarizes the setup of each model regarding data sources and molecules besides H, H<sub>2</sub>, and He.

For this investigation, BART’s only restrictions on the atmospheric models are 1) the sum of molecular abundances must equal 1 for each layer, and 2) the ratio of H<sub>2</sub> to He is held constant by adjusting their abundances to satisfy condition 1. For models that do not use CHIMERA’s assumption about cross sections at  $T > 3000$  K, BART also enforces that the temperature of the atmosphere must remain within the line-list limits.

The atmospheric models consist of 100 log-spaced layers spanning  $10^{-8}$  – 100 bar. For radiative-transfer calculations, only layers above where the optical depth reaches  $\geq 10$  are considered. We assume uniform abundances for the species present, consistent with the previous publications. Each model has a free parameter for the abundance of each opacity-contributing species, as well as five free parameters for the  $T(p)$  profile (the Planck mean infrared opacity, the ratios of the Planck mean visible and infrared opacities for two streams, the partition between the two streams, and a general parameter for albedo/emissivity/energy recirculation; Line et al. 2013). The free parameters for the partition between streams and albedo/emissivity/recirculation have uniform priors; all other free parameters have log-uniform priors. We do not vary the H<sup>-</sup> or e<sup>-</sup> abundances because previous publications indicate the atmosphere is nearly isothermal at  $\sim 3000$  K in the regions with sensitivity, and the abundances change by only a factor of  $\sim 2$  for a change as large as 200 K. H<sup>-</sup> and e<sup>-</sup> are fixed to an abundance of  $10^{-9}$  and  $10^{-6}$ , respectively, which is roughly consistent with thermochemical equilibrium at 3000 K (NASA Chemical Equilibrium with Applications code Gordon & McBride 1994) at pressures probed by the observations. A wide range of values are allowed for each free parameter without consideration of physical plausibility. For this investigation,

we use the DEMCzs sampling algorithm of ter Braak & Vrugt (2008) because we found that the DEMC algorithm of ter Braak (2006) occasionally leads to rogue chains that do not converge. Since DEMCzs only considers the goodness of fit of each proposed model, it can explore both realistic and unrealistic solutions. The initial samples of parameters for the DEMCzs algorithm are drawn randomly from a uniform distribution; most of the parameters are the logarithm of the true parameter, so those true parameters are randomly sampled from a log-uniform space.

We include HITEMP opacities for CO, CO<sub>2</sub>, and H<sub>2</sub>O (Rothman et al. 2010), and HITRAN opacities for NH<sub>3</sub>, C<sub>2</sub>H<sub>2</sub>, and HCN (Rothman et al. 2013). Models 1 – 11 use the Rothman et al. (2013) CH<sub>4</sub> line list for consistency with Line et al. (2014). while models 12 and 13 use the new CH<sub>4</sub> HITEMP line list (Hargreaves et al. 2020). TiO opacities are sourced from Schwenke (1998). We include H<sub>2</sub>-H<sub>2</sub> and H<sub>2</sub>-He collision-induced absorptions (Richard et al. 2012) as well as H<sup>-</sup> bound-free and free-free absorption (John 1988), where appropriate.

Note that Equation 3 of John (1988) does not lead to the correct bound-free opacity values necessary to reproduce Table 1 of John (1988); a factor of 10 for the bound-free cross-sections is required to obtain agreement. We refer the reader to our compendium for a detailed proof. It is unclear whether the given equations or table provide the correct opacities. Considering the numerous fitted constants used in the equations, we have chosen to assume that the table is correct, as it appears to lead to agreement with the H<sup>-</sup> opacities plotted in Figure 1 of Arcangeli et al. (2018) and Figure 4 of Parmentier et al. (2018). Additionally, by assuming the greater of the two possibilities, we can assess an upper limit on the impact of H<sup>-</sup> when retrieving WASP-12b’s atmospheric properties.

### 3.5 Results and Discussion

The accepted  $T(p)$  profiles with  $1\sigma$  and  $2\sigma$  regions, normalized contribution functions, best-fit spectrum, and 1D marginalized posteriors are shown for Models 12 and 13 in Figure 3.1 and 3.2, respectively. The electronic supplement provides additional figures: 2D pairwise posteriors and trace plots for Models 12 and 13, as well as the corresponding set of six plots for Models 1 – 11. Table 3.3 contains the best-fit values and 68.27% interval for the retrieved log abundances of each molecule for all 13 models, the best-fit values and 68.27% interval reported by Line et al. (2014) for the ‘null’ and ‘ellipsoidal’ cases, the best-fit values reported by Stevenson et al. (2014) for the C-rich and O-rich cases, and the lower/upper limits on abundances shown in the top of Figure 3 of Oreshenko et al. (2017). We also report the best-fit and 68.27% credible region for C/O for each case. We have excluded extreme outliers ( $C/O > 1000$ ) from the density estimation, as they represent a small percentage of the total models and cause problems for the density estimation algorithm, and we do not consider HCN or  $C_2H_2$  when calculating C/O due to the lack of evidence to support their inclusion in the model. Values of “...” indicates that the model does not contain that molecule. Table 3.2 lists the SPEIS, ESS, and associated uncertainty in the 68.27% credible region for each model considered. We also provide a compendium with the data and commands necessary to reproduce this work; the link is at the end of the text.

Table 3.2: Summary of Retrieved Credible Region Uncertainties

Model	SPEIS <sup>a</sup>	ESS <sup>b</sup>	68.27% Region Uncertainty (%) <sup>c</sup>
1	1834	517	2.04
2	2276	417	2.27
3	1025	926	1.53
4	953	996	1.47
5	4871	195	3.31
6	1948	487	2.10
7	3463	274	2.80
8	5483	455	2.17
9	6135	407	2.30
10	4673	534	2.01
11	7562	330	2.55
12	3356	327	2.56
13	3359	327	2.56

a: Steps per effective independent sample; the number of iterations needed for a non-correlated sample.

b: Effective sample size; the number of independent samples.

c: For finite ESS, the approximation to the true posterior induces an uncertainty in any determined credible region. See Harrington et al. (2022) for details.

Table 3.3: Retrieved Molecular Log Abundances

Model	CO	CO <sub>2</sub>	CH <sub>4</sub>	H <sub>2</sub> O	HCN	C <sub>2</sub> H <sub>2</sub>	NH <sub>3</sub>	TiO	C/O
1	-4.4 [-10.0, -2.3]	-2.6 [-11.0, -2.2]	-11.4 [-10.5, -2.9]	-5.4 [-4.7, -1.2]	...	...	...	...	0.50 [0.00, 4.81]
2	-11.3 [-10.9, -4.7]	-5.2 [-5.6, -3.2]	-2.8 [-9.7, -2.0]	-11.1 [-11.2, -6.6]	...	...	...	...	115 [0.00, 9.53]
3	-2.0 [-9.6, -2.1]	-1.0 [-10.8, -1.4]	-10.1 [-11.2, -4.2]	-6.6 [-5.5, -1.2]	...	...	...	...	0.52 [0.00, 4.84]
4	-10.4 [-13.2, -5.7]	-5.8 [-6.2, -5.0]	-3.8 [-6.6, -2.8]	-8.6 < -7.7	...	...	...	...	52.8 [0.06, 9.86]
5	-6.8 [-10.6, -4.8]	-5.5 [-6.1, -4.9]	-3.5 [-6.1, -2.6]	-8.3 [-10.6, -6.7]	-7.0 [-10.6, -1.4]	-7.1 [-10.0, -4.1]	...	...	52.2 [0.00, 12.4]
6	-5.9 [-10.9, -4.6]	-5.2 [-5.5, -4.0]	-3.2 [-5.6, -1.6]	-8.4 [-11.1, -6.6]	...	...	...	...	45.2 [0.10, 12.6]
7	-7.8 [-11.3, -5.5]	-5.5 [-6.1, -4.9]	-3.5 [-7.3, -2.7]	-7.2 [-10.8, -6.9]	-6 fixed	-5 fixed	...	...	50.5 [0.04, 23.7]
8	-2.3 [-8.2, -1.3]	-1.0 [-10.0, -1.6]	-5.7 [-10.5, -3.4]	-7.3 [-6.8, -1.0]	...	...	...	...	0.51 [0.00, 5.72]
9	-2.2 [-10.3, -1.5]	-1.7 [-10.1, -1.3]	-9.4 [-10.2, -2.5]	-8.4 [-6.6, -1.1]	-8.5 [-12.1, -2.4]	-2.8 [-11.9, -3.6]	...	...	0.57 [0.00, 5.64]
10	-6.7 [-10.5, -4.9]	-5.5 [-6.1, -4.5]	-3.7 [-5.7, -1.9]	-10.0 [-10.7, -6.5]	...	...	...	...	37.3 [0.05, 21.0]
11	-8.4 [-10.8, -5.2]	-5.9 [-6.0, -4.4]	-3.9 [-6.3, -1.9]	-10.9 [-10.4, -6.6]	-10.5 [-11.6, -3.5]	-6.3 [-12.6, -5.3]	...	...	45.9 [0.05, 20.0]
12	-9.5 [-8.3, -2.1]	-2.1 [-9.6, -2.3]	-1.6 [-10.5, -2.5]	-8.0 [-4.5, -1.2]	-9.4 [-11.8, -3.7]	-12.2 [-11.0, -4.9]	-11.5 [-10.6, -4.9]	-8.1 [-10.0, -5.1]	1.94 [0.00, 5.51]
13	-9.1 [-10.2, -4.8]	-5.2 [-5.8, -4.2]	-4.4 [-6.8, -3.2]	-8.1 [-10.0, -6.0]	-5.4 [-11.2, -3.2]	-6.0 [-12.0, -4.8]	-8.2 [-12.4, -6.4]	-9.8 [-12.3, -8.4]	3.39 [0.31, 1.43]
L14 <sup>a</sup> , null	-2.7 [-9.7, -2.0]	-1.2 [-8.1, -1.3]	-8.6 [-10.8, -3.1]	-8.1 [-8.8, -2.3]	...	...	...	...	0.51 [0.30, 1.00]



Model	CO	CO <sub>2</sub>	CH <sub>4</sub>	H <sub>2</sub> O	HCN	C <sub>2</sub> H <sub>2</sub>	NH <sub>3</sub>	TiO	C/O
L14, ellipsoidal	-2.7 [-10.3, -2.8]	-1.0 [-1.3, -0.8]	-9.7 [-10.6, -5.1]	-3.3 [-10.0, -3.1]	... ...	... ...	... ...	... ...	0.50 [0.11, 0.22]
St14, C-rich	-3.5	-6.0	-4.1	-6.6	-6.0	-5.0	...	...	1.30
St14, O-rich	-3.3	-4.2	-7.0	-3.3	-7.0	-9.8	...	...	0.50
O17 <sup>b</sup>	[-12, -2.5]	[-7, -4]	[-12, -6]	[-12, -6]	[-3, -1]	[-12, -4]	...	...	[0.3, 4]

**Notes:** Models 1 – 13 and the Line et al. (2014) results are given as the best-fit values, followed by the 68.27% credible regions. Stevenson et al. (2014) results are given as the reported best-fit values. Oreshenko et al. (2017) results are given as the estimated minimum and maximum of the posteriors from their Figure 3.

a: Line et al. (2014); b: Oreshenko et al. (2017)

For Model 1 (Figure 3.1, figure set panel 1.1), BART’s results generally agree with that of the ‘null’ case of Line et al. (2014). Our best-fit abundances have  $\text{CO}_2 > \text{CO}$ , as Line et al. (2014) find, which, as mentioned previously here and in their paper, is implausible. However, there is large uncertainty in this result, as indicated by the nearly flat posterior of both CO and  $\text{CO}_2$ . Like Line et al. (2014), we find that CO and  $\text{CH}_4$  are unconstrained, according to the flat posteriors, but we also find that  $\text{CO}_2$  is unconstrained. Except for  $\text{CH}_4$ , the best-fit values fall within the 68.27% region reported by Line et al. (2014). The  $T(p)$  profile  $1\sigma$  regions overlap, with both best-fit profiles favoring an inversion.

For Model 2 (Figure 3.1, figure set panel 1.2), BART’s results differ in some respects from the ‘ellipsoidal’ case of Line et al. (2014). Most notably, the  $T(p)$  profile  $1\sigma$  region shows a non-inverted atmosphere where the upper atmosphere is  $< 3000$  K, whereas Line et al. (2014) find an inverted atmosphere with the upper atmosphere  $> 3000$  K. However, the normalized contribution functions indicate minimal sensitivity below a pressure of  $10^{-6}$  bar; in the best-constrained region (0.1 – 10 bar), the retrieved  $T(p)$  profiles agree. For abundances, the best-fit values disagree, but, except for  $\text{CO}_2$ , the 68.27% intervals overlap. However, the case of Line et al. (2014) where  $\text{CO}_2$  has an upper limit agrees more closely with our retrieved interval. It is unclear why BART does not find the high- $\text{CO}_2$  mode found by Line et al. (2014).

The results for Model 3 (Figure 3.1, figure set panel 1.3) in general agree with Model 1. CO,  $\text{CO}_2$ , and  $\text{CH}_4$  are similarly unconstrained, while the 68.27% region for  $\text{H}_2\text{O}$  overlaps (Table 3.3). A notable difference is that Model 3 finds a lower minimum for that region. This is likely from the additional data point of López-Morales et al. (2010) providing an additional constraint on the background emission.

BART's results for Model 4 (Figure 3.1, figure set panel 1.4) in many ways match that of Model 2. The  $T(p)$  profile  $1\sigma$  regions closely match, and the marginalized posteriors exhibit many similarities. Similar to Model 2, BART does not find the unrealistically high  $\text{CO}_2$  abundance reported by Stevenson et al. (2014) when retrieving with this setup (Table 3.3). Since the parameter space for  $\text{CO}_2$  extended to a log mixing ratio of -1, these models must have been discarded by BART's sampler. It is uncertain whether this difference can be attributed to the sampling algorithm, the opacity sources, or some other difference in the retrieval algorithm. Nevertheless, the best-fit  $T(p)$  profiles of Stevenson et al. (2014) fall within BART's reported  $1\sigma$  region where there is sensitivity. Further, their C-rich best-fit values are within BART's 68.27% regions for all but CO. Our 68.27% interval for  $\text{H}_2\text{O}$  rules out their O-rich result, which features an  $\text{H}_2\text{O}$  abundance of  $10^{-3.3}$ .

The results for Model 5 (Figure 3.1, figure set panel 1.5) mostly agree with the findings of Stevenson et al. (2014). The C-rich best-fit values for  $\text{CO}_2$ ,  $\text{CH}_4$ ,  $\text{H}_2\text{O}$ , and  $\text{C}_2\text{H}_2$  reported by them fall within BART's 68.27% regions (Table 3.3). Our 68.27% interval for  $\text{H}_2\text{O}$  similarly rules out their O-rich result. BART finds HCN to be unconstrained. The retrieved  $T(p)$  profiles closely match the results of Model 4 and are therefore consistent with the retrieved  $T(p)$  profiles of Stevenson et al. (2014). The best-fit values and 68.27% region found for Model 5 closely matches the results of Model 4, which does not include HCN or  $\text{C}_2\text{H}_2$ . Thus, the inclusion of these additional molecules only minorly affects the retrieved result.

Model 6 (Figure 3.1, figure set panel 1.6) examines the effect of allowing  $T(p)$  profiles with  $T > 3000$  K on Model 4. BART's results closely match those of Models 4 and 5. The  $T(p)$  profile  $1\sigma$  regions generally agree, with a general upper limit of  $\sim 3000$  K. The marginalized posteriors are similar, and the 68.27% regions for the molecular abundances closely overlap. This importantly demonstrates that the eclipse data, not the model assumptions, are driving the result.

For Model 7 (Figure 3.1, figure set panel 1.7), BART’s results generally agree with those of Model 5. Even with the HCN and C<sub>2</sub>H<sub>2</sub> abundances fixed to the C-rich best-fit values reported by Stevenson et al. (2014), BART does not favor their reported CO abundance. We suspect that this difference can be attributed to differences in line lists, though we were unable to ascertain the line lists used by Stevenson et al. (2014) via private communication to test this hypothesis.

Models 8, 9, and 12 (Figure 3.1, figure set panels 1.8, 1.9, and 1.12), which are equivalent to Model 3 but with additional opacity sources, bear qualitatively similar results to Model 3. While the 68% regions may differ slightly, these are largely due to numerical differences in the density estimation of the marginalized posteriors. Visually, the posteriors are quite similar; CO, CO<sub>2</sub>, and CH<sub>4</sub> have flat marginalized posteriors, while H<sub>2</sub>O has a preference for log abundances greater than -6. In Model 12, the flat posteriors of HCN, C<sub>2</sub>H<sub>2</sub>, and NH<sub>3</sub> demonstrate that the data are insensitive to their inclusion. While the log abundance TiO is loosely constrained to be  $-7.5 \pm 2.5$ , consistent with thermochemical equilibrium for the retrieved  $T(p)$  profile in the regions probed by the observations, the posterior allows for TiO to be absent from the atmosphere within  $2\sigma$ .

Similarly, Models 10, 11, and 13 (Figure 3.2; see also figure set panels 1.10, 1.11, and 1.13), which are equivalent to Model 6 except with additional opacity sources, yield results similar to Model 6. The resulting thermal profiles agree closely. As with Models 8, 9, and 12, the 68% regions differ slightly due to the density estimation algorithm, though the marginalized posteriors are qualitatively similar to Model 6. That is, CO tends to favor smaller values, CO<sub>2</sub> is tightly constrained within an order of magnitude from  $10^{-5}$ , CH<sub>4</sub> features a constraint about  $10^{-3}$  but could also be absent from the atmosphere (the posterior has a non-negligible tail), and H<sub>2</sub>O has an upper limit of around  $10^{-4}$ . For Model 13, BART finds upper limits of  $10^{-2}$  for NH<sub>3</sub> and  $10^{-6}$  for TiO.

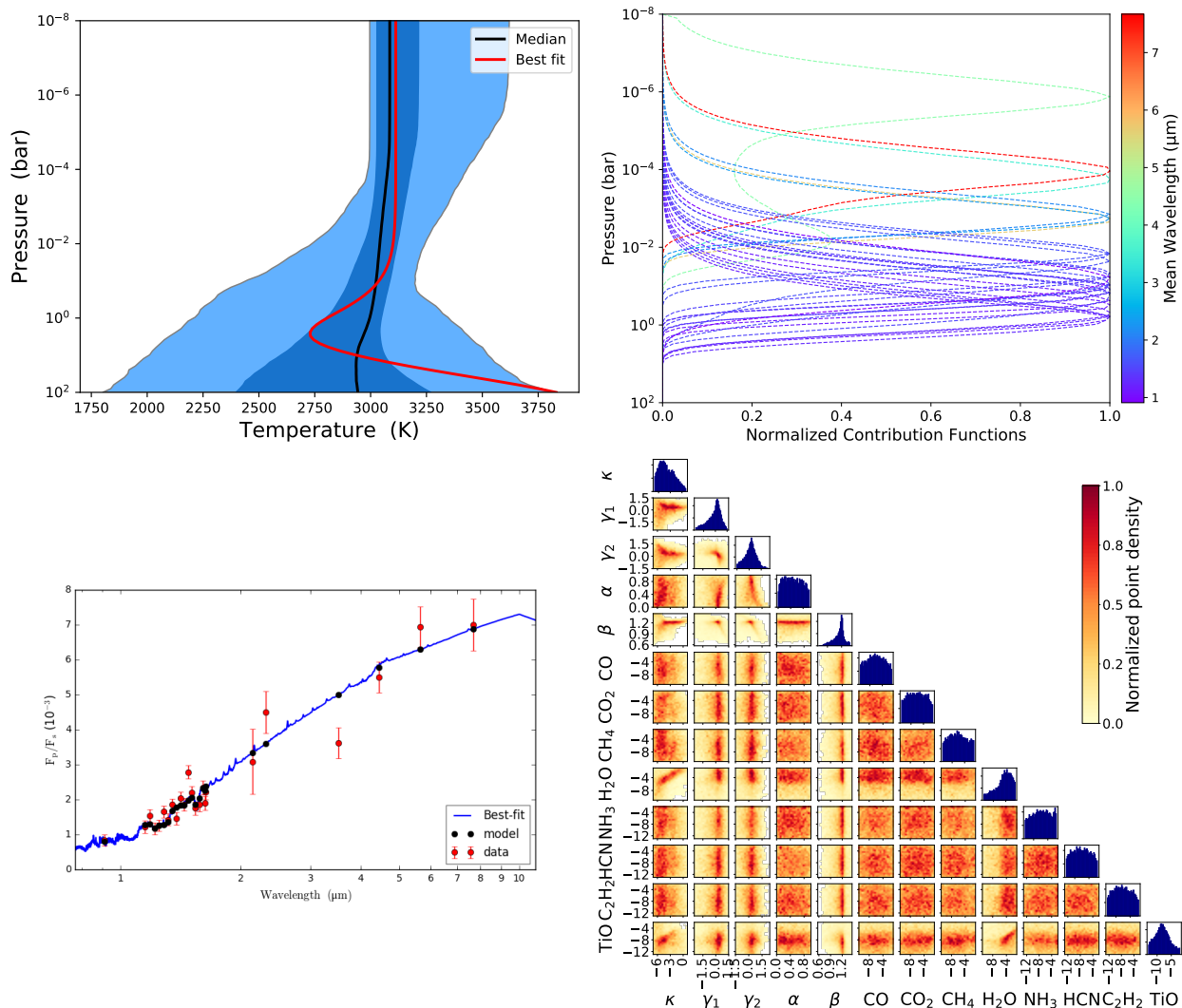


Figure 3.1: BART results for Model 12, which uses the data set of Line et al. (2014) with the additional data point from López-Morales et al. (2010). *Top left*:  $T(p)$  profiles explored by the Markov chain Monte Carlo (MCMC). Red line denotes the best-fit  $T(p)$  profile, the black line denotes the median  $T(p)$  profile, and the dark and light blue regions indicate the  $1\sigma$  and  $2\sigma$  regions, respectively. *Top right*: normalized contribution functions. *Bottom left*: best-fit spectrum. *Bottom right*: marginalized posteriors. BART favors a mostly isothermal  $T(p)$  profile around 3000 K. Of the molecules considered, only  $\text{H}_2\text{O}$  is constrained (log abundance  $> -6$ ). While the log abundance of TiO features a bump near  $-8$ , the non-negligible tail at lesser abundances indicates the possibility that TiO is not present in the atmosphere. (The complete figure set (13 images) is available.)

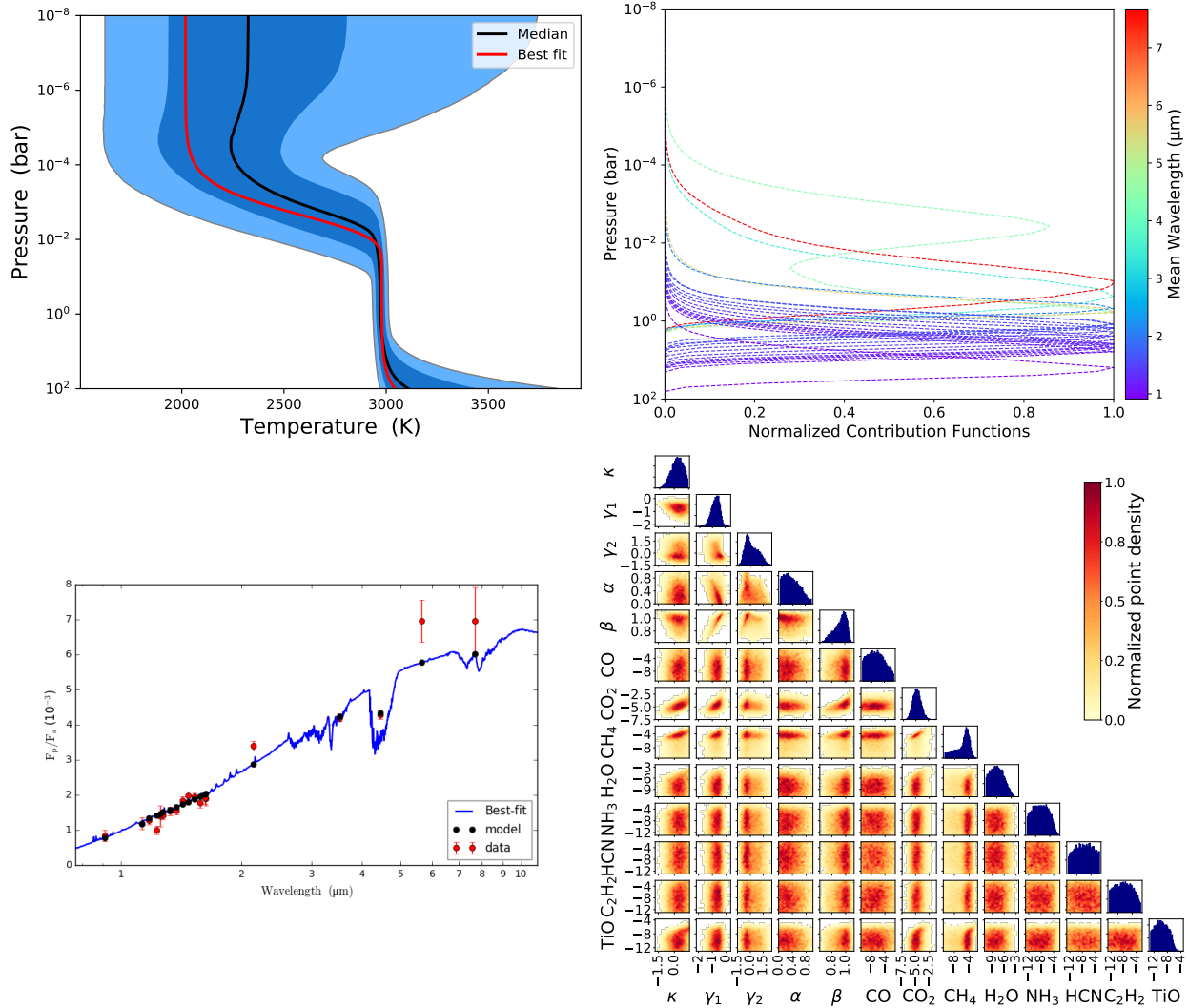


Figure 3.2: Same as Figure 3.1, except for Model 13, which uses the data set of Stevenson et al. (2014). Like Model 12, BART favors a nearly isothermal  $T(p)$  profile in the regions probed by the observations. Contrary to Model 12,  $\text{H}_2\text{O}$  features an upper limit around  $-4$  for its log abundance,  $\text{CO}_2$  is constrained to a log abundance of  $-7 - -3$ ,  $\text{CH}_4$  shows evidence of a log abundance around  $-5 - -4$  (though features a non-negligible tail at lesser abundances, indicating it may not be present in the atmosphere), and  $\text{CO}$  favors log abundances around  $-10 - -6$ . Like Model 12, the posterior for  $\text{TiO}$  indicates an upper limit on the log abundance of around  $-5 - -4$ , and there is similarly no evidence for  $\text{HCN}$  or  $\text{C}_2\text{H}_2$ .

As Models 8 and 10, 9 and 11, and 12 and 13 bear identical setups/assumptions aside from the eclipse depths, the differences between them are thus solely attributable to the data. Both data sets favor a  $\sim 3000$  K atmosphere in the regions probed by the observations, as evidenced by the normalized contribution functions. However, their retrieved abundances, and by extension the inferred  $C/O$ , are incompatible. The data set of Models 8, 9, and 12 yields evidence of a high water abundance and possibly TiO in thermochemical equilibrium, with no evidence of other molecules, while the data set of Models 10, 11, and 13 constrain  $CO_2$  and  $CH_4$ , with upper limits for CO and  $H_2O$ . None of the data sets considered offer meaningful constraints on HCN,  $C_2H_2$ , or  $NH_3$ . Despite the theoretical expectation that  $H^-$  plays an important role in the atmosphere of hot Jupiters like WASP-12b (Parmentier et al. 2018), its inclusion does not make a significant difference in the retrieval results (Figure 3.3).

Our results are broadly consistent with those of Oreshenko et al. (2017), except for HCN (Table 3.3). In general, the models using the Line et al. (2014) data favor an atmosphere that is isothermal or has an inversion, whereas the models using the Stevenson et al. (2014) data favor an atmosphere that has no inversion. Lothringer et al. (2018) show that thermal inversions are likely for this planet class even without VO or TiO. While the results using the Line et al. (2014) ‘null’ data set are consistent with this finding, it is important to consider that these results also favor a high  $H_2O$  abundance, which is inconsistent with the expected thermal dissociation of  $H_2O$ . While the Line et al. (2014) ‘ellipsoidal’ case favors a low  $H_2O$  abundance, it does not show strong evidence of a thermal inversion and more closely matches the results of Stevenson et al. (2014).

Over the reported 68.27% credible regions, the  $C/O$  for these models can take a wide range of values. However, they do not indicate that  $C/O \gg 1$ . Rather, it is due to a combination of reasons. For one, the DEMCzs sampler is free to explore the parameter space without regard for  $C/O$ . For example, in the case of Model 4, the best-fit  $C/O$  value is 52.8, which is noticeably different than the 68.27% region; this is a product of a high best-fit abundance of  $CH_4$  and low best-fit abundance

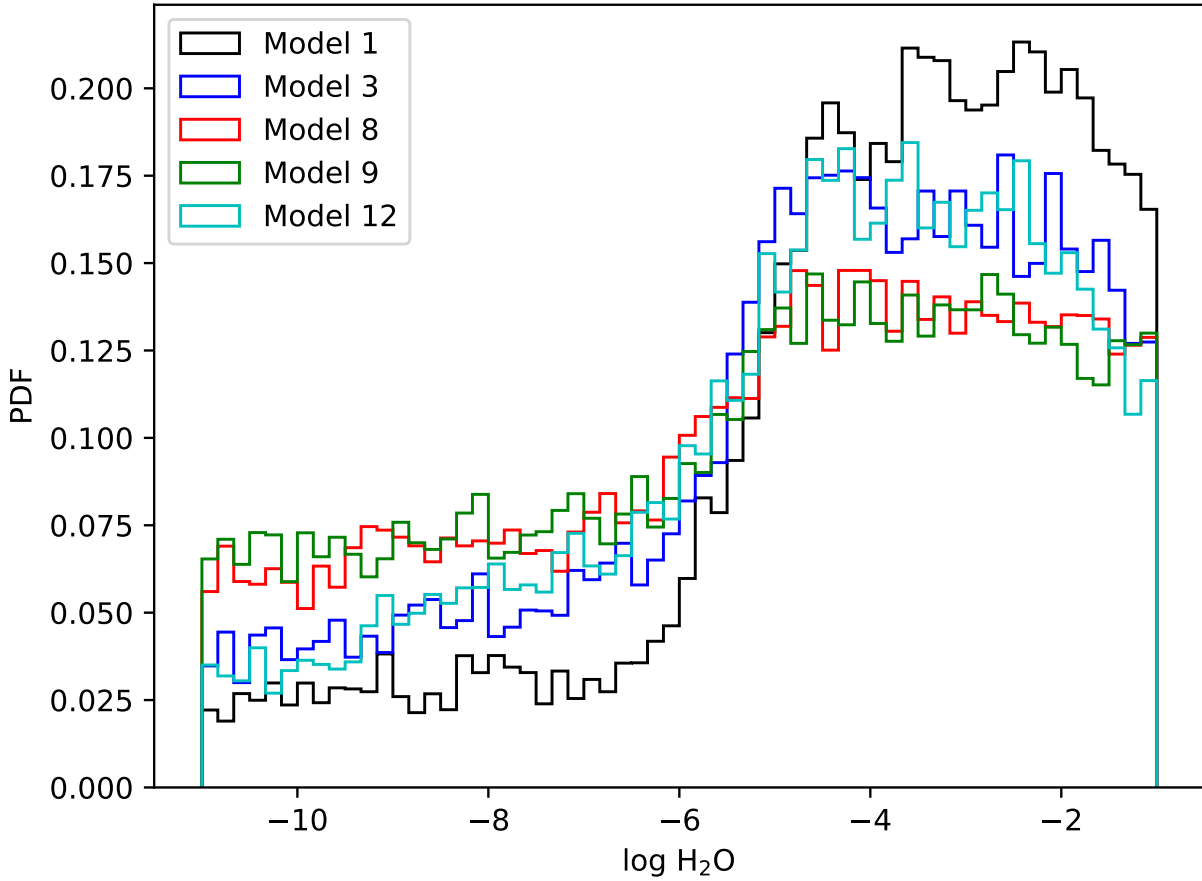


Figure 3.3: Comparison of the retrieved  $\text{H}_2\text{O}$  marginalized posteriors for the models using the Line et al. (2014) ‘null’ data set. The lack of significant variation between the models without  $\text{H}^-$  (1, 3) and those with  $\text{H}^-$  (8, 9, 12) indicates that the inclusion of  $\text{H}^-$  does not significantly affect the retrieved  $\text{H}_2\text{O}$  abundance.



of  $\text{H}_2\text{O}$ , which provides a better statistical fit than models with more reasonable C/O values. Additionally, thermal dissociation of  $\text{H}_2\text{O}$  would lead to the formation of O (Parmentier et al. 2018), causing an apparent increase in the measurable C/O but not the true C/O. There are also other oxygen-bearing species not considered here, particularly condensates at the limb (Wakeford et al. 2017), that would contribute to C/O, if included in the model. However, Oreshenko et al. (2017) demonstrated that the cloud composition is degenerate with gas mixing ratios; higher-resolution data is necessary to explore a model with various species of condensates.

Typically, when considering multiple models, the Bayes factor of each model is compared to choose the ‘best’ model. In this investigation, however, this would be erroneous: the 13 models presented do not use the same data sets, and they are not competing for the ‘best’ model. Rather, the models demonstrate that the retrieval results are data driven and independent of the model selected. Consequently, we do not compute the Bayes factor as it would be a misleading metric. We emphasize that the results show that the previous retrieval analyses of Line et al. (2014) and Stevenson et al. (2014) are consistent, when considering the data set used in each investigation. Follow-up observations are required to determine which data set, if either, represents the true nature of WASP-12b.

### 3.6 Conclusions

In general, we are able to reproduce the published results of Line et al. (2014), Stevenson et al. (2014), and Oreshenko et al. (2017) using BART. We confirm the finding of Stevenson et al. (2014) that excludes an isothermal profile when mimicking their setup. By following the model assumption of Line et al. (2014) allowing temperatures above 3000 K with the Stevenson et al. (2014) data, the range of possible  $T(p)$  profiles expands to include inverted profiles but still favors a non-inverted atmosphere. Note that an inverted profile is expected for ultra-hot Jupiters like WASP-12b (Lothringer et al. 2018).

We find that current data does not support the inclusion of HCN, C<sub>2</sub>H<sub>2</sub>, H<sup>-</sup>, e<sup>-</sup>, NH<sub>3</sub>, and TiO, as they do not significantly affect the posterior. As new telescopes in the near future provide higher quality data, these molecules should be reconsidered, as in this investigation, to determine if they inform the results.

Some aspects were unable to be reproduced, namely, the unrealistically high CO<sub>2</sub> abundances reported in the Line et al. (2014) ‘ellipsoidal’ case and the Stevenson et al. (2014) case without HCN or C<sub>2</sub>H<sub>2</sub>, and the unrealistically high HCN abundance reported by Oreshenko et al. (2017). While we suspect this discrepancy is due to differences in the retrieval model, further investigation is necessary to definitively determine the origin.

We have demonstrated that differences in eclipse depth data sets primarily drive the differences between the results of Line et al. (2014) and Stevenson et al. (2014), with more subtle differences likely attributable to the retrieval model and input data sources (e.g., line lists). Many of the eclipse depths come from the same set of observations but are analyzed using different reduction pipelines. Our study shows that subtle differences in the reduction pipelines used (e.g., the binning of WFC3 spectra into discrete channels) can drive radically different results. This emphasizes the need for standard data sets to be used for benchmarking photometry and spectroscopy reduction pipelines.

The conflicting results of previous publications highlight the complexities of retrieval modeling and the importance of clearly communicating model assumptions. This will be especially important as retrieval models become more sophisticated with the introduction of more complicated techniques such as 3D modeling and machine learning (Márquez-Neila et al. 2018, Zingales & Waldmann 2018, Waldmann & Griffith 2019, Cobb et al. 2019). Transparency allows for published analyses to be easily reproduced by others and encourages quicker resolution of conflicting results, which can lead to better work in the field.

Future retrieval studies should use multiple binnings of the same data to explore whether the retrievals are consistent across different binnings of the data, including the unbinned data. This requires that data analyses publish unbinned spectra to support future retrieval studies. Retrieval results dependent on the binning, such as those shown here, indicate the need for higher quality spectroscopic data. A comprehensive retrieval analysis of WASP-12b with additional data from future flagship observatories, such as the James Webb Space Telescope, will provide deeper insight into the nature of this extreme exoplanet.

The Reproducible Research Compendium for this work is available for download<sup>1</sup>.

---

<sup>1</sup>Available at <https://doi.org/10.5281/zenodo.5777204>.  
Note: the compendium is 16.5 GB compressed and 34 GB uncompressed.

### 3.7 Acknowledgments

We thank Michael Line and Julianne Moses for helpful discussions during the preparation of this manuscript. We also thank the anonymous referee for valuable comments that improved the quality of this manuscript. We thank contributors to SciPy, Matplotlib, and the Python programming language, the free and open-source community, and the NASA Astrophysics Data System for software and services. Part of this work is based on observations made with the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA. This work was supported by NASA Planetary Atmospheres grant NNX12AI69G, NASA Astrophysics Data Analysis Program grant NNX13AF38G, and NASA Fellowship Activity under NASA Grant 80NSSC20K0682.

### 3.8 List of References

- Arcangeli, J., Désert, J.-M., Line, M. R., Bean, J. L., Parmentier, V., Stevenson, K. B., Kreidberg, L., Fortney, J. J., Mansfield, M., & Showman, A. P. 2018, *ApJ*, 855, L30
- Batalha, N. M. 2014, *Proceedings of the National Academy of Sciences*, 111, 12647
- Bechter, E. B., Crepp, J. R., Ngo, H., Knutson, H. A., Batygin, K., Hinkley, S., Muirhead, P. S., Johnson, J. A., Howard, A. W., Montet, B. T., Matthews, C. T., & Morton, T. D. 2014, *ApJ*, 788, 2
- Bergfors, C., Brandner, W., Henning, T., & Daemgen, S. 2011, in *IAU Symposium*, Vol. 276, *The Astrophysics of Planetary Systems: Formation, Structure, and Dynamical Evolution*, ed. A. Sozzetti, M. G. Lattanzi, & A. P. Boss, 397–398

- Blecic, J., Harrington, J., Cubillos, P. E., Bowman, M. O., Rojo, P. M., Stemm, M., Challener, R. C., Himes, M. D., Foster, A. J., Dobbs-Dixon, I., Foster, A. S. D., Lust, N. B., Blumenthal, S. D., Bruce, D., & Loredó, T. J. 2022, *The Planetary Science Journal*, 3, 82
- Campo, C. J., Harrington, J., Hardy, R. A., Stevenson, K. B., Nymeyer, S., Ragozzine, D., Lust, N. B., Anderson, D. R., Collier-Cameron, A., Blecic, J., Britt, C. B. T., Bowman, W. C., Wheatley, P. J., Loredó, T. J., Deming, D., Hebb, L., Hellier, C., Maxted, P. F. L., Pollaco, D., & West, R. G. 2011, *ApJ*, 727, 125
- Cobb, A. D., Himes, M. D., Soboczenski, F., Zorzan, S., O'Beirne, M. D., Güneş Baydin, A., Gal, Y., Domagal-Goldman, S. D., Arney, G. N., & Angerhausen, D. 2019, *AJ*, 158, 33
- Collins, K. A., Kielkopf, J. F., & Stassun, K. G. 2017, *AJ*, 153, 78
- Cowan, N. B., Machalek, P., Croll, B., Shekhtman, L. M., Burrows, A., Deming, D., Greene, T., & Hora, J. L. 2012, *ApJ*, 747, 82
- Croll, B., Lafreniere, D., Albert, L., Jayawardhana, R., Fortney, J. J., & Murray, N. 2011, *AJ*, 141, 30
- Crossfield, I. J. M., Barman, T., Hansen, B. M. S., Tanaka, I., & Kodama, T. 2012, *ApJ*, 760, 140
- Cubillos, P., Harrington, J., Loredó, T. J., Lust, N., Blecic, J., & Stemm, M. 2017, *AJ*, 153, 3
- Cubillos, P. E., Harrington, J., Blecic, J., Himes, M. D., Rojo, P. M., Loredó, T. J., Lust, N. B., Challener, R. C., Foster, A. J., Stemm, M. M., Foster, A. S. D., & Blumenthal, S. D. 2022, *The Planetary Science Journal*, 3, 81
- Deming, L. D. & Seager, S. 2017, *Journal of Geophysical Research (Planets)*, 122, 53
- Föhring, D., Dhillon, V. S., Madhusudhan, N., Marsh, T. R., Copperwheat, C. M., Littlefair, S. P., & Wilson, R. W. 2013, *MNRAS*, 435, 2268

- Gordon, S. & McBride, B. J. 1994, Computer Program for Calculation of Complex Chemical Equilibrium Compositions and Applications: I. Analysis, Tech. Rep. Reference Publication 1311, National Aeronautics and Space Administration, Washington, DC
- Hargreaves, R. J., Gordon, I. E., Rey, M., Nikitin, A. V., Tyuterev, V. G., Kochanov, R. V., & Rothman, L. S. 2020, *ApJS*, 247, 55
- Harrington, J., Himes, M. D., Cubillos, P. E., Blecic, J., Rojo, P. M., Challener, R. C., Lust, N. B., Bowman, M. O., Blumenthal, S. D., Dobbs-Dixon, I., Foster, A. S. D., Foster, A. J., Green, M. R., Lored, T. J., McIntyre, K. J., Stemm, M. M., & Wright, D. C. 2022, *The Planetary Science Journal*, 3, 80
- Hebb, L., Collier-Cameron, A., Loeillet, B., Pollacco, D., Hébrard, G., Street, R. A., Bouchy, F., Stempels, H. C., Moutou, C., Simpson, E., Udry, S., Joshi, Y. C., West, R. G., Skillen, I., Wilson, D. M., McDonald, I., Gibson, N. P., Aigrain, S., Anderson, D. R., Benn, C. R., Christian, D. J., Enoch, B., Haswell, C. A., Hellier, C., Horne, K., Irwin, J., Lister, T. A., Maxted, P., Mayor, M., Norton, A. J., Parley, N., Pont, F., Queloz, D., Smalley, B., & Wheatley, P. J. 2009, *ApJ*, 693, 1920
- Heng, K. & Lyons, J. R. 2016, *ApJ*, 817, 149
- John, T. L. 1988, *A&A*, 193, 189
- Kreidberg, L., Line, M. R., Parmentier, V., Stevenson, K. B., Louden, T., Bonnefoy, M., Faherty, J. K., Henry, G. W., Williamson, M. H., & Stassun, K. 2018, *AJ*, 156, 17
- Laraia, A. L., Gamache, R. R., Lamouroux, J., Gordon, I. E., & Rothman, L. S. 2011, *Icarus*, 215, 391
- Line, M. R., Knutson, H., Wolf, A. S., & Yung, Y. L. 2014, *ApJ*, 783, 70

- Line, M. R., Wolf, A. S., Zhang, X., Knutson, H., Kammer, J. A., Ellison, E., Deroo, P., Crisp, D., & Yung, Y. L. 2013, *ApJ*, 775, 137
- López-Morales, M., Coughlin, J. L., Sing, D. K., Burrows, A., Apai, D., Rogers, J. C., Spiegel, D. S., & Adams, E. R. 2010, *ApJ*, 716, L36
- Lothringer, J. D., Barman, T., & Koskinen, T. 2018, *ApJ*, 866, 27
- Madhusudhan, N. 2012, *ApJ*, 758, 36
- . 2018, in *Handbook of Exoplanets*, ed. H. J. Deeg & J. A. Belmonte (Springer International Publishing AG), 104
- Madhusudhan, N., Harrington, J., Stevenson, K., Nymeyer, S., Campo, C., Wheatley, P. J., Deming, D., Blecic, J., Hardy, R., Lust, N., Anderson, D. R., Collier Cameron, A., Britt, C., Bowman, W., Hebb, L., Hellier, C., Maxted, P. F. L., Pollacco, D., & West, R. G. 2011, *Nature*, 469, 64
- Márquez-Neila, P., Fisher, C., Sznitman, R., & Heng, K. 2018, *Nature Astronomy*
- Moses, J. I. 2014, *Philosophical Transactions of the Royal Society of London Series A*, 372, 20130073
- Moses, J. I., Madhusudhan, N., Visscher, C., & Freedman, R. S. 2013, *ApJ*, 763, 25
- Oreshenko, M., Lavie, B., Grimm, S. L., Tsai, S.-M., Malik, M., Demory, B.-O., Mordasini, C., Alibert, Y., Benz, W., Quanz, S. P., Trotta, R., & Heng, K. 2017, *ApJ*, 847, L3
- Parmentier, V., Line, M. R., Bean, J. L., Mansfield, M., Kreidberg, L., Lupu, R., Visscher, C., Désert, J.-M., Fortney, J. J., & Deleuil, M. 2018, *A&A*, 617, A110
- Richard, C., Gordon, I. E., Rothman, L. S., Abel, M., Frommhold, L., Gustafsson, M., Hartmann, J.-M., Hermans, C., Lafferty, W. J., Orton, G. S., Smith, K. M., & Tran, H. 2012, *J. Quant. Spec. Radiat. Transf.*, 113, 1276

Royo, P. M. 2006, PhD thesis, Cornell University

Rothman, L. S., Gordon, I. E., Babikov, Y., Barbe, A., Benner, D. C., Bernath, P. F., Birk, M., Bizzocchi, L., Boudon, V., Brown, L. R., et al. 2013, *J. Quant. Spec. Radiat. Transf.*, 130, 4

Rothman, L. S., Gordon, I. E., Barber, R. J., Dothe, H., Gamache, R. R., Goldman, A., Perevalov, V. I., Tashkun, S. A., & Tennyson, J. 2010, *J. Quant. Spec. Radiat. Transf.*, 111, 2139

Schwenke, D. W. 1998, *Faraday Discussions*, 109, 321

Stevenson, K. B., Bean, J. L., Madhusudhan, N., & Harrington, J. 2014, *ApJ*, 791, 36

Swain, M., Deroo, P., Tinetti, G., Hollis, M., Tessenyi, M., Line, M., Kawahara, H., Fujii, Y., Showman, A. P., & Yurchenko, S. N. 2013, *Icarus*, 225, 432

ter Braak, C. 2006, *Statistics and Computing*, 16, 239

ter Braak, C. J. F. & Vrugt, J. A. 2008, *Statistics and Computing*, 18, 435

Teske, J. K., Cunha, K., Smith, V. V., Schuler, S. C., & Griffith, C. A. 2014, *ApJ*, 788, 39

Wakeford, H. R., Visscher, C., Lewis, N. K., Kataria, T., Marley, M. S., Fortney, J. J., & Mandell, A. M. 2017, *MNRAS*, 464, 4247

Waldmann, I. P. & Griffith, C. A. 2019, *Nature Astronomy*, 308

Winn, J. N. & Fabrycky, D. C. 2015, *Annual Review of Astronomy and Astrophysics*, 53, 409

Zhao, M., Monnier, J. D., Swain, M. R., Barman, T., & Hinkley, S. 2012, *ApJ*, 744, 122

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



## CHAPTER 4: FROM BIOHINTS TO CONFIRMED EVIDENCE OF LIFE: POSSIBLE METABOLISMS WITHIN EXTRATERRESTRIAL ENVIRONMENTAL SUBSTRATES

**Michael D. Himes<sup>1</sup>, Molly D. O’Beirne<sup>2</sup>, Frank Soboczenski<sup>3</sup>, Simone Zorzan<sup>4</sup>, Atılım Güneş Baydin<sup>5</sup>, Adam D. Cobb<sup>5</sup>, Daniel Angerhausen<sup>6</sup>, Giada N. Arney<sup>7,8</sup>, Shawn D. Domagal-Goldman<sup>7,9</sup>**

<sup>1</sup> *Planetary Sciences Group, Department of Physics, University of Central Florida, Orlando, FL 32816-2385*

<sup>2</sup> *Department of Geology and Environmental Science, University of Pittsburgh, Pittsburgh, PA 15260, USA*

<sup>3</sup> *SPHES, King’s College London, UK*

<sup>4</sup> *ERIN Department, Luxembourg Institute of Science and Technology*

<sup>5</sup> *Department of Engineering Science, University of Oxford, UK*

<sup>6</sup> *Center for Space and Habitability (CSH), Universitat Bern, Sidlerstrasse 5, 3012 Bern, Switzerland*

<sup>7</sup> *NASA Astrobiology Institute, Virtual Planetary Laboratory Team, Seattle, Washington, USA* <sup>8</sup> *Planetary Systems Laboratory, NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA*

<sup>9</sup> *Planetary Environments Laboratory, NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA*

Published as part of the NASA Frontier Development Lab Proceedings 2018

Himes, M. D., M. D. O’Beirne, F. Soboczenski, S. Zorzan, A. G. Baydin, A. D. Cobb, D. Angerhausen, G. N. Arney, and S. D. Domagal-Goldman. 2018, *NASA FDL Proceedings 2018*, 158–181.

©SETI. Reproduced with permission.

## 4.1 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.

## 4.2 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).

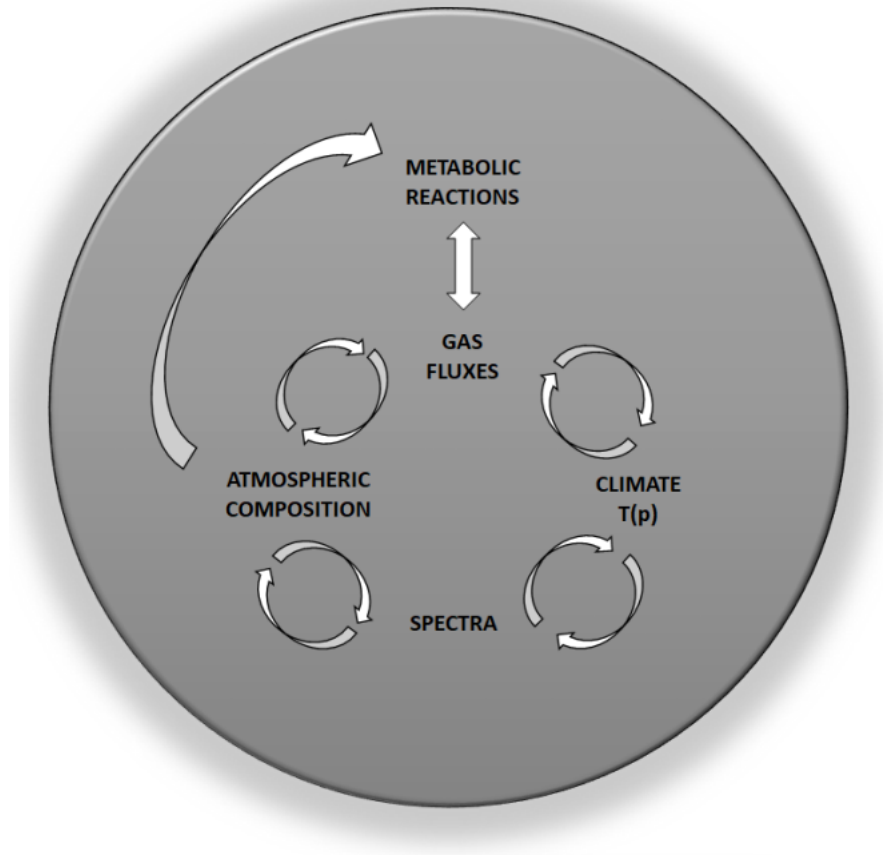


Figure 4.1: Schematic overview of the problem. Planetary spectra are a product of the pressure–temperature profile and atmospheric composition of an exoplanet. The atmospheric composition is influenced by gas fluxes resulting from geological processes, biological processes, or a combination of the two. 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.

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 4.1).

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 & 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 & 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.

## 4.3 Methods

### 4.3.1 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<sup>1</sup> 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 instantiate multiple such PSG instances via a single `Python` command<sup>2</sup>. Figure 4.2 displays the described software landscape.

---

<sup>1</sup>An environment `YAML Ain’t Markup Language (YML)` file is available at the INARA `GitLab` repository

<sup>2</sup>See the documentation at <https://gitlab.com/frontierdevelopmentlab/astrobiology/inara>

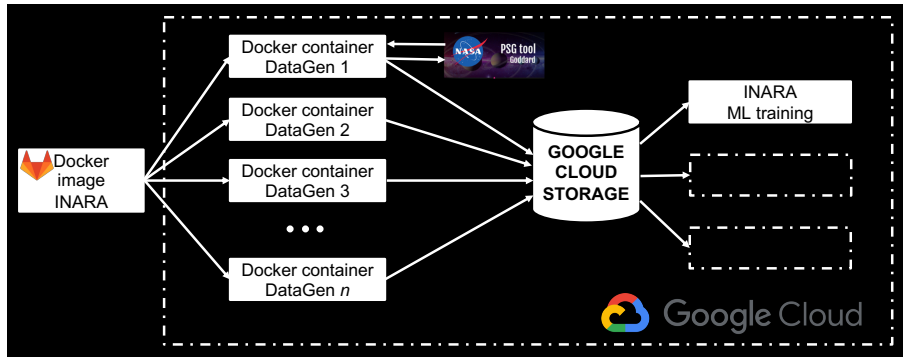


Figure 4.2: Overview of the entire implementation landscape. Once instantiated as a Docker container in the cloud, INARA communicates with a designated PSG VM to generate the data and store it in the cloud. INARA Docker containers are then able to read the data back in for ML training and inference purposes.

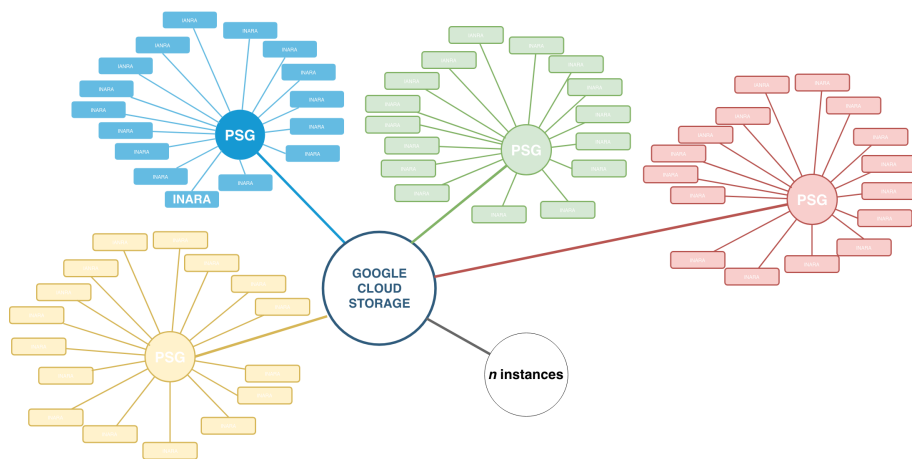


Figure 4.3: Structure of the instantiated INARA docker container and PSG VMs that act as nodes connected to the cloud storage.

Google Cloud also served as the main computational platform for data generation, model training and inference. Via Google Cloud, we were able to instantiate  $\sim 2000$  VMs (groups of 16 INARA instances connected to one PSG node) for data generation. Figure 4.3 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.

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.

#### *4.3.2 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.

##### *4.3.2.1 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  $\mu\text{m}$  with low read noise ( $1 \text{ e}^-/\text{pixel}$ ) and dark current ( $0.001 \text{ e}^-/\text{s}$ ). The coronagraph's inner working angle is  $2 \lambda/D$ . 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  $\mu\text{m}$  to provide a ground-truth spectrum across a wide wavelength range which may be used for other investigations beyond what is considered here.

#### 4.3.2.2 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 & 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 & Catling (2014) that the planets in the solar system all have a tropopause around 0.1 bars.

We consider 12 molecules:  $\text{H}_2\text{O}$ ,  $\text{CO}_2$ ,  $\text{O}_2$ ,  $\text{N}_2$ ,  $\text{CH}_4$ ,  $\text{N}_2\text{O}$ ,  $\text{CO}$ ,  $\text{O}_3$ ,  $\text{SO}_2$ ,  $\text{NH}_3$ ,  $\text{C}_2\text{H}_6$ , and  $\text{NO}_2$ . Table 4.1 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 ( $\text{CO}_2$ ,  $\text{O}_2$ ,  $\text{N}_2$ ). 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.  $\text{CO}_2$  and  $\text{N}_2$  have been observed in excess of 90% in atmospheres in the solar system, hence their upper limits.  $\text{O}_2$  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,  $\text{H}_2\text{O}$  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.  $\text{O}_3$ , a photochemical product of  $\text{O}_2$ , has been shown to be at most  $\sim 1\%$  the amount of  $\text{O}_2$  in prebiotic Earth-like atmospheres (Domagal-Goldman et al. 2014). We impose a similar limit for  $\text{C}_2\text{H}_6$  as it is a photochemical product of  $\text{CH}_4$ .  $\text{N}_2\text{O}$  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  $\text{N}_2\text{O}$  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;  $\text{H}_2\text{O}$  and  $\text{NH}_3$  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;  $\text{H}_2\text{O}$  covers a range of 255.9 to 573 K, and  $\text{NH}_3$  covers a range of 164 to 371.5 K. For an Earth-like planet, this method leads to the abundance of  $\text{H}_2\text{O}$  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 to the computational burden, as even poor modeling efforts increase computational time by a factor of  $\sim 50$ .

Table 4.1: Upper limits on random uniform distribution draw for each molecule

Molecule	H <sub>2</sub> O	CO <sub>2</sub>	O <sub>2</sub>	N <sub>2</sub>	CH <sub>4</sub>	N <sub>2</sub> O	CO	O <sub>3</sub>	SO <sub>2</sub>	NH <sub>3</sub>	C <sub>2</sub> H <sub>6</sub>	NO <sub>2</sub>
Upper limit	0.1	1.0	1.0	1.0	0.1	0.02	0.02	0.01*O <sub>2</sub>	0.02	0.01	0.01*CH <sub>4</sub>	2e-5

### 4.3.3 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 4.2 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 4.2: 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 <sub>2</sub> O, CO <sub>2</sub> , O <sub>2</sub> , N <sub>2</sub> , CH <sub>4</sub> , N <sub>2</sub> O, CO, O <sub>3</sub> , SO <sub>2</sub> , NH <sub>3</sub> , C <sub>2</sub> H <sub>6</sub> , and NO <sub>2</sub> )
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 <sup>2</sup> )
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

#### 4.3.4 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  $x = (x_1, x_2, x_3, \dots, x_n)$  and the predicted values  $\bar{x} = (\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_n)$  as displayed in:

$$E_{Tr}(\bar{x}_i, x_i) \tag{4.1}$$

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,  $f_\theta$ , as function that can accurately infer the parameters  $x_i$  for a given observation  $o_i$ :

$$x_i = f_\theta(o_i) \tag{4.2}$$

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

$$E_{Tr} = \sum_{i=1}^{Tr} E(\bar{x}_i = f_{\theta}(o_i), x_i) \quad (4.3)$$

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.

$$E_V = \sum_{i=1}^V E(\bar{x}_i = f_{\theta}(o_i), x_i) \quad (4.4)$$

#### 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 4.4). 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 algorithms (ADAM, SGD, ADAdelta, RMSProp) toward the selection of a model with good performance (see Section 4.4). 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.

#### 4.3.4.2 *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 & 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 4.6 shows the mean prediction compared to the true value for each planet in the test set, and Figure 4.7 show a predictive distribution resulting from dropout for a specific test planet.

## 4.4 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 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 4.5) 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.

### 4.4.1 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 4.6, where each dot represents the average of 600 runs of our model with dropout, for each of the 1000 predicted planetary parameters.

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 4.3. Low mean squared error (MSE) values indicate better predictions of the true value.



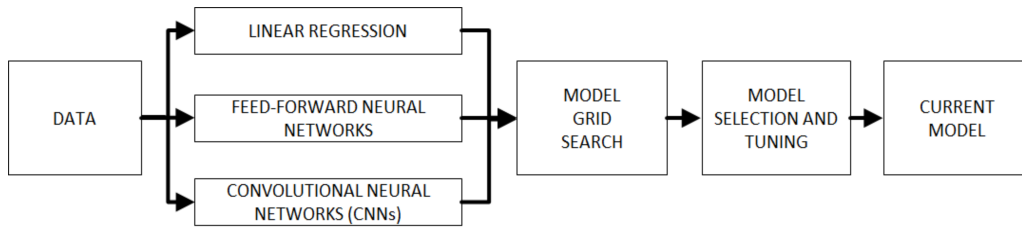


Figure 4.4: Evaluated ML architectures. We started with a simple linear regression (single layer), a standard feed forward and a convolutional neural network. Model grid search was led evaluating learning rates of 0.0001, 0.001 and 0.01, ADAM, SGD, ADAdelta and RMSProp optimizers, and as activation functions Tanh, Softmax, ReLU and linear. Once we found a reasonably performing model we selected the model used for further steps.

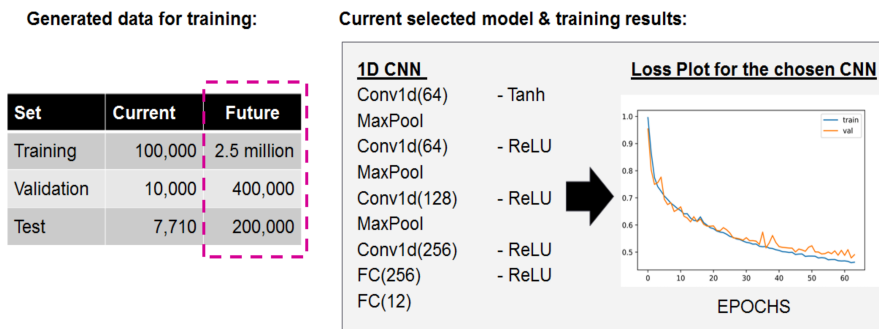


Figure 4.5: Available and used data in ML phases (left) and model architecture (on the right). Loss plot over the 64 epochs, with details on the relation between train and validation loss function.

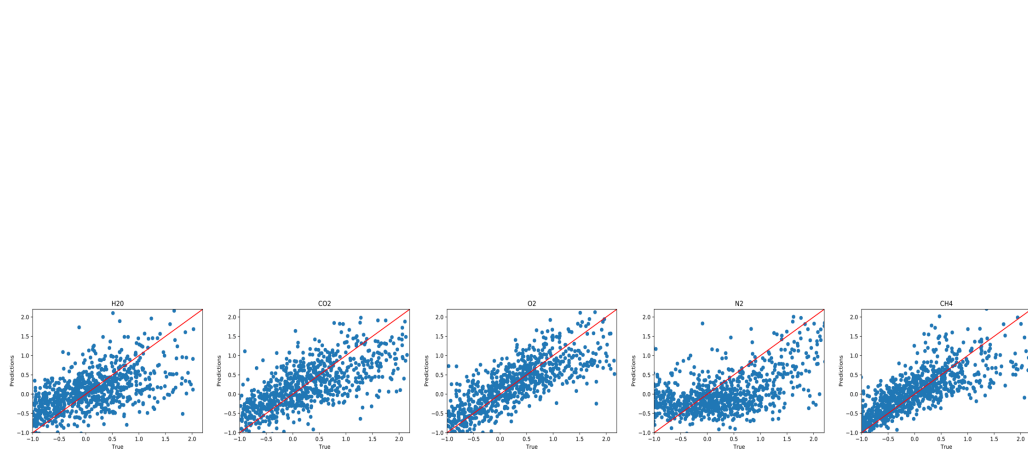


Figure 4.6: INARA prediction performance for the logarithm of the normalized abundance of  $\text{H}_2\text{O}$ ,  $\text{CO}_2$ ,  $\text{O}_2$ ,  $\text{N}_2$  and  $\text{CH}_4$  across 1000 test planets. We normalize values by subtracting the mean and dividing by the standard deviation across all 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.

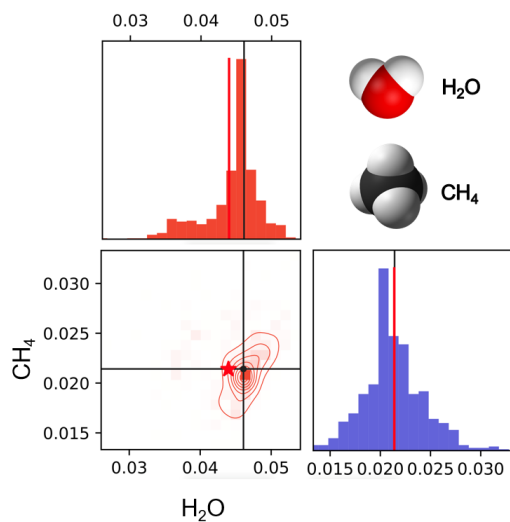


Figure 4.7: Detailed results of INARA in predicting one random planet's values for  $\text{H}_2\text{O}$  and  $\text{CH}_4$ . The distribution is obtained by 600 predictions with dropout where nodes were randomly removed from the network. The black lines represent INARA's median prediction, and the red lines and star represent, respectively, the real value in the dimension of  $\text{H}_2\text{O}$  and  $\text{CH}_4$ , and in the two-dimensional space with one molecule per axis.

Table 4.3: Reported error for five retrieved molecules

Error	$\text{H}_2\text{O}$	$\text{CO}_2$	$\text{O}_2$	$\text{N}_2$	$\text{CH}_4$
MSE	$3.43\text{e-}4$	$1.02\text{e-}2$	$7.00\text{e-}3$	$2.05\text{e-}2$	$1.93\text{e-}4$
$\pm 2\sigma$	$2.28\text{e-}3$	$3.53\text{e-}2$	$2.59\text{e-}2$	$5.21\text{e-}2$	$1.07\text{e-}3$

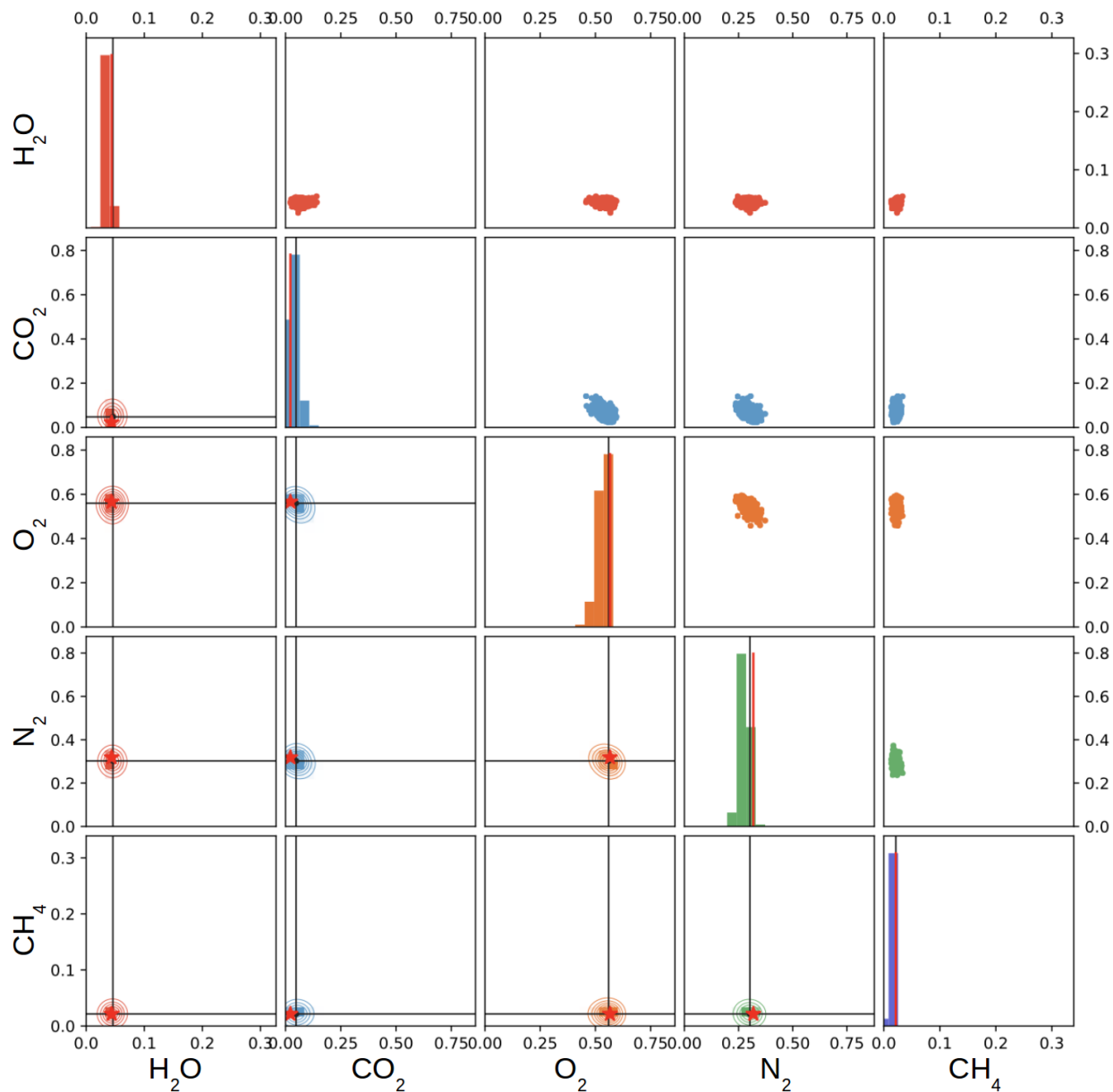


Figure 4.8: Detailed results of INARA in predicting one random planet’s values for H<sub>2</sub>O, CO<sub>2</sub>, O<sub>2</sub>, N<sub>2</sub>, and CH<sub>4</sub>. The distribution is obtained by 600 predictions with dropout where nodes were randomly removed from the network. Plots along the top-left 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.

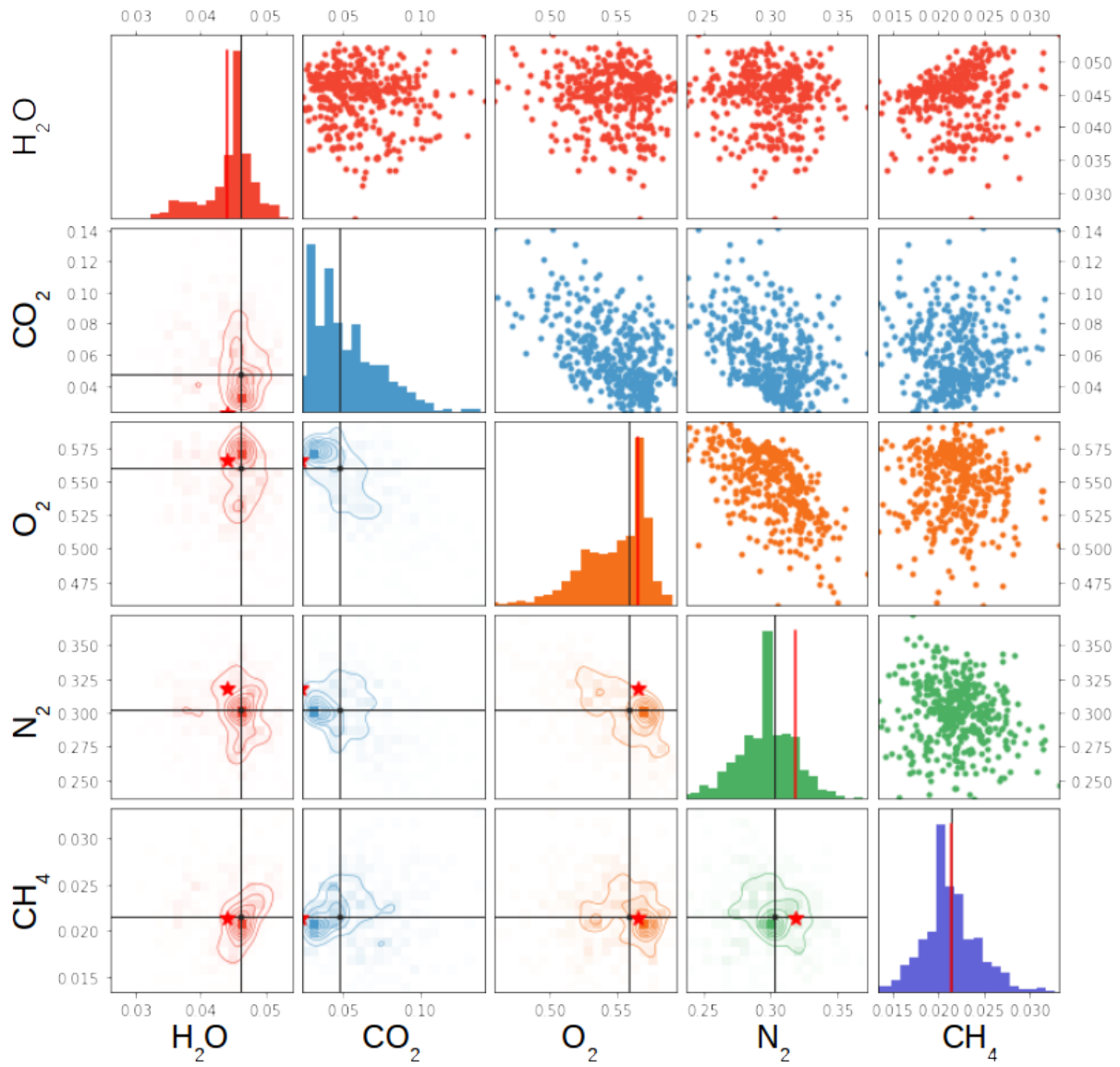


Figure 4.9: Same as Figure 4.8, but zoomed in around each distribution.

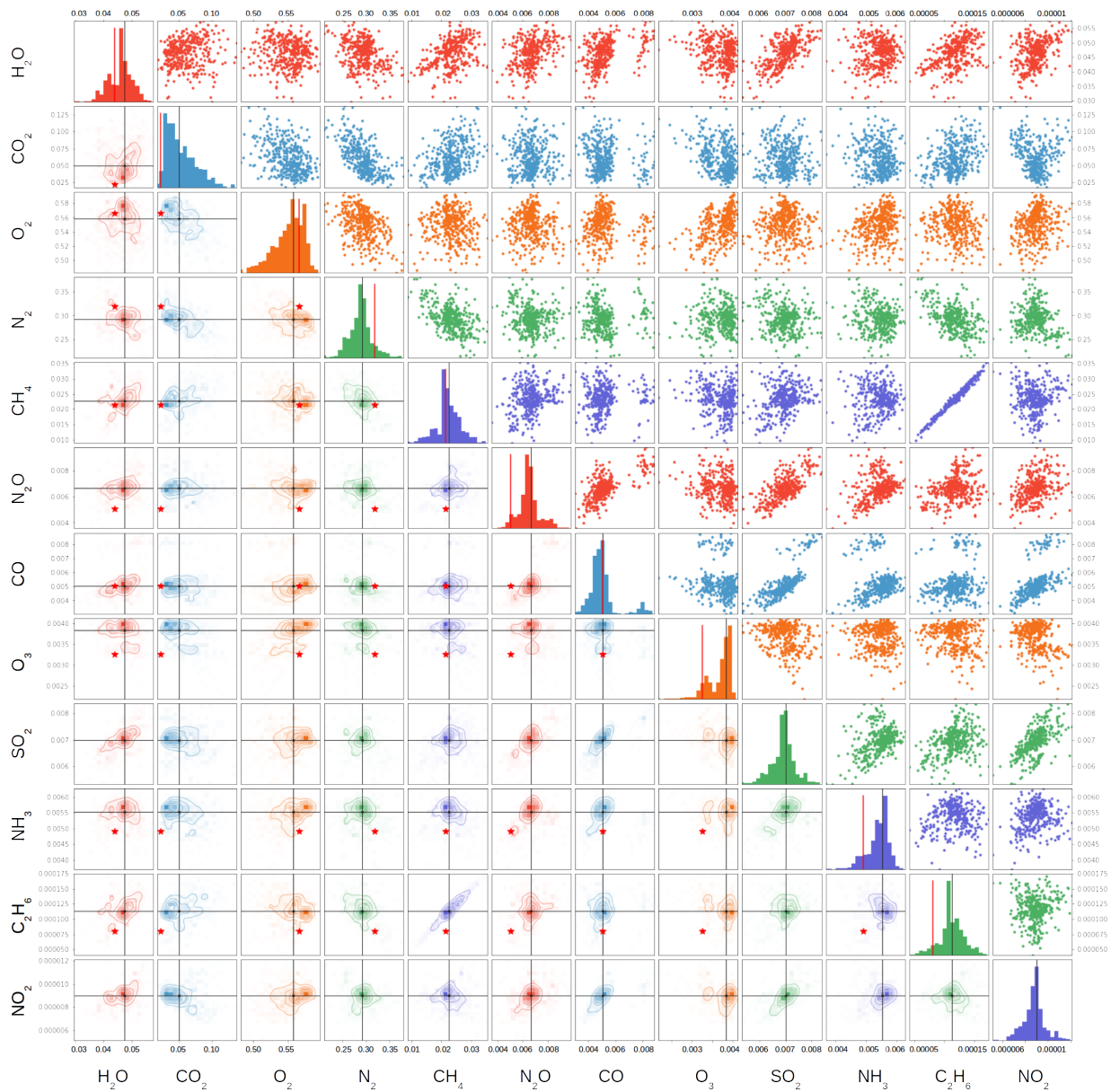


Figure 4.10: Same as Figure 4.8, but for all 12 molecules considered in the atmospheric model.

A more detailed representation of INARA's results is presented in Figure 4.7, which shows the predictive distributions for CH<sub>4</sub> and H<sub>2</sub>O 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 4.8, 4.9, and 4.10.

#### 4.4.2 Links to Code Repositories

The entire developed code palette is available open source at the *Frontier Development Lab* GitLab repository<sup>3</sup> 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>
- **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 2022 via the NASA Exoplanet Archive

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.

---

<sup>3</sup><https://gitlab.com/frontierdevelopmentlab>

## 4.5 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 4.4).

Table 4.4: Comparison of atmospheric retrieval methods

Method	CPU Time	# of Molecules Retrieved
Traditional	Hundreds of hours	User-specified
ExoGAN <sup>1</sup>	Minutes	4
HELA <sup>2</sup>	Seconds	3
INARA	Seconds	12

<sup>1</sup> Zingales & Waldmann (2018)

<sup>2</sup> Márquez-Neila et al. (2018)

Our present data set is the largest collection of rocky exoplanet spectra to date. The analytical temperature–pressure profile we adopted (see Section 4.3.2) allows the prediction of those parameters as well. While the simulated observed spectra span 0.2 to 2  $\mu\text{m}$ , our model still managed to predict the abundance of  $\text{N}_2$  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 & 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 4.3, the standard deviation is rather high for molecules with convoluted features in the 0.2 to 2  $\mu\text{m}$  range considered. We attribute this to the limited data set used for the present results, and we expect 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.

## 4.6 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<sup>4</sup> 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.

---

<sup>4</sup><http://www.kaggle.com>

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.

#### 4.7 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.

#### 4.8 Acknowledgements

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.

#### 4.9 List of References

- Boyajian, T. S., von Braun, K., van Belle, G., Farrington, C., Schaefer, G., Jones, J., White, R., McAlister, H. A., Theo, A., Ridgway, S., et al. 2013, *The Astrophysical Journal*, 771, 40
- Boyajian, T. S., Von Braun, K., Van Belle, G., McAlister, H. A., Theo, A., Kane, S. R., Muirhead, P. S., Jones, J., White, R., Schaefer, G., et al. 2012, *The Astrophysical Journal*, 757, 112
- Crossfield, I. J. M. 2015, *PASP*, 127, 941
- Domagal-Goldman, S. D., Segura, A., Claire, M. W., Robinson, T. D., & Meadows, V. S. 2014, *The Astrophysical Journal*, 792, 90

- Fujii, Y., Angerhausen, D., Deitrick, R., Domagal-Goldman, S., Grenfell, J. L., Hori, Y., Kane, S. R., Pallé, E., Rauer, H., Siegler, N., Stapelfeldt, K., & Stevenson, K. B. 2018, *Astrobiology*, 18, 739
- Gal, Y. & Ghahramani, Z. 2016, in international conference on machine learning, 1050–1059
- Goodfellow, I., Bengio, Y., & Courville, A. 2016, *Deep Learning* (MIT Press), <http://www.deeplearningbook.org>
- Graves, A., Mohamed, A.-r., & Hinton, G. 2013, in Acoustics, speech and signal processing (icassp), 2013 IEEE international conference on, IEEE, 6645–6649
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. 2012, arXiv preprint arXiv:1207.0580
- Kopparapu, R. K., Ramirez, R., Kasting, J. F., Eymet, V., Robinson, T. D., Mahadevan, S., Terrien, R. C., Domagal-Goldman, S., Meadows, V., & Deshpande, R. 2013, *The Astrophysical Journal*, 770, 82
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, in *Advances in neural information processing systems*, 1097–1105
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, *Proceedings of the IEEE*, 86, 2278
- Line, M. R., Wolf, A. S., Zhang, X., Knutson, H., Kammer, J. A., Ellison, E., Deroo, P., Crisp, D., & Yung, Y. L. 2013, *ApJ*, 775, 137
- Madhusudhan, N. 2018, *Handbook of Exoplanets*, 1
- Márquez-Neila, P., Fisher, C., Sznitman, R., & Heng, K. 2018, *Nature astronomy*, 2, 719
- Robinson, T. D. & Catling, D. C. 2014, *Nature Geoscience*, 7, 12

Rogers, L. A. 2015, *The Astrophysical Journal*, 801, 41

Schwieterman, E. W., Kiang, N. Y., Parenteau, M. N., Harman, C. E., DasSarma, S., Fisher, T. M., Arney, G. N., Hartnett, H. E., Reinhard, C. T., Olson, S. L., Meadows, V. S., Cockell, C. S., Walker, S. I., Grenfell, J. L., Hegde, S., Rugheimer, S., Hu, R., & Lyons, T. W. 2018, *Astrobiology*, 18, 663

Skilling, J. 2004in , *AIP*, 395–405

Sotin, C., Grasset, O., & Mocquet, A. 2007, *Icarus*, 191, 337

ter Braak, C. J. & Vrugt, J. A. 2008, *Statistics and Computing*, 18, 435

Villanueva, G. L., Smith, M. D., Protopapa, S., Faggi, S., & Mandell, A. M. 2018, *Journal of Quantitative Spectroscopy and Radiative Transfer*

Zahnle, K. J. & Catling, D. C. 2017, *The Astrophysical Journal*, 843, 122

Zingales, T. & Waldmann, I. P. 2018, arXiv preprint arXiv:1806.02906

# CHAPTER 5: ACCURATE MACHINE LEARNING ATMOSPHERIC RETRIEVAL VIA A NEURAL NETWORK SURROGATE MODEL FOR RADIATIVE TRANSFER

Michael D. Himes<sup>1</sup>, Joseph Harrington<sup>2</sup>, Adam D. Cobb<sup>3</sup>, Atılım Güneş Baydin<sup>3</sup>, Frank Soboczenski<sup>4</sup>, Molly D. O’Beirne<sup>5</sup>, Simone Zorzan<sup>6</sup>, David C. Wright<sup>1</sup>, Zacchaeus Scheffer<sup>1</sup>, Shawn D. Domagal-Goldman<sup>7</sup>, Giada N. Arney<sup>7</sup>

<sup>1</sup> *Planetary Sciences Group, Department of Physics, University of Central Florida, Orlando, FL 32816, USA*

<sup>2</sup> *Planetary Sciences Group, Department of Physics and Florida Space Institute, University of Central Florida, Orlando, FL 32816, USA*

<sup>3</sup> *Department of Engineering Science, University of Oxford, UK*

<sup>4</sup> *SPHES, King’s College London, UK*

<sup>5</sup> *Department of Geology and Environmental Science, University of Pittsburgh, Pittsburgh, PA 15260, USA*

<sup>6</sup> *ERIN Department, Luxembourg Institute of Science and Technology*

<sup>7</sup> *NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA*

Received 4 March 2020.

Accepted 4 February 2021.

Published in *The Planetary Science Journal* 25 April 2022.

Himes, M. D., J. Harrington, A. D. Cobb, A. G. Baydin, F. Soboczenski, M. D. O’Beirne, S. Zorzan, D. C. Wright, Z. Scheffer, S. D. Domagal-Goldman, and G. N. Arney. 2022, PSJ, 3, 91

<https://doi.org/10.3847/PSJ/abe3fd>

## 5.1 Abstract

Atmospheric retrieval determines the properties of an atmosphere based on its measured spectrum. The low signal-to-noise ratios of exoplanet observations require a Bayesian approach to determine posterior probability distributions of each model parameter, given observed spectra. This inference is computationally expensive, as it requires many executions of a costly radiative transfer (RT) simulation for each set of sampled model parameters. Machine learning (ML) has recently been shown to provide a significant reduction in runtime for retrievals, mainly by training inverse ML models that predict parameter distributions, given observed spectra, albeit with reduced posterior accuracy. Here we present a novel approach to retrieval by training a forward ML surrogate model that predicts spectra given model parameters, providing a fast approximate RT simulation that can be used in a conventional Bayesian retrieval framework without significant loss of accuracy. We demonstrate our method on the emission spectrum of HD 189733 b and find good agreement with a traditional retrieval from the Bayesian Atmospheric Radiative Transfer (BART) code (Bhattacharyya coefficients of 0.9843–0.9972, with a mean of 0.9925, between 1D marginalized posteriors). This accuracy comes while still offering significant speed enhancements over traditional RT, albeit not as much as ML methods with lower posterior accuracy. Our method is  $\sim 9\times$  faster per parallel chain than BART when run on an AMD EPYC 7402P central processing unit (CPU). Neural-network computation using an NVIDIA Titan Xp graphics processing unit is 90–180 $\times$  faster per chain than BART on that CPU.



## 5.2 Introduction

Over the past decades, exoplanet studies have expanded from their detection to include characterization of their atmospheres via retrieval (see reviews by Seager & Deming 2010, Deming & Seager 2017). Retrieval is the inverse modeling technique whereby forward models of a planet's spectrum are compared to observational data in order to constrain the model parameters (Madhusudhan 2018). These typically include the shape of the thermal profile, abundances of species, and condensate properties. While some solar system objects can be characterized with simpler approaches (such as Levenberg–Marquardt minimization) due to their high signal-to-noise ratios (e.g., Koskinen et al. 2016), retrieval on noisy exoplanet spectra require Bayesian methods to provide a distribution of models that can explain the observed data. The posterior distribution resulting from a Bayesian retrieval places limits on each model parameter (within some range, an upper or lower limit, or equally probable for all values considered), informing the statistical significance of the result.

Bayesian retrieval methods involve evaluating thousands to millions of spectra, integrating over the observational bandpasses, and comparing to observations. Depending on model complexity, this requires hundreds to thousands of parallelizable compute hours, resulting in hours to days of runtime. Calculating the model spectra by solving the radiative transfer (RT) equation takes the vast majority of compute time.

Machine learning (ML) encompasses algorithms that learn representations of and uncover relationships within a collection of data samples. Deep learning (Goodfellow et al. 2016) is a subfield of ML that is based on neural networks, which are highly flexible differentiable functions that can be fit to data. Neural networks can classify images (e.g., Krizhevsky et al. 2012, Simonyan & Zisserman 2015, Szegedy et al. 2015, He et al. 2016, Huang et al. 2017), recognize speech (e.g., Chorowski et al. 2014, Amodei et al. 2016, Chan et al. 2016, Xiong et al. 2016), and translate

between languages (e.g., Cho et al. 2014, Bahdanau et al. 2015, Ranzato et al. 2016, Sennrich et al. 2016, Wu et al. 2016). Neural networks consist of a hierarchy of layers that contain nodes performing weighted (non)linear transformations of their inputs, through a series of hidden layers, to the desired output. For example, for a retrieval, one might have the input layer receive the observed spectrum, hidden layers extract features, and the output layer predict the underlying atmospheric parameters. Neural-network training conventionally uses gradient-based optimization, iteratively adjusting the weights of the connections between nodes to minimize the error between the neural network’s prediction and the desired output (Rumelhart et al. 1986).

Recent applications of ML to atmospheric retrieval reduced compute time from hundreds of hours to minutes or less. Márquez-Neila et al. (2018) presented a random forest of regression trees to build predictive distributions comparable to the posterior distributions of traditional Bayesian retrievals. Zingales & Waldmann (2018) utilized a generative adversarial network (GAN; Goodfellow et al. 2014) to retrieve distributions for model parameters. Waldmann & Griffith (2019) used a convolutional neural network (CNN) to map spatial and spectral features across Saturn. In Cobb et al. (2019), we introduced `plan-net`, an ensemble of Bayesian neural networks that uses parameter correlations to inform the uncertainty on retrieved parameters. Hayes et al. (2019) demonstrated a new approach to ML retrieval by applying *k*-means clustering to a principal component analysis of the observed spectrum to inform a standard Bayesian retrieval. Johnsen & Marley (2019) showed that a dense neural network can provide quick estimations of atmospheric properties.

While these approaches are promising, all except Hayes et al. (2019) suffer from a common deficiency: the reduction in computational time is accompanied by a reduction in posterior accuracy because they make significant approximations when performing Bayesian inference. For ML to become an integral part of atmospheric retrieval, the accuracy of the posterior approximation must be preserved.

The solution lies in simulation-based inference methods (Cranmer et al. 2019). While directly using a simulator (e.g., RT code) requires a consistent amount of compute time for each new inference (e.g., retrieval), surrogate models that emulate the simulator (e.g, neural networks) allow new data to be quickly evaluated after an upfront computational cost to train the surrogate (Kasim et al. 2021, Munk et al. 2019). ML- and simulation-based inference approaches have been successfully applied to a variety of tasks ranging from quantum chemistry (Gilmer et al. 2017) to particle physics (Brehmer et al. 2018, Baydin et al. 2019), resulting in significant reductions in compute cost with minimal loss in accuracy. Similar approaches have been used by the Earth science community to reduce the computational burden of forward modeling of spectra, retrieval of surface conditions, and atmospheric correction (e.g., Atzberger 2004, Garcia-Cuesta et al. 2009, Rivera et al. 2015, Verrelst et al. 2015, Gmez-Dans et al. 2016, Verrelst et al. 2016, Verrelst et al. 2017, Chernetskiy et al. 2018, Yin et al. 2018, Vicent et al. 2018, Bue et al. 2019).

Here we present a novel application of this approach to retrieval, which uses a neural-network model of RT within a Bayesian framework, apply it to the emission spectrum of HD 189733 b, and compare the results to a classical retrieval using the RT code that trained the surrogate model. Our general method is to (1) generate a data set over some parameter space, (2) train a surrogate forward model on the generated data, and (3) infer the inverse process via a Bayesian sampler (Figure 5.1). Our approach circumvents the existing limitations of ML retrieval methods, which seek to directly learn the inverse process, by learning the forward, deterministic process (RT) and using the simulator surrogate in a standard inference pipeline. This approach preserves the accuracy of the Bayesian inference and, while slower than direct ML retrieval, is still much faster than computing RT.

In Section 5.3, we describe our approach in detail as well as introduce the software packages that implement the method. Section 5.4 discusses the results. Finally, Section 5.5 presents conclusions.

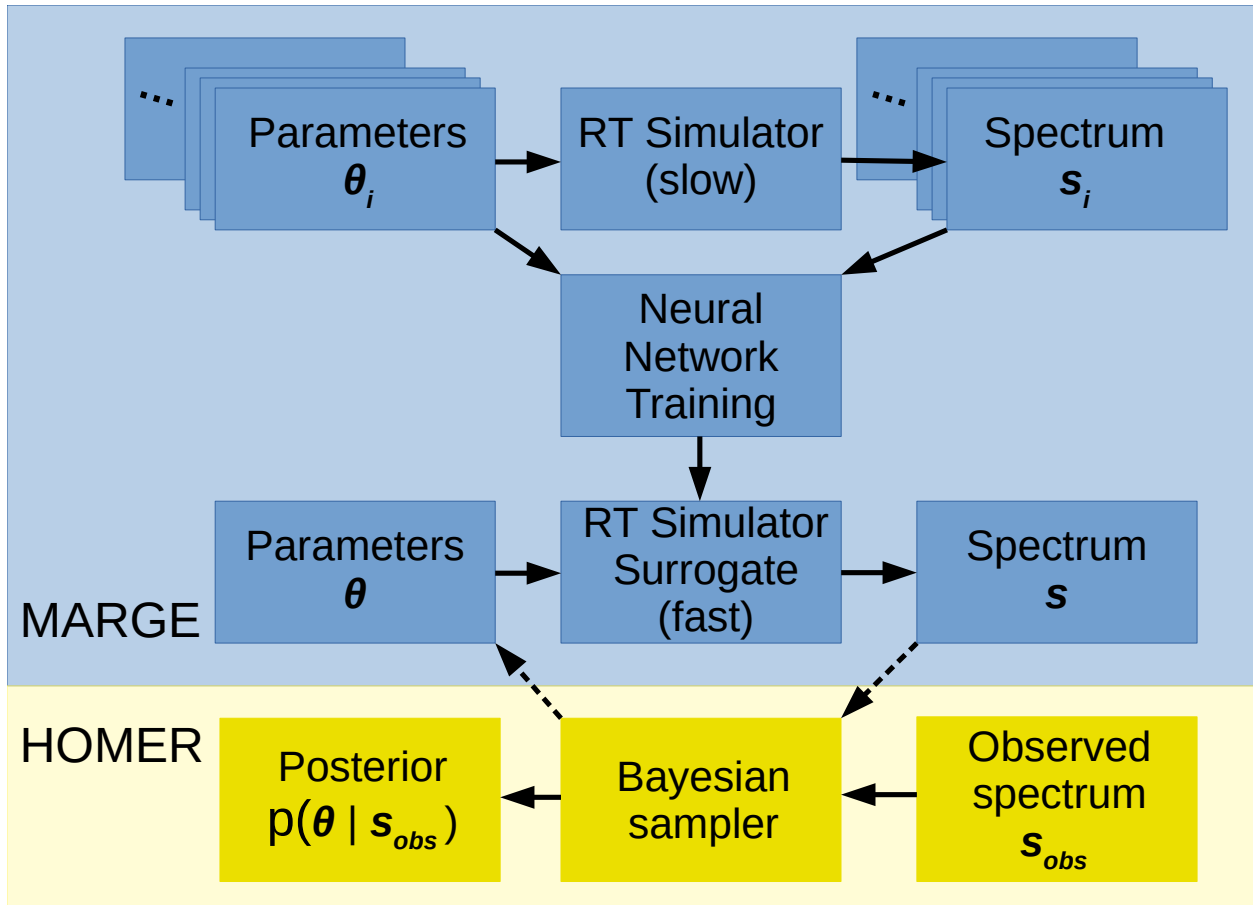


Figure 5.1: Schematic diagram of our inverse modeling method, color-coded based on the scope of our software packages. MARGE (Section 5.3.3.1) generates a data set based on a deterministic, forward process (e.g., RT) and trains a surrogate model to approximate that process. Using the trained surrogate, HOMER (Section 5.3.3.2) infers the inverse process (e.g., atmospheric retrieval) by simulating many forward models and comparing them to the target data (e.g., an observed spectrum) in a Bayesian framework.

## 5.3 Methods

### 5.3.1 Model Training

To train a neural network for our approach (Figure 5.1), we generate a data set of spectra using the Bayesian Atmospheric Radiative Transfer (BART) code (Harrington et al. 2022, Cubillos et al. 2022, Blecic et al. 2022).

The atmospheric models consist of 100 log-uniform layers spanning pressures from  $10^{-8}$  to 100 bar, and we assume that the planet radius corresponds to a pressure of 0.1 bar. We use the five-parameter temperature–pressure profile,  $T(p)$ , parameterization of Line et al. (2013):  $\kappa$ , the Planck mean infrared opacity;  $\gamma_1$  and  $\gamma_2$ , the ratios of the Planck mean visible and infrared opacities for each of two streams;  $\alpha$ , which controls the contribution of the two streams; and  $\beta$ , which represents albedo, emissivity, and energy recirculation. We allow the radius ( $R_p$ ), mass ( $M_p$ ), and semimajor axis ( $a$ , adjusts the temperature at the top of the atmosphere due to stellar irradiation) of the planet to vary to encompass a range of hot Jupiters. We also include a free parameter for each of the uniform vertical abundance profiles of  $\text{H}_2\text{O}$ ,  $\text{CO}_2$ ,  $\text{CO}$ , and  $\text{CH}_4$ .

We allow a wide range of values without regard for physical plausibility, except by enforcing that (1) the  $\text{H}_2/\text{He}$  ratio remains constant, (2) the total relative abundances of molecules in the atmosphere equals 1, and (3) the  $T(p)$  profile does not exceed the temperature range of the line lists. For example, this could lead to models with  $\text{H}_2\text{O}$  at conditions where it dissociates (Arcangeli et al. 2018), though in the case of HD 189733 b, such models would be rejected with a high probability due to a poor fit. We note that these are not fundamental constraints of our approach; other constraints (e.g., enforcing thermochemical equilibrium, keeping elemental ratios within some range) may be used when generating the data set to train the surrogate model.

Table 5.1: Forward Model Parameter Space

Parameter	Minimum	Maximum
$\log \kappa$	-5.0	1.0
$\log \gamma_1$	-2.0	2.0
$\log \gamma_2$	-1.3	1.3
$\alpha$	0.0	1.0
$\beta$	0.7	1.3
$R_p (R_J)$	0.8	1.5
$M_p (M_J)$	0.8	1.5
$a$ (AU)	0.2	0.4
$\log \text{H}_2\text{O}$	-13	-0.5
$\log \text{CO}_2$	-13	-0.5
$\log \text{CO}$	-13	-0.5
$\log \text{CH}_4$	-13	-0.5

For opacities, we use HITEMP for  $\text{H}_2\text{O}$ ,  $\text{CO}$ , and  $\text{CO}_2$  (Goorvitch 1994, Tashkun et al. 2003, Barber et al. 2006, Rothman et al. 2010), HITRAN for  $\text{CH}_4$  (Niederer et al. 2008, Boudon et al. 2010, Nikitin et al. 2010, 2011, Brown et al. 2013, Campargue et al. 2013, Daumont et al. 2013, Niederer et al. 2013, Nikitin et al. 2013, Rothman et al. 2013), and collision-induced absorptions of  $\text{H}_2\text{-H}_2$  and  $\text{H}_2\text{-He}$  (Borysow et al. 2001, Borysow 2002, Abel et al. 2012, Richard et al. 2012). While there are newer line lists available with a greater number of lines (e.g., Hargreaves et al. 2020), these tests are meant to demonstrate consistency between neural-network-based and non-ML retrievals; we therefore use the setup described in Harrington et al. (2022), which uses this set of line lists to compare with previous studies. As our approach learns RT from a data set of spectra, it is not tied to any specific line lists.

To train our neural-network surrogate model, we generate 3,458,432 spectra, which are subdivided into 2,446,784 spectra ( $\sim 70\%$ ) for training, 689,536 spectra ( $\sim 20\%$ ) for validation, and 322,112 spectra ( $\sim 10\%$ ) for testing (for considerations about data set size, see Appendix C). Model parameters come from the uniform distribution bound by the limits listed in Table 5.1. Each spectrum spans 280–7100  $\text{cm}^{-1}$  at a resolution of 1.0  $\text{cm}^{-1}$  and corresponds to the planet’s emitted flux in  $\text{erg s}^{-1} \text{cm}^{-1}$ .

When processing the BART inputs/outputs for our neural network, we simplify the neural-network inputs by transforming the planet mass into the surface gravity, because this is a factor in the integration to calculate the spectrum. We assume a host star of radius  $0.756 R_{\odot}$  with a temperature of 5000 K to calculate the  $T(p)$  profiles; because  $\beta$  acts as a scaling factor on the related term (Eq. 15 of Line et al. 2013), it can compensate for different stellar fluxes.

We normalize the input and output data by (1) taking the logarithm of the output spectra, (2) standardizing the inputs and (log) outputs by subtracting the training mean and dividing by the training standard deviation, and (3) scaling the standardized inputs and outputs to be in the range [-1, 1]. The neural network’s input layer corresponds to the 12 inputs described above, with surface gravity replacing planetary mass. The hidden layers consist of Conv1d(256)L(0.05) – Dense(4096)L(0.05) – Dense(4096)L(0.05) – Dense(4096)L(0.05) – Dense(4096)L(0.05). Conv1d( $n$ ) indicates a 1D convolutional layer with  $n$  feature maps and a kernel size of 3. L( $m$ ) indicates a leaky rectified linear unit (ReLU) activation function with slope  $m$  for  $x < 0$ . The dense output layer has 6821 nodes, corresponding to the emitted spectrum over the defined wavenumber grid, with a ReLU activation function. For details on our model selection process, see Appendix B.

We train with a batch size of 64 using a mean-squared-error loss function, the Adam optimizer, and early stopping with a patience of 30 epochs based on the validation loss. We employ a cyclical learning rate that increases from  $8 \times 10^{-6}$  to  $5 \times 10^{-3}$  over 4 epochs, then decreases over the same window. After each complete cycle (8 epochs), the maximum learning rate decays by half the difference between the maximum and minimum learning rates (*triangular2* policy, Smith 2015). The boundaries were chosen according to the method described in Smith (2015), except that we consider the loss instead of accuracy (see Appendix B for details). To evaluate the model’s performance, we compute the root-mean-squared error (RMSE; comparable to the standard deviation of the differences between the predicted and true values) and the coefficient of determination ( $R^2$ ; measures the linear correlation between the predicted and true values) between the data and the predictions, both for the full high-resolution output and the band-integrated spectra corresponding to the observations of HD 189733 b.

### 5.3.2 Retrieval

Following the setup of Harrington et al. (2022), we perform a retrieval of the dayside atmosphere of HD 189733 b based on the measurements by the Hubble Space Telescope Near Infrared Camera MultiObject Spectrograph (Swain et al. 2009); Spitzer Space Telescope Infrared Spectrograph (IRS Grillmair et al. 2008); Spitzer InfraRed Array Camera (IRAC) channels 1 and 2 values of  $0.1533 \pm 0.0029\%$  and  $0.1886 \pm 0.0071\%$  (M. Line, priv. comm.); IRAC channel 3, IRS 16  $\mu\text{m}$  photometry, and Multiband Imaging Photometer for Spitzer (Charbonneau et al. 2008); and IRAC channel 4 (Agol et al. 2010). We use a K2 solar-abundance Kurucz stellar model for the host star’s emission (Castelli & Kurucz 2003). Using the differential evolution Markov chain with snooker updating



algorithm of ter Braak & Vrugt (2008), 2,500,000 iterations are spread across 10 parallel chains, with a burn-in of 50,000 iterations per chain. When retrieving, we fix the semimajor axis to 0.031 au and the planetary radius and gravity at 0.1 bar to  $1.138 R_J$  and  $2187.762 \text{ cm s}^{-2}$ , respectively. The remaining neural-network input parameters are allowed to freely vary over the entire training space.

We compute the Bhattacharyya coefficient (Bhattacharyya 1943, Aherne et al. 1998) to compare the similarity of 1D marginalized posteriors, where a value of 0 indicates no overlap and a value of 1 indicates identical distributions. We choose this metric over others, such as the Kullback-Leibler divergence, because it is both intuitive to understand and defined for all distributions, even those that do not overlap.

For this investigation, we focus on a neural network as a faster replacement for an RT code for retrieval; we therefore only compare the results of BART and the neural-network approximation. For a discussion of these results in the context of previous retrievals of HD 189733 b’s dayside atmosphere, see Harrington et al. (2022).

### 5.3.3 *Software*

We have developed two Python packages for this investigation. Both are open-source software, with full documentation, under the Reproducible Research Software License<sup>1</sup>. We encourage users to contribute to the code via pull requests on Github.

---

<sup>1</sup><https://planets.ucf.edu/resources/reproducible-research/software-license/>

### 5.3.3.1 *MARGE*

The Machine learning Algorithm for Radiative transfer of Generated Exoplanets<sup>2</sup> (*MARGE*, Figure 5.1) (1) generates a data set based on a user-supplied function, (2) processes the generated data using a user-supplied function, and (3) trains, validates, and tests a user-specified neural-network architecture on a data set. The software package allows independent execution of any of the three modes, enabling a wide range of applications beyond exoplanet retrieval.

*MARGE*'s design allows it to be applied to any deterministic model. For 1D data (such as spectra), *MARGE*'s desired format is NumPy binary (.npy) files of 2D arrays, where each row corresponds to a single case. Each row is a data vector of the input parameters followed by the output data (e.g., spectrum). *MARGE* currently includes data-generation and -processing functions for BART as well as a data-processing function for the `pypsg`<sup>3</sup> Python interface (Soboczenski et al. 2018) for the NASA Planetary Spectrum Generator (Villanueva et al. 2018). We encourage users to contribute code via pull request to handle the processing of the inputs/outputs of other software packages.

We implement neural-network model training in Keras (version 2.2.4, Chollet et al. 2015), using a Tensorflow (version 1.13.1, Abadi et al. 2016) backend. *MARGE* enables early stopping by default to prevent overfitting, and the user can halt or resume training. *MARGE* allows for cyclical learning rates for more efficient training (Smith 2015, see also Appendix B). Users specify the model architecture details and the data location, and the software handles the data normalization, training, validation, and testing. *MARGE* preprocesses the data into Tensorflow's TFRecords format for efficient handling. Users have multiple options when preprocessing the data, which include taking the logarithm of the inputs and/or outputs, standardizing the data according to its mean and

---

<sup>2</sup>*MARGE* is available at <https://github.com/exosports/marge>

<sup>3</sup><https://gitlab.com/frontierdevelopmentlab/astrobiology/pypsg>

standard deviation, and/or scaling the data to be within a specified range. The mean and standard deviation of the data set are computed using Welford's method (Welford 1962) to avoid the need to load the entire data set into memory at once. MARGE computes the RMSE and  $R^2$  for predictions on the validation and test sets to evaluate model performance; these metrics can optionally be calculated over integrated filter bandpasses. Finally, users may specify cases from the test set to plot the predicted and true spectra, with residuals (e.g., Figure 5.2).

### 5.3.3.2 HOMER

The Helper Of My Eternal Retrievals<sup>4</sup> (HOMER) utilizes a MARGE-trained model to infer the underlying inputs corresponding to some observed outputs (Figure 5.1). For its Bayesian framework, HOMER uses a Python wrapper for Markov chain and nested-sampling algorithms. The user specifies data, uncertainties, observational filters, a parameter space, and a few related inputs, which are passed to the Bayesian sampler to perform the inference. If available, a graphics processing unit (GPU) calculates neural-network predictions, though the central processing unit (CPU) can do this at the cost of increased runtime. For each iteration of the Bayesian inference, the trained neural network predicts on the proposed input parameters, which are modified as necessary (descale, denormalize, divide by the stellar spectrum, unit conversions, and/or integrated over bandpasses).

---

<sup>4</sup>HOMER is available at <https://github.com/exosports/homer>

HOMER produces plots of the best-fit spectrum, 1D marginalized posteriors, 2D pairwise posteriors, and parameter history traces. The best-fit spectrum plot contains the data (with observational bandpasses indicated by uncertainties in  $x$ ) and, if the DataSketches<sup>5</sup> library is installed, the  $1\sigma$ ,  $2\sigma$ , and  $3\sigma$  spectra. We use the streaming quantiles method of Karnin et al. (2016) as implemented in DataSketches to compute the 1-2-3 $\sigma$  spectra. This approach avoids needing to load all of the evaluated models at once, which could exceed system memory.

HOMER calculates the steps per effective independent sample (SPEIS) and effective sample size (ESS) as described in Harrington et al. (2022). Markov chains make small, correlated steps; while a chain may perform 100,000 iterations, if it takes 5000 steps to materialize a completely independent sample (steps per effective independent sample, SPEIS), then there have only been 20 effective samples. SPEIS is calculated from the autocorrelation function of each parameter for each chain; as a conservative estimate, we use the highest SPEIS value when calculating the ESS of the Bayesian inference to ensure we do not underestimate credible region uncertainties. By rearranging Equation 1 of Harrington et al. (2022), an uncertainty  $s_{\hat{C}}$  can be calculated on a given credible region  $\hat{C}$  based on the ESS:

$$s_{\hat{C}} \approx \sqrt{\frac{\hat{C}(1 - \hat{C})}{ESS}} \quad (5.1)$$

For example, if the ESS is 20, then the determined 68.27% credible region is actually the  $68.27 \pm 10\%$  credible region; running the inference for more iterations would increase the ESS and accordingly decrease the uncertainty on that credible region.

For easy comparison with other retrieval results, HOMER can overplot the 1D and 2D posteriors for multiple retrievals (e.g., Figure 5.3) and compute the Bhattacharyya coefficients between the 1D posteriors.

---

<sup>5</sup><https://datasketches.apache.org/>

Table 5.2: Model Evaluation: High-resolution Spectra

Metric	Min.	Median	Mean	Max.
Norm. RMSE	0.00153	0.00224	0.00247	0.01040
Norm. $R^2$	0.99999	1.00000	1.00000	1.00000
Denorm. $R^2$	0.99885	0.99993	0.99990	0.99997

**Notes.** RMSE and  $R^2$  are calculated for each of the 6821 outputs corresponding to the wavenumber grid of 280–7100  $\text{cm}^{-1}$  with a resolution of 1.0  $\text{cm}^{-1}$ . For conciseness, we present statistics about these values. The  $R^2$  values are slightly less than 1, but they round to 1 at the reported precision.

## 5.4 Results and Discussion

The normalized RMSE, normalized  $R^2$ , and denormalized  $R^2$  metrics for the MARGE-trained model on the test set for the high-resolution and band-integrated spectra are detailed in Tables 5.2 and 5.3, respectively. The normalized RMSE  $\ll 1$  and  $R^2 \sim 1$  indicate an accurate model for RT over the parameter space. Rather than waiting for early stopping to engage, we manually stopped training at 130 epochs because there was an insignificant improvement in the loss for dozens of epochs. For considerations on how this affects model performance, see Appendix C.

Figure 5.2 shows example comparisons between the spectra predicted by MARGE and true spectra calculated by BART. While residuals tend to be around a few percent, they generally fluctuate around 0; when band-integrated over the observational filters, these errors usually cancel, as shown by the lower normalized RMSE and higher denormalized  $R^2$  metrics (Tables 5.2 and 5.3). We observe that in some cases, there are regions where the spectrum is consistently over- or underestimated by a few percent (e.g., the top-left panel of Figure 5.2 around 4250  $\text{cm}^{-1}$ ), thereby introducing error in the band-integrated value. However, the small deviations appear to have only a minor effect on this retrieval’s result; see Section 5.4.1 for considerations when retrieving at high spectral resolutions or in cases where a traditional retrieval result is not available for comparison.

Table 5.3: Model Evaluation: Band-integrated Spectra

Metric	Min.	Median	Mean	Max.
Norm. RMSE	0.00123	0.00147	0.00148	0.00183
Norm. $R^2$	1.00000	1.00000	1.00000	1.00000
Denorm. $R^2$	0.99995	0.99997	0.99997	0.99998

**Notes.** Same as Table 5.2, except integrated over the 66 bandpasses corresponding to the referenced observations of HD 189733 b.

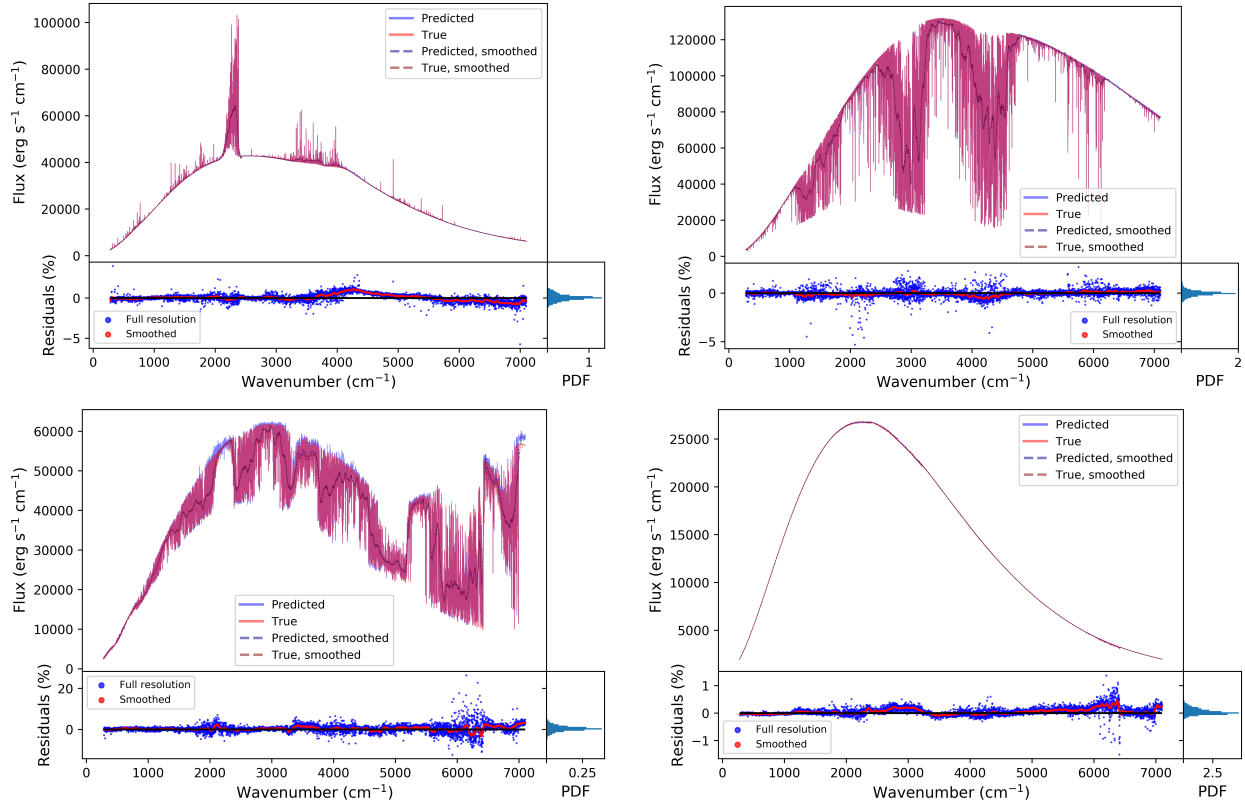


Figure 5.2: Four comparisons of planetary emission spectra predicted by MARGE and calculated by BART. The smoothed curves use a Savitzky–Golay filter with a third-order polynomial across a window of 101 elements ( $100 \text{ cm}^{-1}$ ). The purple color arises due to a detailed match between the red and blue spectra at high resolution. For the residuals, a black line is plotted at 0 to show regions where the neural network consistently over- or underpredicts the spectrum. A histogram of the high-resolution residuals appears to the right of the residual scatter plot, where the x-axis shows the probability density function (PDF) for the range of residual percentages. **Top left:** case with  $T(p)$  profile that increases in temperature with altitude, with  $\text{H}_2\text{O}$  and  $\text{CO}_2$  emission lines. **Top right:** case with  $T(p)$  profile that decreases in temperature with altitude, with absorption primarily due to  $\text{CH}_4$  and  $\text{H}_2\text{O}$ . **Bottom left:** cases with  $T(p)$  profile that has an inversion around 0.1 bar, with  $\text{CH}_4$ ,  $\text{CO}$ , and  $\text{CO}_2$  absorption and emission features. **Bottom right:** case with  $T(p)$  profile that is nearly isothermal at the pressures with sensitivity.

When applying HOMER to the emission spectrum of HD 189733 b, the results are consistent with BART. The retrieved  $T(p)$  profiles (bottom-left panel Figure 5.3) agree in the regions probed by the observations ( $<1$  bar, bottom-right panel Figure 5.3) and only begin to deviate deeper in the atmosphere, where little to no signal is measured according to the contribution functions. By nature, HOMER cannot calculate contribution functions, as the MARGE model does not solve RT. While they could be included for each case in the training set, this would require significantly more compute resources. Computing the contribution functions for the single best-fit case using the RT code that trained MARGE more efficiently uses compute resources.

Table 5.4 compares HOMER’s retrieved 68.27% (“ $1\sigma$ ”), 95.45% (“ $2\sigma$ ”), and 99.73% (“ $3\sigma$ ”) credible regions with BART’s retrieved credible regions. All regions closely agree, with differences attributable to a combination of uncertainty from a finite ESS and the neural network’s imperfect nature (Figure 5.3, top-right panel). For CO, both BART and HOMER favor large abundances, though BART finds a greater probability for log abundances  $\geq -2$  (Figure 5.3, top-right panel). Despite this, the credible regions agree (Table 5.4). Similarly, HOMER favors lower values for  $\gamma_1$  and  $\alpha$ , though the resulting thermal profiles agree (Figure 5.3, bottom-left panel).

Table 5.5 compares the SPEIS, ESS values, and associated uncertainties in the  $1\sigma$ ,  $2\sigma$ , and  $3\sigma$  credible regions for HOMER and BART. HOMER yields an SPEIS that is less than BART’s, attributable to the conservative estimate of SPEIS as being the greatest among all chains and parameters. The highest SPEIS values fluctuate between runs, though the median SPEIS remains relatively constant. HOMER’s median SPEIS of 627 and BART’S 615 better reflect the close agreement between the two retrievals. The Bhattacharyya coefficients between the 1D marginalized posteriors of HOMER and BART indicate agreement in the range 0.9843–0.9972, with a mean of 0.9925 (Table 5.6).



Table 5.4: Retrieved Credible Regions

Parameter	Code	68.27%	95.45%	99.73%
$\log \kappa$	HOMER	[-1.63, -1.06]	[-1.84, -0.71]	[-2.07, -0.33]
	BART	[-1.58, -1.09]	[-1.81, -0.79]	[-1.99, -0.46]
$\log \gamma_1$	HOMER	[-1.98, -1.65]	[-1.99, -1.34]	[-1.99, -1.06]
	BART	[-1.98, -1.62]	[-2.00, -1.33]	[-2.00, -1.07]
$\log \gamma_2$	HOMER	[0.34, 0.77]	[0.21, 1.10]	[0.11, 1.29]
	BART	[0.35, 0.73]	[0.21, 1.02]	[-0.07, 1.30]
$\alpha$	HOMER	[0.07, 0.39]	[0.03, 0.60]	[0.01, 0.74]
	BART	[0.11, 0.42]	[0.06, 0.60]	[0.02, 0.74]
$\beta$	HOMER	[1.01, 1.07]	[0.99, 1.12]	[0.96, 1.15]
	BART	[1.01, 1.06]	[0.99, 1.10]	[0.97, 1.15]
$\log \text{H}_2\text{O}$	HOMER	[-3.11, -2.37]	[-3.37, -1.82]	[-3.70, -1.27]
	BART	[-3.12, -2.44]	[-3.37, -1.92]	[-3.63, -1.41]
$\log \text{CO}_2$	HOMER	[-3.39, -2.73]	[-3.78, -2.36]	[-4.26, -2.01]
	BART	[-3.33, -2.71]	[-3.66, -2.32]	[-4.05, -2.03]
$\log \text{CO}$	HOMER	[-6.89, -0.51]	[-12.02, -0.51]	[-12.90, -0.51]
	BART	[-6.60, -0.50]	[-12.55, -0.50]	[-12.90, -0.50]
$\log \text{CH}_4$	HOMER	[-5.16, -3.53]	[-10.25, -3.20]	[-12.95, -3.12]
	BART	[-4.71, -3.67]	[-10.53, -3.14]	[-12.73, -3.07]

Table 5.5: Credible Region Accuracy

Code	SPEIS	ESS	$1\sigma$ Uncertainty	$2\sigma$ Uncertainty	$3\sigma$ Uncertainty
HOMER	1668	1199	1.34%	0.60%	0.15%
BART	2084	959	1.50%	0.67%	0.17%

Table 5.6: Bhattacharyya Coefficients

Parameter	Value
$\kappa$	0.9948
$\gamma_1$	0.9972
$\gamma_2$	0.9950
$\alpha$	0.9909
$\beta$	0.9879
H <sub>2</sub> O	0.9968
CO <sub>2</sub>	0.9968
CO	0.9888
CH <sub>4</sub>	0.9843
Mean	0.9925

#### 5.4.1 Limitations

HOMER’s accuracy is, by nature, bound by the accuracy of the neural-network model. Model inaccuracies may slightly bias the results, as seen in the minor differences between the posteriors of HOMER and BART. In our application, this discrepancy does not significantly affect the scientific conclusions at the spectral resolution of these observations for the current neural-network accuracy. However, this does not necessarily hold for all cases. It is possible that at higher resolutions this neural network’s minor inaccuracies can drive the Bayesian sampler to radically different results. While in theory MARGE works for any spectral resolution, users will need to carefully select the model architecture to ensure that it can accurately model the spectra over the desired phase space. In situations lacking a physics-based retrieval to compare with, we advise testing to ensure that forward models are reasonably accurate over the retrieval’s phase space, as some regions may not be sufficiently sampled for accurate predictions.

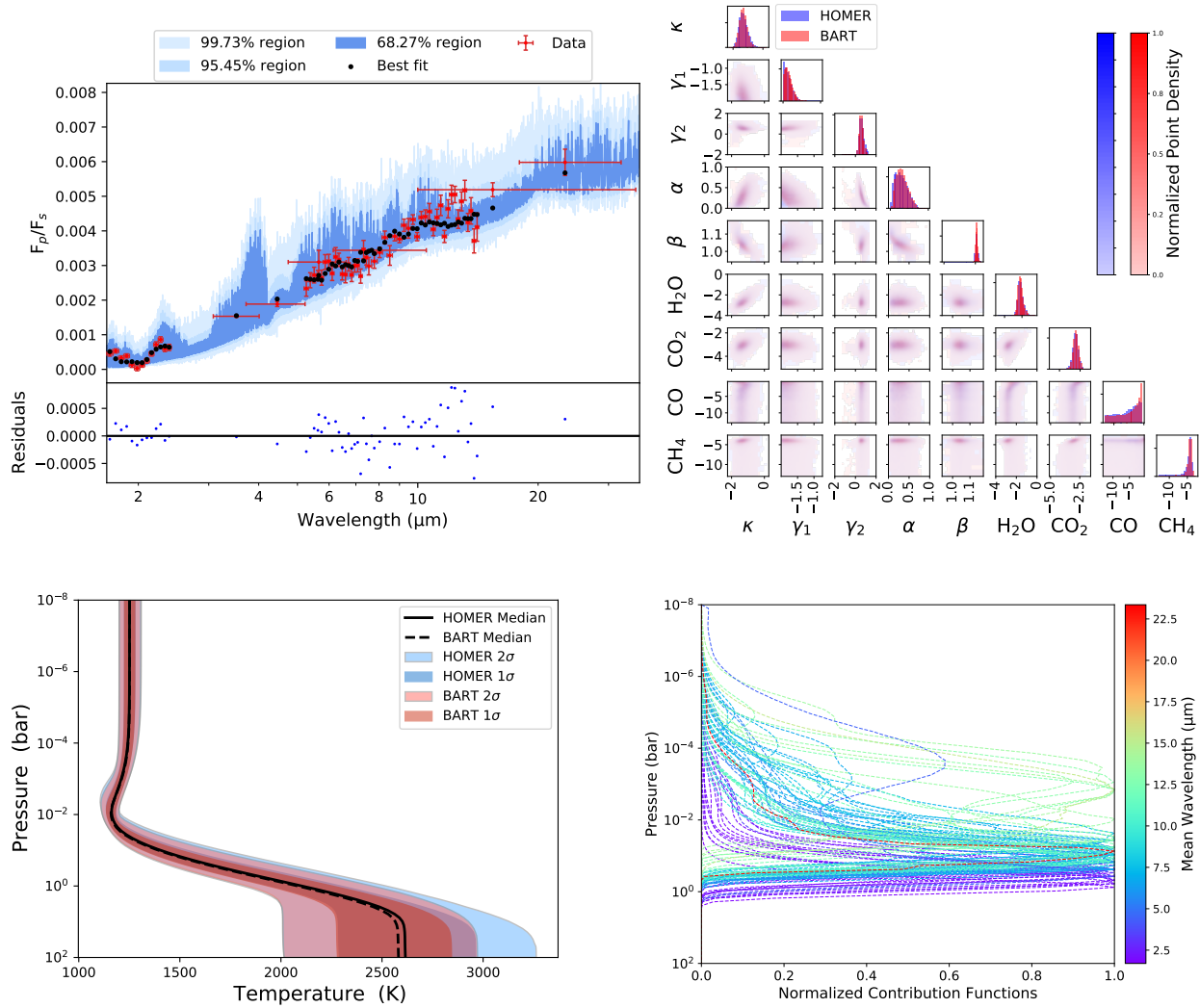


Figure 5.3: Comparisons between HOMER and BART posteriors. **Top left:** best-fit spectrum of HOMER, with  $1\sigma$ ,  $2\sigma$ , and  $3\sigma$  regions. **Top right:** normalized probability density functions of the 2D marginalized pairwise posteriors retrieved for HD 189733 b, with the 1D marginalized posteriors along the diagonal. The purple color arises from the close match between HOMER and BART. **Bottom left:** posterior median,  $1\sigma$  from the median, and  $2\sigma$  from the median  $T(p)$  profile. In the regions with sensitivity, HOMER closely matches BART, with a slightly greater uncertainty. **Bottom right:** normalized contribution functions, which show the pressure range each filter probes, for the best-fit BART model.

#### 5.4.2 Compute Cost

The performance differences between HOMER and BART highlight HOMER’s computational benefits. For a single Markov chain iteration, BART requires around 1.8 seconds per parallel chain on an AMD EPYC 7402P CPU, and multiple chains parallelize linearly across CPUs. By comparison, a single iteration with HOMER on the same CPU — which includes preprocessing (e.g., input normalization), prediction, post-processing (e.g., output denormalization, scaling according to the stellar spectrum), and band-integration — requires just  $\sim 0.2$  s for any number of chains fewer than 32. For a single chain, this is thus a  $\sim 9\times$  speedup. In our setup, we considered 10 parallel chains, translating to a  $\sim 90\times$  speedup for the function evaluated at each step of the Markov chain. Using an NVIDIA Titan Xp for predictions, the model evaluations at each Markov chain step require 0.01–0.02 seconds, a 10–20 $\times$  speedup over predictions with the aforementioned CPU and, when using 10 parallel chains, a 900–1800 $\times$  speedup over the same function evaluation in BART. We note that if BART were capable of utilizing a GPU, this speedup factor would be much less. Further investigation is necessary to determine whether HOMER offers speed improvements over a GPU-accelerated RT code. Nevertheless, the CPU results emphasize the speed improvements of our approach; the significant reduction in compute time enables retrievals to be executed on an average laptop. We note that the memory footprints for both approaches were comparable, though the parameters of each approach can strongly affect the required memory.

Here, the upfront compute cost to generate a data set and train a MARGE model is greater than the time to execute a single BART retrieval. In our example, we generated around 1.5–2 $\times$  the number of spectra typically computed during a BART retrieval with small credible region uncertainties, plus a few dozen hours to train the neural network. However, additional retrievals within the trained parameter space execute in around 30 minutes on our GPU (less when neglecting to compute spectral quantiles). Thus, when carrying out even two retrievals within some shared phase space, the

compute cost of MARGE+HOMER is less than two classical retrievals. In certain circumstances, such as where the radius and mass of the planet do not need to be varied (e.g., retrievals on different data sets of the same exoplanet), the number of spectra required to approximate the phase space accurately would be less than in our example, which may lead to MARGE+HOMER requiring less compute time than a single BART retrieval. Beyond the scope of retrieval, this approach could also provide a benefit to situations where it is advantageous to trade one set of computing resources for another. For example, spaceflight missions may be limited by thermal, power, and/or on-board computational resources; it may be advantageous to increase the total compute time, if it can decrease the power, thermal, and/or on-board computing required for the calculation.

Another benefit of our approach is that the compute-cost scaling is less than linear: increasing from 10 to 256 chains results in just a  $\sim 12.5\times$  increase in compute cost per iteration when using a GPU, compared to  $25.6\times$  as much for BART. Additional chains enable faster exploration of the parameter space, and, if executed for the same number of iterations per chain, increases the ESS, which reduces the uncertainty in the bounds of credible regions (Harrington et al. 2022). Thus, the combination of MARGE and HOMER saves valuable compute resources when performing retrievals and reduces total runtime when performing multiple retrievals.

## 5.5 Conclusions

This paper presents a novel technique for ML retrieval that uses a neural-network model of RT within a Bayesian framework to reduce the runtime of a retrieval. Our open-source codes, MARGE and HOMER, provide the community with an easy-to-use implementation of this approach, and they are readily applicable to any forward model and its inversion — not strictly BART or even RT. They are available on Github with full user documentation.

Our method enables fast retrievals that are consistent with algorithms that solve the RT equation. The approach circumvents limitations of current ML retrieval models by using an RT surrogate in place of the RT code found in classical retrieval algorithms, thereby preserving the accuracy of the Bayesian inference. Like BART, MARGE and HOMER work at both the low resolutions of Spitzer and the high spectral resolutions of advanced ground-based spectrographs.

On our hardware, HOMER reduces the runtime of each MCMC iteration by  $\sim 9\times$  per parallel chain using a CPU and  $90\text{--}180\times$  per chain using a GPU, compared to BART. For the case of HD 189733 b, the Bhattacharyya coefficients of the 1D marginalized posteriors of BART and HOMER are  $>0.984$ , indicating a close match. This reduction in compute time enables using more realistic (and computationally expensive) RT models, such as those including scattering and condensates. Additionally, 3D retrievals with  $\sim 200$  cells could be completed in a matter of days.

Our approach is particularly well suited to planning studies for future observations, telescopes, and instruments, like the James Webb Space Telescope and the Large UltraViolet Optical InfraRed Surveyor (e.g., Rocchetto et al. 2016, Feng et al. 2018). Using a single MARGE model trained over the desired parameter space, HOMER can perform dozens to hundreds of retrievals in the time it takes to run a single retrieval with an RT solver.

More generally, our technique and tools can be applied to problems beyond the scope of this investigation. MARGE provides a generalized method to train a neural network to model any deterministic process, while HOMER uses a MARGE-trained model to infer the inverse process. MARGE models could be trained for cloud/haze formation or photochemistry within general circulation models, for example. MARGE and HOMER could also be used to map gravitationally lensed galaxies (e.g., Perreault Levasseur et al. 2017).

With the plethora of ML retrieval algorithms that have emerged in recent years, standard data sets should be created and used for benchmarking. Ideally, such a data set would cover a wide range of wavelengths at high resolution and include all available opacity sources, scattering, clouds/hazes, and, in the case of terrestrial planets, surface properties. This would allow easy comparisons among current and future ML retrieval codes.

The Reproducible Research Compendium for this work is available for download<sup>6</sup>. It includes all of the code, configuration files, data, and plots used in support of this work.

## 5.6 Acknowledgements

We thank Jon Malkin, Lee Rhodes, and Edo Liberty for valuable contributions to the streaming quantiles method used in this work. We thank James Mang and Nicholas Susemihl for useful feedback on the software developed for this work. We thank Michael Lund and the NASA Exoplanet Archive for preparing and hosting the online RRC. We thank Jennifer Adams for helpful discussions on radiative transfer emulation in Earth science. We also thank contributors to the Datasketches library, NumPy, SciPy, Matplotlib, Tensorflow, Keras, the Python Programming Language, the free and open-source community, and the NASA Astrophysics Data System for software and services. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. This research was supported by the NASA Fellowship Activity under NASA Grant 80NSSC20K0682 and NASA Exoplanets Research Program grant NNX17AB62G. We thank FDL (<http://www.frontierdevelopmentlab.org/>) and SETI (<https://www.seti.org>) for making this collaboration possible.

---

<sup>6</sup>Available at <https://exoplanetarchive.ipac.caltech.edu/docs/marge-homer.html>

## 5.7 List of References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. 2016, in OSDI, Vol. 16, 265–283
- Abel, M., Frommhold, L., Li, X., & Hunt, K. L. C. 2012, *The Journal of Chemical Physics*, 136, 044319
- Agol, E., Cowan, N. B., Knutson, H. A., Deming, D., Steffen, J. H., Henry, G. W., & Charbonneau, D. 2010, *ApJ*, 721, 1861
- Aherne, F. J., Thacker, N. A., & Rockett, P. I. 1998, *Kybernetika*, 34, 363
- Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., & Zhu, Z. 2016, in *Proceedings of Machine Learning Research*, Vol. 48, *Proceedings of The 33rd International Conference on Machine Learning*, ed. M. F. Balcan & K. Q. Weinberger (New York, New York, USA: PMLR), 173–182
- Arcangeli, J., Désert, J.-M., Line, M. R., Bean, J. L., Parmentier, V., Stevenson, K. B., Kreidberg, L., Fortney, J. J., Mansfield, M., & Showman, A. P. 2018, *ApJ*, 855, L30
- Atzberger, C. 2004, *Remote Sensing of Environment*, 93, 53



- Bahdanau, D., Cho, K., & Bengio, Y. 2015, in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, ed. Y. Bengio & Y. LeCun
- Barber, R. J., Tennyson, J., Harris, G. J., & Tolchenov, R. N. 2006, MNRAS, 368, 1087
- Baydin, A. G., Heinrich, L., Bhimji, W., Shao, L., Naderiparizi, S., Munk, A., Liu, J., Gram-Hansen, B., Louppe, G., Meadows, L., Torr, P., Lee, V., Prabhat, Cranmer, K., & Wood, F. 2019, in Advances in Neural Information Processing Systems 33
- Bhattacharyya, A. 1943, Bull. Calcutta Math. Soc., 35, 99
- Blecic, J., Harrington, J., Cubillos, P. E., Bowman, M. O., Rojo, P. M., Stemm, M., Challener, R. C., Himes, M. D., Foster, A. J., Dobbs-Dixon, I., Foster, A. S. D., Lust, N. B., Blumenthal, S. D., Bruce, D., & Loredó, T. J. 2022, The Planetary Science Journal, 3, 82
- Borysow, A. 2002, A&A, 390, 779
- Borysow, A., Jorgensen, U. G., & Fu, Y. 2001, J. Quant. Spec. Radiat. Transf., 68, 235
- Boudon, V., Pirali, O., Roy, P., Brubach, J.-B., Manceron, L., & Auwera], J. V. 2010, Journal of Quantitative Spectroscopy and Radiative Transfer, 111, 1117 , special Issue Dedicated to Laurence S. Rothman on the Occasion of his 70th Birthday.
- Brehmer, J., Cranmer, K., Louppe, G., & Pavez, J. 2018, Phys. Rev. D, 98, 052004
- Brown, L., Sung, K., Benner, D., Devi, V., Boudon, V., Gabard, T., Wenger, C., Campargue, A., Leshchishina, O., Kassi, S., Mondelain, D., Wang, L., Daumont, L., Rgaglia, L., Rey, M., Thomas, X., Tyuterev, V. G., Lyulin, O., Nikitin, A., Niederer, H., Albert, S., Bauerecker, S., Quack, M., O'Brien, J., Gordon, I., Rothman, L., Sasada, H., Coustenis, A., Smith, M., Carrington, T., Wang, X.-G., Mantz, A., & Spickler, P. 2013, Journal of Quantitative Spectroscopy and Radiative Transfer, 130, 201 , hITRAN2012 special issue

- Bue, B. D., Thompson, D. R., Deshpande, S., Eastwood, M., Green, R. O., Natraj, V., Mullen, T., & Parente, M. 2019, *Atmospheric Measurement Techniques*, 12, 2567
- Campargue, A., Leshchishina, O., Wang, L., Mondelain, D., & Kassi, S. 2013, *Journal of Molecular Spectroscopy*, 291, 16, methane spectroscopy and its applications to planetary atmospheres, including the Earth's
- Castelli, F. & Kurucz, R. L. 2003, in *IAU Symposium*, Vol. 210, *Modelling of Stellar Atmospheres*, ed. N. Piskunov, W. W. Weiss, & D. F. Gray, A20
- Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. 2016, in *ICASSP*
- Charbonneau, D., Knutson, H. A., Barman, T., Allen, L. E., Mayor, M., Megeath, S. T., Queloz, D., & Udry, S. 2008, *ApJ*, 686, 1341
- Chernetskiy, M., Gobron, N., Gmez-Dans, J., Morgan, O., Disney, M., Lewis, P., & Schmullius, C. 2018, *Advances in Space Research*, 62, 1654
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. 2014, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar: Association for Computational Linguistics), 1724–1734
- Chollet, F. et al. 2015, *Keras*, <https://github.com/keras-team/keras>
- Chorowski, J., Bahdanau, D., Cho, K., & Bengio, Y. 2014, in *NIPS 2014 Workshop on Deep Learning*, December 2014
- Cobb, A. D., Himes, M. D., Soboczenski, F., Zorzan, S., O'Beirne, M. D., Güneş Baydin, A., Gal, Y., Domagal-Goldman, S. D., Arney, G. N., Angerhausen, D., & 2018 NASA FDL Astrobiology Team, I. 2019, *AJ*, 158, 33
- Cranmer, K., Brehmer, J., & Louppe, G. 2019, arXiv preprint arXiv:1911.01429

- Cubillos, P. E., Harrington, J., Blečić, J., Himes, M. D., Rojo, P. M., Loredó, T. J., Lust, N. B., Challener, R. C., Foster, A. J., Stemm, M. M., Foster, A. S. D., & Blumenthal, S. D. 2022, *The Planetary Science Journal*, 3, 81
- Daumont, L., Nikitin, A., Thomas, X., Rgaglia, L., der Heyden], P. V., Tyuterev, V., Rey, M., Boudon, V., Wenger, C., Lote, M., & Brown, L. 2013, *Journal of Quantitative Spectroscopy and Radiative Transfer*, 116, 101
- Deming, L. D. & Seager, S. 2017, *Journal of Geophysical Research (Planets)*, 122, 53
- Feng, Y. K., Robinson, T. D., Fortney, J. J., Lupu, R. E., Marley, M. S., Lewis, N. K., Macintosh, B., & Line, M. R. 2018, *AJ*, 155, 200
- Garcia-Cuesta, E., de la Torre, F., & de Castro, A. J. 2009, in *Advances in Computational Algorithms and Data Analysis*, ed. S.-I. Ao, B. Rieger, & S.-S. Chen (Dordrecht: Springer Netherlands), 319–331
- Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. 2017, in *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML17 (JMLR.org)*, 12631272
- Goodfellow, I., Bengio, Y., & Courville, A. 2016, *Deep Learning* (MIT Press)
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. 2014, in *Advances in Neural Information Processing Systems*, 2672–2680
- Goorvitch, D. 1994, *ApJS*, 95, 535
- Grillmair, C. J., Burrows, A., Charbonneau, D., Armus, L., Stauffer, J., Meadows, V., van Cleve, J., von Braun, K., & Levine, D. 2008, *Nature*, 456, 767
- Gmez-Dans, J., Lewis, P., & Disney, M. 2016, *Remote Sensing*, 8, 119

- Hargreaves, R. J., Gordon, I. E., Rey, M., Nikitin, A. V., Tyuterev, V. G., Kochanov, R. V., & Rothman, L. S. 2020, *ApJS*, 247, 55
- Harrington, J., Himes, M. D., Cubillos, P. E., Blečić, J., Rojo, P. M., Challener, R. C., Lust, N. B., Bowman, M. O., Blumenthal, S. D., Dobbs-Dixon, I., Foster, A. S. D., Foster, A. J., Green, M. R., Loredó, T. J., McIntyre, K. J., Stemm, M. M., & Wright, D. C. 2022, *The Planetary Science Journal*, 3, 80
- Hayes, J. J. C., Kerins, E., Awiphan, S., McDonald, I., Morgan, J. S., Chuanraksasat, P., Komonjinda, S., Sanguansak, N., & Kittara, P. 2019, arXiv e-prints, arXiv:1909.00718
- He, K., Zhang, X., Ren, S., & Sun, J. 2016, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. 2017, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269
- Johnsen, T. K. & Marley, M. S. 2019, arXiv e-prints, arXiv:1911.03997
- Karnin, Z., Lang, K., & Liberty, E. 2016, in *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, 71–78
- Kasim, M. F., Watson-Parris, D., Deaconu, L., Oliver, S., Hatfield, P., Froula, D. H., Gregori, G., Jarvis, M., Khatiwala, S., Korenaga, J., Topp-Mugglestone, J., Viezzer, E., & Vinko, S. M. 2021, *Machine Learning: Science and Technology*, 3, 015013
- Koskinen, T. T., Moses, J. I., West, R. A., Guerlet, S., & Jouchoux, A. 2016, *Geophys. Res. Lett.*, 43, 7895
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, in *Advances in neural information processing systems*, 1097–1105

- Line, M. R., Wolf, A. S., Zhang, X., Knutson, H., Kammer, J. A., Ellison, E., Deroo, P., Crisp, D., & Yung, Y. L. 2013, *ApJ*, 775, 137
- Madhusudhan, N. 2018, in *Handbook of Exoplanets*, ed. H. J. Deeg & J. A. Belmonte (Springer International Publishing AG), 104
- Márquez-Neila, P., Fisher, C., Sznitman, R., & Heng, K. 2018, *Nature Astronomy*, 2, 719
- Munk, A., Ścibior, A., Baydin, A. G., Stewart, A., Fernlund, G., Poursartip, A., & Wood, F. 2019, arXiv preprint arXiv:1910.11950
- Niederer, H.-M., Albert, S., Bauerecker, S., Boudon, V., Champion, J.-P., & Quack, M. 2008, *CHIMIA International Journal for Chemistry*, 62, 273
- Niederer, H.-M., Wang, X.-G., Carrington, T., Albert, S., Bauerecker, S., Boudon, V., & Quack, M. 2013, *Journal of Molecular Spectroscopy*, 291, 33 , methane spectroscopy and its applications to planetary atmospheres, including the Earth's
- Nikitin, A., Brown, L., Sung, K., Rey, M., Tyuterev, V., Smith, M., & Mantz, A. 2013, *Journal of Quantitative Spectroscopy and Radiative Transfer*, 114, 1
- Nikitin, A., Daumont, L., Thomas, X., Rgalia, L., Rey, M., Tyuterev, V., & Brown, L. 2011, *Journal of Molecular Spectroscopy*, 268, 93 , philip R. Bunker and A. Robert W. McKellar
- Nikitin, A., Lyulin, O., Mikhailenko, S., Perevalov, V., Filippov, N., Grigoriev, I., Morino, I., Yokota, T., Kumazawa, R., & Watanabe, T. 2010, *Journal of Quantitative Spectroscopy and Radiative Transfer*, 111, 2211 , xVIth Symposium on High Resolution Molecular Spectroscopy (HighRus-2009)
- Perreault Levasseur, L., Hezaveh, Y. D., & Wechsler, R. H. 2017, *ApJ*, 850, L7

- Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. 2016, in 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, ed. Y. Bengio & Y. LeCun
- Richard, C., Gordon, I., Rothman, L., Abel, M., Frommhold, L., Gustafsson, M., Hartmann, J.-M., Hermans, C., Lafferty, W., Orton, G., Smith, K., & Tran, H. 2012, *Journal of Quantitative Spectroscopy and Radiative Transfer*, 113, 1276 , three Leaders in Spectroscopy
- Rivera, J., Verrelst, J., Gmez-Dans, J., Muoz-Mar, J., Moreno, J., & Camps-Valls, G. 2015, *Remote Sensing*, 7, 93479370
- Rocchetto, M., Waldmann, I. P., Venot, O., Lagage, P. O., & Tinetti, G. 2016, *ApJ*, 833, 120
- Rothman, L. S., Gordon, I. E., Babikov, Y., Barbe, A., Benner, D. C., Bernath, P. F., Birk, M., Bizzocchi, L., Boudon, V., Brown, L. R., et al. 2013, *J. Quant. Spec. Radiat. Transf.*, 130, 4
- Rothman, L. S., Gordon, I. E., Barber, R. J., Dothe, H., Gamache, R. R., Goldman, A., Perevalov, V. I., Tashkun, S. A., & Tennyson, J. 2010, *J. Quant. Spec. Radiat. Transf.*, 111, 2139
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, *Nature*, 323, 533
- Seager, S. & Deming, D. 2010, *ARA&A*, 48, 631
- Sennrich, R., Haddow, B., & Birch, A. 2016, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Berlin, Germany: Association for Computational Linguistics), 1715–1725
- Simonyan, K. & Zisserman, A. 2015, in International Conference on Learning Representations
- Smith, L. N. 2015, arXiv e-prints, arXiv:1506.01186

- Soboczenski, F., Himes, M. D., O'Beirne, M. D., Zorzan, S., Gunes Baydin, A., Cobb, A. D., Gal, Y., Angerhausen, D., Mascaro, M., Arney, G. N., & Domagal-Goldman, S. D. 2018, arXiv e-prints, arXiv:1811.03390
- Swain, M. R., Vasisht, G., Tinetti, G., Bouwman, J., Chen, P., Yung, Y., Deming, D., & Deroo, P. 2009, *ApJ*, 690, L114
- Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. 2015, in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1–9
- Tashkun, S., Perevalov, V., Teffo, J.-L., Bykov, A., & Lavrentieva, N. 2003, *Journal of Quantitative Spectroscopy and Radiative Transfer*, 82, 165, the HITRAN Molecular Spectroscopic Database: Edition of 2000 Including Updates of 2001.
- ter Braak, C. J. F. & Vrugt, J. A. 2008, *Statistics and Computing*, 18, 435
- Verrelst, J., Dethier, S., Rivera, J. P., Munoz-Mari, J., Camps-Valls, G., & Moreno, J. 2016, *IEEE Geoscience and Remote Sensing Letters*, 13, 1012
- Verrelst, J., Rivera, J. P., Gmez-Dans, J., Camps-Valls, G., & Moreno, J. 2015, in 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 633–636
- Verrelst, J., Rivera Caicedo, J., Muoz-Mar, J., Camps-Valls, G., & Moreno, J. 2017, *Remote Sensing*, 9, 927
- Vicent, J., Verrelst, J., Rivera-Caicedo, J. P., Sabater, N., Muoz-Mar, J., Camps-Valls, G., & Moreno, J. 2018, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11, 4918
- Villanueva, G. L., Smith, M. D., Protopapa, S., Faggi, S., & Mandell, A. M. 2018, *J. Quant. Spec. Radiat. Transf.*, 217, 86

Waldmann, I. P. & Griffith, C. A. 2019, *Nature Astronomy*, 3, 620

Welford, B. P. 1962, *Technometrics*, 4, 419

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. 2016, arXiv preprint arXiv:1609.08144

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., & Zweig, G. 2016, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP

Yin, F., Gomez-Dans, J., & Lewis, P. 2018, in *2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 1804–1807

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



**CHAPTER 6: TOWARDS 3D RETRIEVAL OF EXOPLANET  
ATMOSPHERES: ASSESSING THERMOCHEMICAL EQUILIBRIUM  
ESTIMATION METHODS**

**Michael D. Himes<sup>1</sup>, Joseph Harrington<sup>2</sup>**

<sup>1</sup> *Planetary Sciences Group, Department of Physics, University of Central Florida, Orlando, FL 32816, USA*

<sup>2</sup> *Planetary Sciences Group, Department of Physics and Florida Space Institute, University of Central Florida, Orlando, FL 32816, USA*

In preparation for submission to *The Planetary Science Journal*.

## 6.1 Abstract

Characterizing exoplanetary atmospheres via Bayesian retrievals requires assuming some chemistry model, such as thermochemical equilibrium or constant-with-altitude abundances. The higher-resolution data offered by upcoming telescopes enables more complex chemistry models within retrieval frameworks. Yet, many chemistry codes that model more complex processes like vertical transport are computationally expensive, and directly incorporating them into a 1D retrieval model can result in prohibitively long execution times. Additionally, phase curve observations with upcoming telescopes motivate 2D and 3D retrieval models, further exacerbating the lengthy runtime for retrieval frameworks with complex chemistry models. Here, we compare thermochemical equilibrium approximation methods based on their speed and accuracy with respect to a Gibbs energy-minimization code. We find that, while all methods offer orders of magnitude reductions in computational cost, neural network surrogate models perform more accurately than the other approaches considered, achieving a median absolute dex error  $< 0.03$  for the phase space considered. Our findings suggest that more complex and computationally expensive chemistry models could be incorporated into retrieval models via this surrogate modeling approach.

## 6.2 Introduction

Improvements in signal-to-noise ratio and spectral resolution of exoplanet observations, such as those offered by the James Webb Space Telescope (JWST), motivate and enable the use of more sophisticated models to characterize exoplanetary atmospheres via retrieval (see review by Madhusudhan 2018). Early retrieval studies made simplifying assumptions about atmospheric chemistry, such as gas abundances that were constant with altitude or by scaling equilibrium profiles calculated assuming solar metallicity (e.g., Madhusudhan & Seager 2009, Line et al. 2014). Later,

as new data began to offer a more detailed picture of exoplanet atmospheres, groups incorporated more complex chemistry models into retrieval frameworks, such as equilibrium abundances for sub- and super-solar metallicities (e.g., Oreshenko et al. 2017). Many exoplanet atmospheres are likely in disequilibrium (Moses et al. 2011, Line & Yung 2013, Moses et al. 2013b, Venot et al. 2015, Roudier et al. 2021). Stevenson et al. (2010) presented the first detection of disequilibrium chemistry in an exoplanet atmosphere, finding  $\text{CH}_4$  depleted relative to thermochemical equilibrium calculations. Other more recent studies considering disequilibrium processes, such as photochemistry and vertical quenching, have found evidence of disequilibrium in multiple hot Jupiter atmospheres (e.g., Mollière et al. 2020, Kawashima & Min 2021, Roudier et al. 2021). While these recent studies still find some equilibrium models consistent with the data, JWST will provide sufficient precision to more definitively differentiate between equilibrium and disequilibrium atmospheres (Blumenthal et al. 2018).

Both the more detailed measurements of JWST and future telescopes as well as recent results from 1D retrieval studies motivate more complex retrieval models. While the findings of Blečić et al. (2017) show that 1D retrieval models can recover a thermal profile comparable to the 3D structure's arithmetic average, Caldas et al. (2019) and Pluriel et al. (2022) found biases in the retrieved gas abundances. A more complete understanding of the atmospheric chemistry thus requires a 2D or 3D model to properly capture longitudinal variations. Yet, that increase in dimensionality is tied to an increase in computational cost. With 1D retrievals requiring on the order of  $10^5$ – $10^6$  forward model evaluations (Madhusudhan 2018), the difference of one second per forward model evaluation adds up to days of computing time, and higher-dimensional models will multiply this further. While any chemistry model could be incorporated into a retrieval framework, the lengthy execution time of many models would be prohibitive.

Recently, Himes et al. (2022) presented a new machine learning approach to atmospheric retrieval, where the radiative transfer forward model is replaced with a neural network (NN) surrogate model. They found that this approach achieves similar quantitative results as the classical approach, but at a fraction of the computational cost. Similar results have been reported for other scientific problems using similar approaches (e.g., Gilmer et al. 2017, Brehmer et al. 2018, Baydin et al. 2019, Munk et al. 2019, Kasim et al. 2021).

In this study, we compare thermochemical equilibrium estimation methods and consider their applicability to 3D retrieval based on runtime and accuracy. In Section 6.3, we describe the analytical, interpolation-based, and NN-based models considered and detail our methodology. In Section 6.4, we present and discuss the results. Finally, we draw conclusions in Section 6.5.

### 6.3 Methods

We utilize four equilibrium estimation methods: minimization of the Gibbs free energy via the Thermochemical Equilibrium Abundances (TEA) code (Blecic et al. 2016); the analytical approximation for equilibrium used in the Reliable Analytic Thermochemical Equilibrium (RATE) code (Cubillos et al. 2019); interpolation within a grid of models produced by TEA; and a surrogate model based on an NN trained on data produced by TEA.

### 6.3.1 *Equilibrium via Gibbs Energy Minimization*

Blecic et al. (2016) presented the open-source TEA code, which calculates thermochemical equilibrium via minimizing the Gibbs free energy in a Lagrangian optimization framework. They demonstrate that TEA reproduces the results of other thermochemical equilibrium implementations when utilizing the same thermodynamic data. Here, we use TEA to calculate the “ground truth” data to compare with the other methods as well as the data set used for interpolation and NN surrogate model training. For more details on TEA’s implementation and validation, see Blecic et al. (2016).

### 6.3.2 *Equilibrium via Analytical Formulae*

Cubillos et al. (2019) presented the open-source RATE code, an analytical formalism to approximate thermochemical equilibrium, which built upon and resolved the instability issues present in the formalism of Heng et al. (2016), Heng & Lyons (2016), and Heng & Tsai (2016). They determined that the RATE approximation is valid over a parameter space of roughly 200–2000 K,  $10^{-8}$ – $10^3$  bar, and  $10^{-3}$ – $10^2 \times$  solar elemental abundances. These stability improvements enable broad application to arbitrary combinations of parameters within this phase space, such as those considered in a Bayesian retrieval on optical and/or infrared spectra of most observed exoplanetary atmospheres. For more details on their approach, see Cubillos et al. (2019).

Table 6.1: TEA Model Grid Parameters

Parameter	Minimum	Maximum	Number of bins
log pressure	-8	3	100
log temperature	2.305	3.778	80
log C/H	-6.57	-0.57	21
log N/H	-7.17	-1.17	21
log O/H	-6.31	-0.31	21

### 6.3.3 Equilibrium via Interpolation

We generate a grid of 74,088,000 points based on pressure, temperature, C/H, N/H, and O/H; the parameter minima, maxima, and number of uniformly log-spaced bins are given in Table 6.1. We consider the same molecules as Cubillos et al. (2019), with the addition of helium: H<sub>2</sub>O, CO, CH<sub>4</sub>, CO<sub>2</sub>, HCN, C<sub>2</sub>H<sub>2</sub>, C<sub>2</sub>H<sub>4</sub>, NH<sub>3</sub>, N<sub>2</sub>, H<sub>2</sub>, H, and He. This enables direct applicability of the models to gas-giant atmospheres.

We consider linear and inverse-distance weighting (IDW) interpolation. For IDW, we calculate the Euclidean distance between the logarithm of the inputs, and we vary the exponent on the inverse distance,  $p$ , between 1 and 40. To interpolate, we consider the  $n$  nearest neighbors along each axis, where  $n$  can vary between 1 and 4. In the interest of the ability to scale the considered models to higher dimensionalities, we do not consider spline interpolation due to longer runtimes than TEA in this setup, likely attributable to the inability to hold this data set in cached memory. Similarly, we do not consider radial basis function interpolation due to the amount of memory required for the necessary  $N \times M \times M$  matrices, where  $N$  is the number of dimensions and  $M$  is the number of data points. For this problem’s dimensionality, using with a data set of 100,000 points — around  $\sim 0.13\%$  of the total TEA data set considered here — would require around 372.5 GB of memory to calculate the radial basis functions in double precision.

### 6.3.4 *Equilibrium via NN-based Surrogate Model*

Himes et al. (2022) presented a NN-based surrogate modeling method along with a software package, MARGE, that implements the technique. Here, we use MARGE to approximate TEA.

We use two different data sets to train the NN models: (1) a grid of TEA models like described in the previous section but with only 40 temperatures and 11 elemental abundance ratios (5,324,000 total grid points), and (2) a set of models whose inputs were randomly drawn from uniform distributions. For (1), we randomly split the grid of TEA models into training, validation, and test sets, which contain 70%, 20%, and 10% of the total data, respectively. For (2), we make 125,000 random draws of temperatures, C/H, N/H, and O/H from uniform distributions with boundaries defined in Table 6.1. Each of those cases are computed over the log-uniform grid of pressures defined in Table 6.1, producing a total of 12,500,000 combinations of the 5 free parameters. We randomly split these cases such that 64%, 16%, and 20% of the cases are in the training, validation, and test sets, respectively.

Using these data sets, we train four NN models. NN1 is trained for 864 epochs on the grid of TEA models. NN2 is trained for 750 epochs on the random data. NN3 is trained for 500 epochs on the random data, but sized to match the gridded data set. NN4 is trained for 500 epochs on the combination of both the grid of TEA models as well as the random data. As in Himes et al. (2022), we find that the models indefinitely improve by minutia and therefore choose to stop training after the above numbers of epochs, as further training offers minimal practical improvement.

Each data case is comprised of the 5 input parameters given in Table 6.1 and the corresponding 12 output gas abundances. We pre-process the data by (1) logarithmically scaling inputs and outputs, (2) standardizing according to the training set mean and standard deviation, and (3) scaling the data to be within [-1, 1] based on the training set extrema.

The NN model consists of an input layer with 5 nodes corresponding to the 5 input parameters; a 1D convolutional layer with a kernel size of 3 and 256 feature maps, which uses a rectified linear unit (“ReLU”) activation function; 3 dense layers with 4096 nodes, each of which use ReLU activation functions; and an output dense layer with 12 nodes corresponding to the output gas abundances. This architecture was selected through an extensive model grid search. We train using a batch size of 1024, the mean-squared-error loss on the validation set, the Adam optimizer, and the `triangular2` learning rate policy (Smith 2015, Himes et al. 2022) with a learning rate cycle of 12 epochs and range determined from the range test (see Appendix A of Himes et al. 2022).

In training the model to predict equilibrium abundances for a given pressure, temperature, C/H, N/H, and O/H, this NN surrogate model can be applied to arbitrary pressure–temperature profiles, and the elemental abundance ratios are not required to be uniform. To apply this NN surrogate model to some pressure–temperature profile with some assumed elemental abundance ratios, we normalize the inputs as described above, make the predictions, and denormalize the outputs. We convolve each denormalized output gas abundance profile with a 1D Gaussian filter with a kernel standard deviation of 1 index ( $\sim 0.11$  dex, in log-pressure) to smooth out minor model inaccuracies, as Himes et al. (2022) found that the residuals for RT NN surrogate models were typically distributed about 0.



### 6.3.5 Performance Assessment

We treat TEA as the control to compare the other approaches against. We assess each non-TEA model by computing (1) the root-mean-square error (RMSE), coefficient of determination ( $R^2$ ), and absolute dex difference over a grid of temperatures, pressures, metallicities, and C/O, and (2) the factor speedup compared to TEA. For (1), we use a grid similar to that used in Cubillos et al. (2019), except with the temperature and metallicity grids shifted by a half cell. This results in metallicities spanning  $-1.5 - 2.5$  dex and temperatures spanning  $221 - 5485$  K. We also consider the original grid used in Cubillos et al. (2019) to test the NN performance at the edge of the phase space.

## 6.4 Results & Discussion

Table 6.2 summarizes the median RMSE, median  $R^2$ , median absolute dex error, and speedup factor relative to TEA for the considered models over the grid of test cases. Note that for this setup, TEA required roughly  $11.7 \pm 1$  seconds to estimate equilibrium at a given temperature, C/H, N/H, and O/H over the 100 pressures considered. In general, we find that all of the considered models offer orders of magnitude reductions in computational cost compared to TEA, while still accurately approximating TEA over many regions of the phase space. Figures 6.1 – 6.12 visualize the model errors, as in Cubillos et al. (2019) but with a different grid of metallicities. Overall, we find that the NNs more accurately approximate TEA than the other models over the phase space considered. Visually, the results of NN2, NN3, and NN4 look similar; consequently, we omit NN4 from the figures here, though they are available in the online compendium (see Section 6.5).

Though the best linear interpolation model achieves the lowest median absolute dex error, its RMSE and  $R^2$  indicate it to be less accurate than the best IDW interpolation model as well as most of the NNs. This is confirmed via the mean absolute dex errors: the best linear model's average error is about 0.38 dex, while the worst of the NN models features an average error of about 0.33 dex (NN2 and NN3 are even less, around 0.13 dex). Thus, while the linear model performs well in many cases (as indicated by the low median dex error), the cases where it is inaccurate results in significant relative errors (as indicated by the greater average dex error). This is confirmed by inspecting the error plots, as the model performs accurately in many regions but features significant errors for carbon- and oxygen-bearing species for  $C/O = 0.9$ .

Based on the calculated metrics, the best-performing linear and IDW interpolation models are comparable. While linear interpolation is generally faster than IDW and achieves a lower median absolute dex error, IDW achieves a lower median RMSE and slightly higher median  $R^2$ . Despite this, both models achieve similar average absolute dex errors of  $\sim 0.38$ . Table 6.3 details the performance of the other linear interpolation models considered in this investigation, while Table 6.4 details the performance of the other IDW interpolation models considered. For simple linear interpolation, we find that increasing the grid density leads to improved model performance, as expected. For IDW, we generally find that the value of the exponent is anti-correlated with the absolute dex error. However, at large values for the exponent, overflow can occur when the tested point is near one of the grid nodes. We also find that, for a given exponent, the average absolute dex error is typically minimized when considering the 2 nearest neighbors along each axis. While considering only the nearest neighbor along each axis performs equivalently or only slightly worse than the  $n = 2$  case, considering 3 or more of the nearest neighbors along each axis performs

significantly worse than in the  $n = 2$  case. This is consistent with intuition: at  $n = 1$ , the interpolated result assumes the behavior is linear from the closest points, while at  $n = 2$  the interpolated result factors in the behavior of the closest points both above and below the target parameters. For  $n > 2$ , the additional neighboring points are farther away and therefore less likely to follow a linear relation with respect to the point of interest, resulting in reduced accuracy.

While RATE achieves a greater speedup factor than the NNs, it is less accurate than those NNs, as indicated by the other performance metrics. Additionally, the speedup factors are computed for the NN using the central processing unit (CPU); on our machine, using an Nvidia Titan Xp graphics processing unit (GPU) resulted in a  $\sim 4\times$  speedup compared to using the AMD EPYC 7402P CPU, nearly matching RATE's speedup factor. Further, when computing the speedup factors, we only considered the computation time of the NN for a single atmospheric model (100 pressure layers), which utilizes only a fraction of the GPU's resources. For applications where multiple atmospheric models can be calculated in parallel (e.g., a Bayesian retrieval with  $N$  parallel chains, or a multi-dimensional retrieval model), the NNs offer further speedup improvements, as their computational time scales less than linearly (Himes et al. 2022).

Among the NN models, we find that NN3 performs best, with NN2 and NN4 not far behind. NN1, which was trained on gridded data alone, performs worst among the NN models. Visually, NN1's error plots feature topography similar to the interpolation error plots. For example, in Figure 6.1, at  $C/O = 0.9$ , the interpolation plots feature a region in the temperature–pressure space with significant error. However, its errors in these regimes are generally less than the interpolation approaches despite being trained on a data grid  $\sim 7\%$  the size of the grid used for interpolation. These similar topographies suggest that the model grid is undersampled in this region. As  $C/O$  approaches 1, oxygen-dominated chemistry gives way to carbon-dominated chemistry (Madhusudhan 2012, Moses et al. 2013a). Thus, it is likely that the model error in this regime is due to nonlinear changes in the chemical abundances, preventing the linear model from capturing it accurately.

This is supported by the fact that the species lacking carbon and oxygen do not display this topography (Figures 6.8, 6.9, 6.10, 6.11, 6.12), while all carbon- and oxygen-bearing species bear this topography at varying magnitudes (Figures 6.1, 6.2, 6.3, 6.4, 6.5, 6.6, 6.7). Similarly, NN1’s poorer performance at  $M/H = 2.5$  compared to the other NNs is likely also attributable to undersampling in that regime.

The other NNs, which were trained on random data, do not feature this topography. NN3 was trained on a data set with  $\sim 7\%$  of the number of cases in the gridded data set used by the interpolation methods. Despite the same training data size as NN1 and training for fewer epochs, NN3 outperforms NN1 according to the performance metrics. This suggests that NN surrogate models more efficiently learn a problem when the training data is randomly generated, rather than generated on a fixed gridding. NN4 utilized both NN1’s gridded data and NN3’s random data—the most data out of any of the NN models. NN3 and NN4 were trained for the same number of epochs, yet despite the additional data considered by NN4 (which led to a greater number of training iterations than NN3), NN4 performs worse than both NN2 and NN3. NN4’s performance falls between NN1 and NN2, which, together with NN1’s performance, suggests that the gridded data force the NN towards a solution that, while optimal for that data, does not properly generalize to approximate thermochemical equilibrium in all regimes. Future studies should examine this in more detail to determine the optimal approach to generating data to train NN surrogate models.

We additionally considered the NN’s performance at the edge of the phase space; Figure 6.13 shows NN2’s and RATE’s predictions for the grid considered in Cubillos et al. (2019) for  $H_2O$  and  $CH_4$ . We find that, unsurprisingly, its accuracy diminishes near the edges of the phase space. At the metallicity extrema (3 orders of magnitude above and below solar metallicity), the NN’s error increases significantly, especially at the C/O extrema. By comparison, RATE’s accuracy at  $Z = -3$  is more or less consistent with less extreme metallicities, though it too becomes inaccurate at  $Z = 3$ , as reported in Cubillos et al. (2019). We attribute the NN’s behavior to its training data set: when

Table 6.2: Model Performance Comparison

Model	RMSE	$R^2$	Dex	Speedup Factor
RATE	1.3386	0.9302	0.1002	891
Linear	1.1131	0.9821	0.0037	8330
IDW	0.8314	0.9879	0.0696	2691
NN1	1.4942	0.9613	0.0316	206
NN2	0.4872	0.9967	0.0250	207
NN3	0.4390	0.9973	0.0280	207
NN4	0.6343	0.9950	0.0258	205

**Notes.** Here we present only the best-performing linear and IDW interpolation models. For each model listed, we present only the median RMSE,  $R^2$ , and absolute dex error for conciseness. For the full data, see the RRC download link at the end of Section 6.5. Note also that the speedup factors for the NNs are calculated for a single atmosphere using the CPU; multiple atmospheric models could be calculated simultaneously and/or a GPU could be used for calculations, increasing the speedup factor.

making random draws from the phase space, few samples will have multiple parameters at extrema (e.g., both  $Z = 3$  and  $C/O = 0.1$ ), resulting in reduced accuracy in that regime. Of course, this can be resolved by generating data over a larger phase space than needed or preferentially drawing samples that are in these regions, thus ensuring the NN can accurately predict thermochemical equilibrium in the regime of interest.

Table 6.3: Linear Interpolation Model Performance Comparison

Number of temperatures	Number of elemental abundances	RMSE	$R^2$	Dex	Speedup Factor
40	11	1.4937	0.9623	0.0251	8434
80	11	1.4719	0.9641	0.0127	8518
80	21	1.1131	0.9821	0.0037	8332

**Notes.** For each model listed, we present only the median RMSE,  $R^2$ , and absolute dex error for conciseness. For the full data, see the RRC download link at the end of Section 6.5.

Table 6.4: IDW Model Performance Comparison

Exponent	Neighbors per axis	RMSE	$R^2$	Dex	Speedup Factor
1	1	1.1341	0.9830	0.0706	15190
2	1	1.0780	0.9799	0.0619	15250
2	2	1.0795	0.9802	0.0650	2763
2	3	5.6737	0.3177	0.0809	886
3	1	1.0068	0.9803	0.0650	15000
5	1	0.8495	0.9840	0.0780	14920
5	2	0.8502	0.9839	0.0722	2716
5	3	5.6693	0.3243	0.0767	841
5	4	7.5990	-0.0859	0.0993	299
10	1	0.8301	0.9884	0.0781	14930
10	2	0.8194	0.9886	0.0721	2188
10	3	5.6713	0.3345	0.0741	835
10	4	7.6006	-0.0631	0.0996	299
20	1	0.8405	0.9877	0.0761	15160
20	2	0.8322	0.9879	0.0687	2710
20	3	5.6721	0.3344	0.0717	871
30	1	0.8392	0.9877	0.0761	15020
30	2	0.8314	0.9879	0.0696	2691
30	3	5.6721	0.3343	0.0748	868
40	1	0.8380	0.9876	0.0756	13680
40	2	0.8316	0.9878	0.0719	2689

**Notes.** For each model listed, we present only the median RMSE,  $R^2$ , and absolute dex error for conciseness. For the full data, see the RRC download link at the end of Section 6.5.

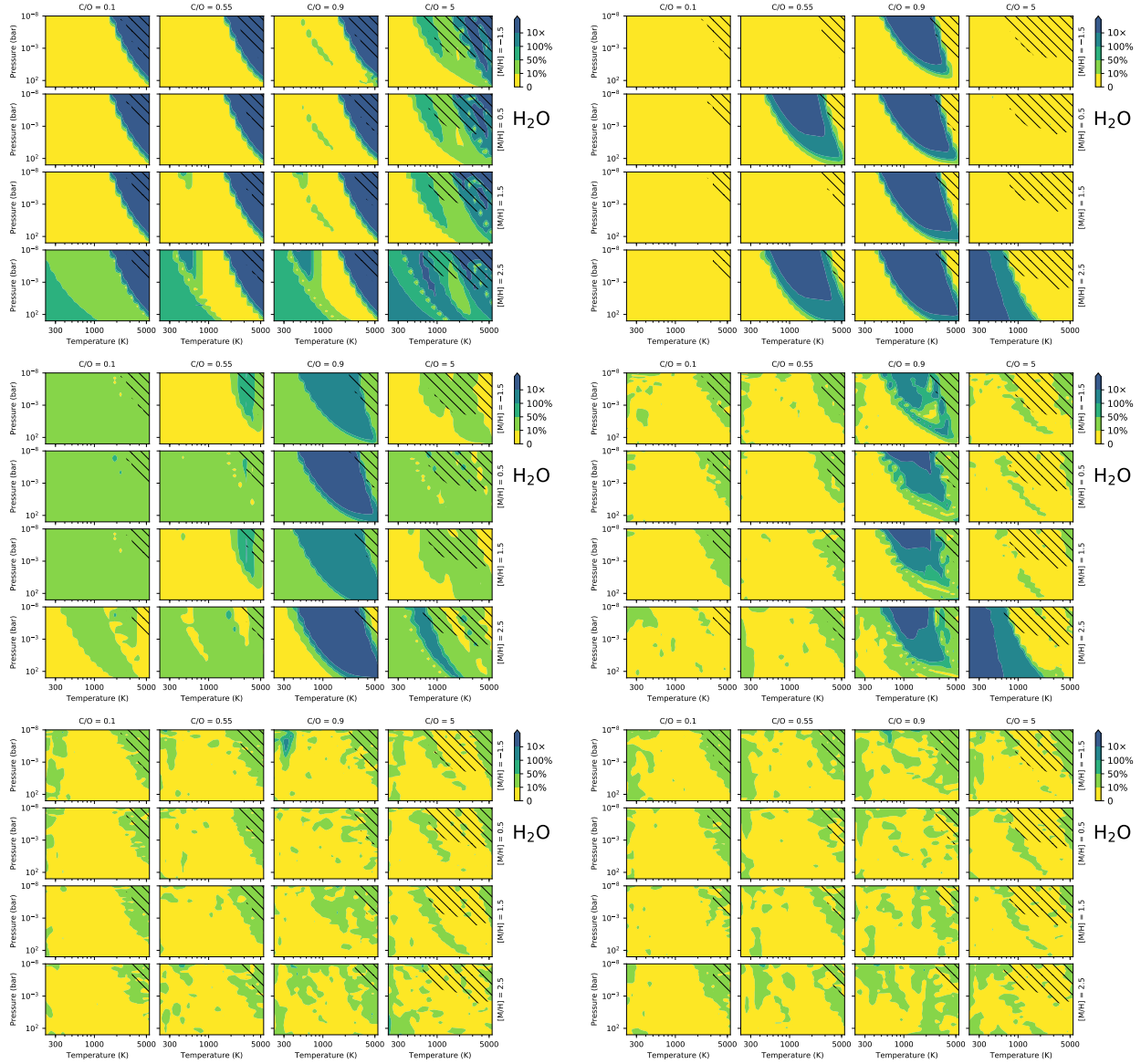


Figure 6.1: Performance comparison between the various models and TEA for  $\text{H}_2\text{O}$ . The black shading indicates regimes where the log abundance is  $\leq -10$ . **Top left:** RATE. **Top right:** Best-performing linear interpolation model. **Middle left:** Best-performing IDW interpolation model. **Middle right:** NN1. **Bottom left:** NN2. **Bottom right:** NN3.



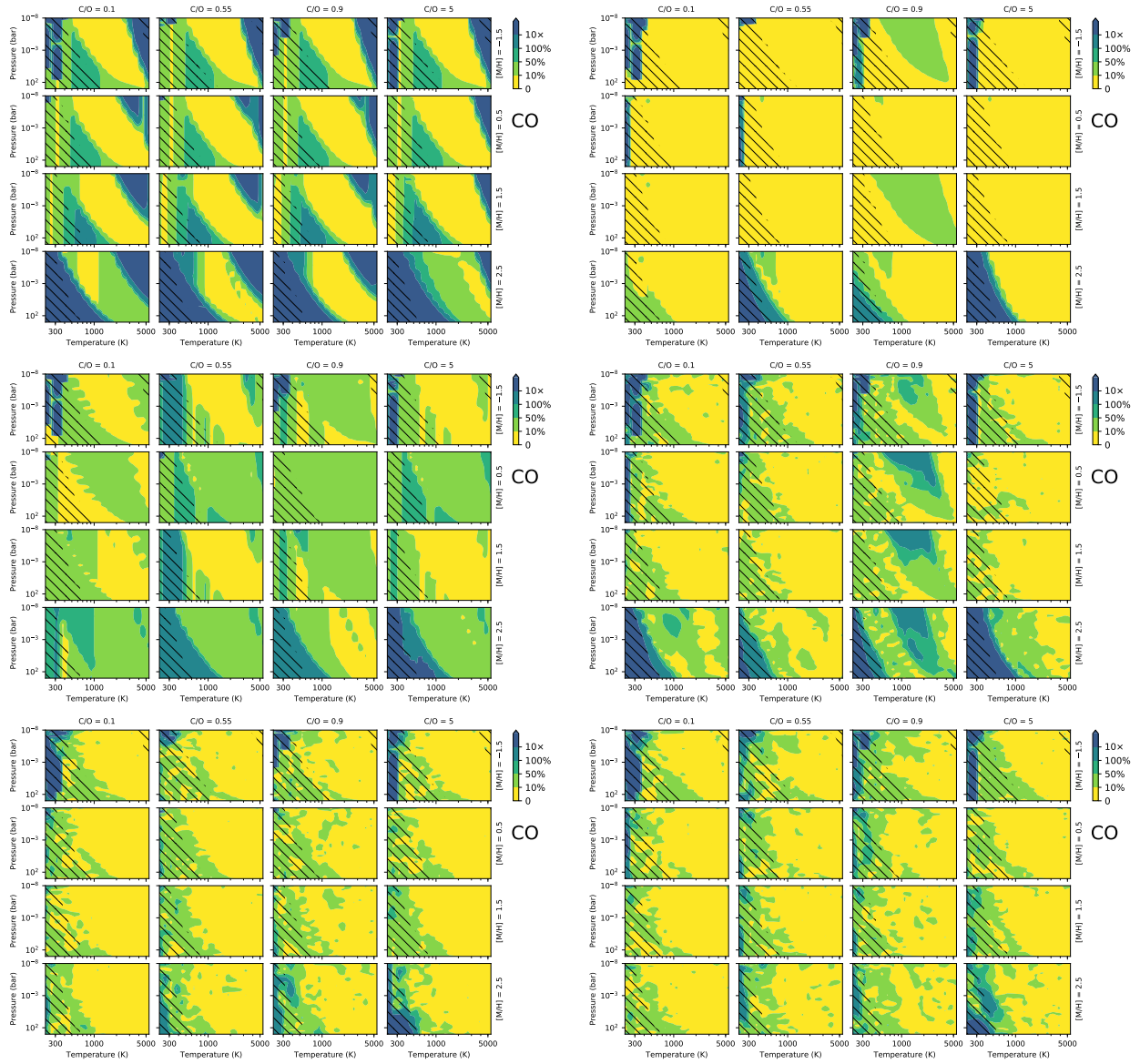


Figure 6.2: As in Figure 6.1, but for CO.

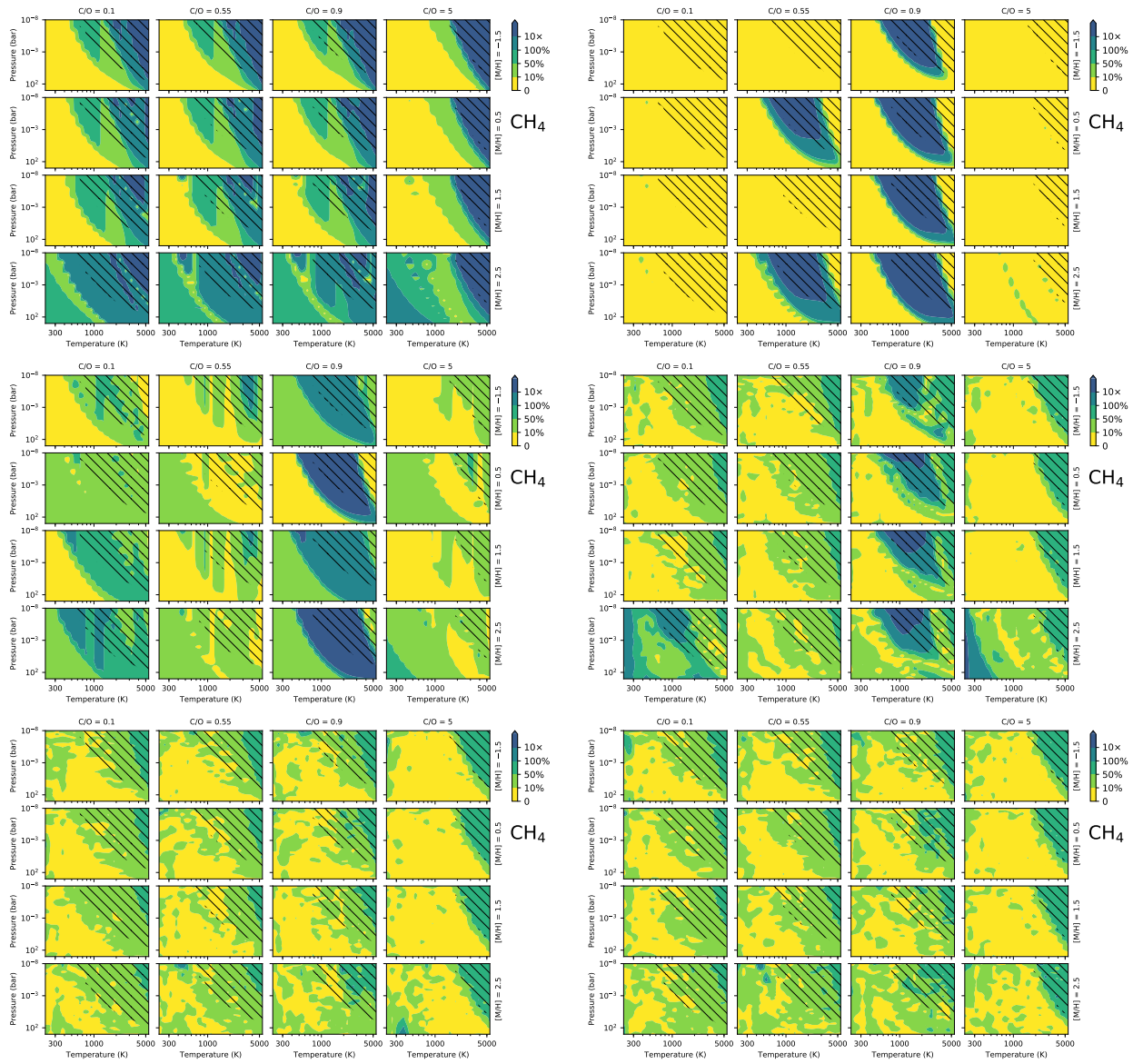


Figure 6.3: As in Figure 6.1, but for  $\text{CH}_4$ .

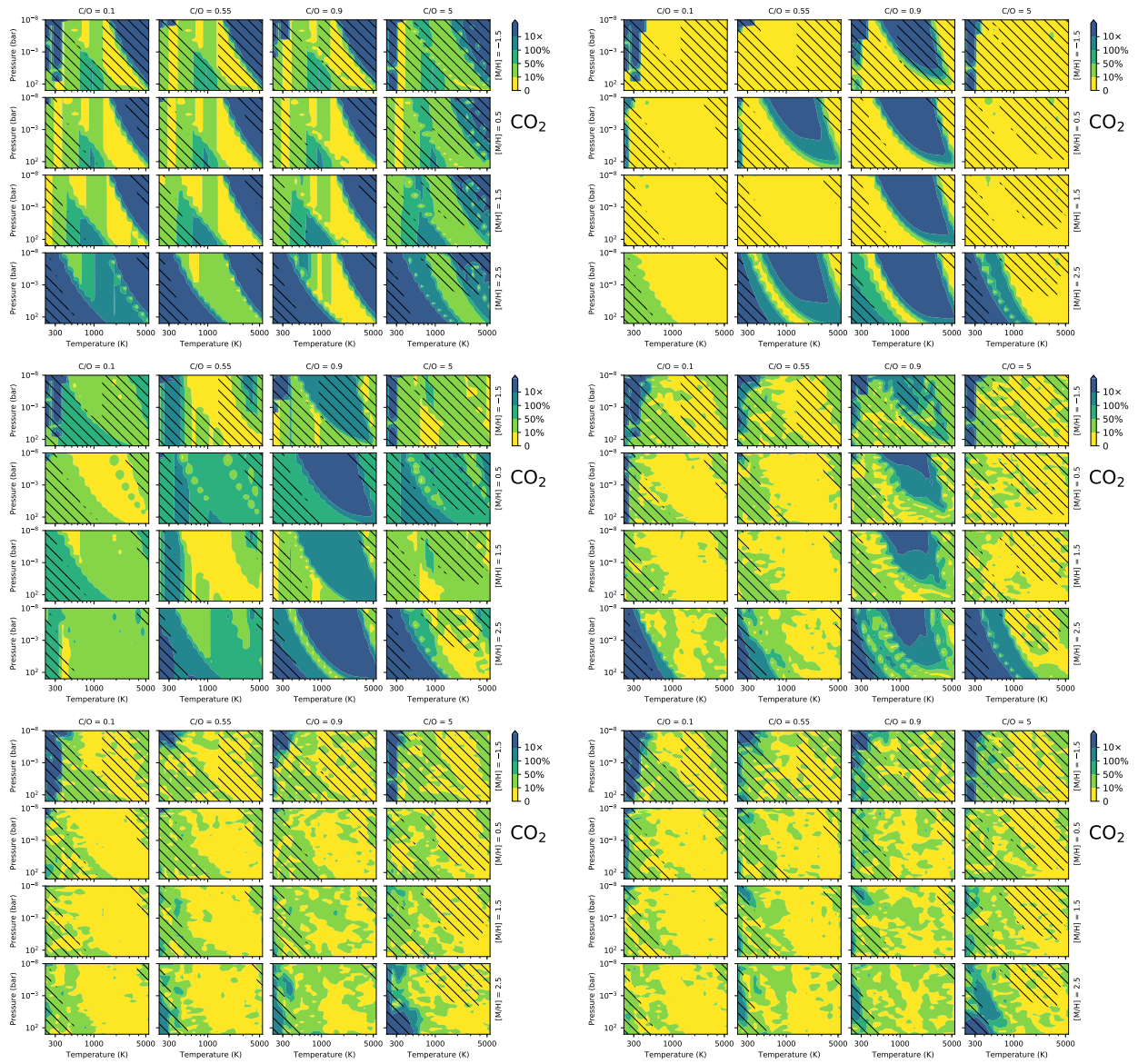


Figure 6.4: As in Figure 6.1, but for  $\text{CO}_2$ .

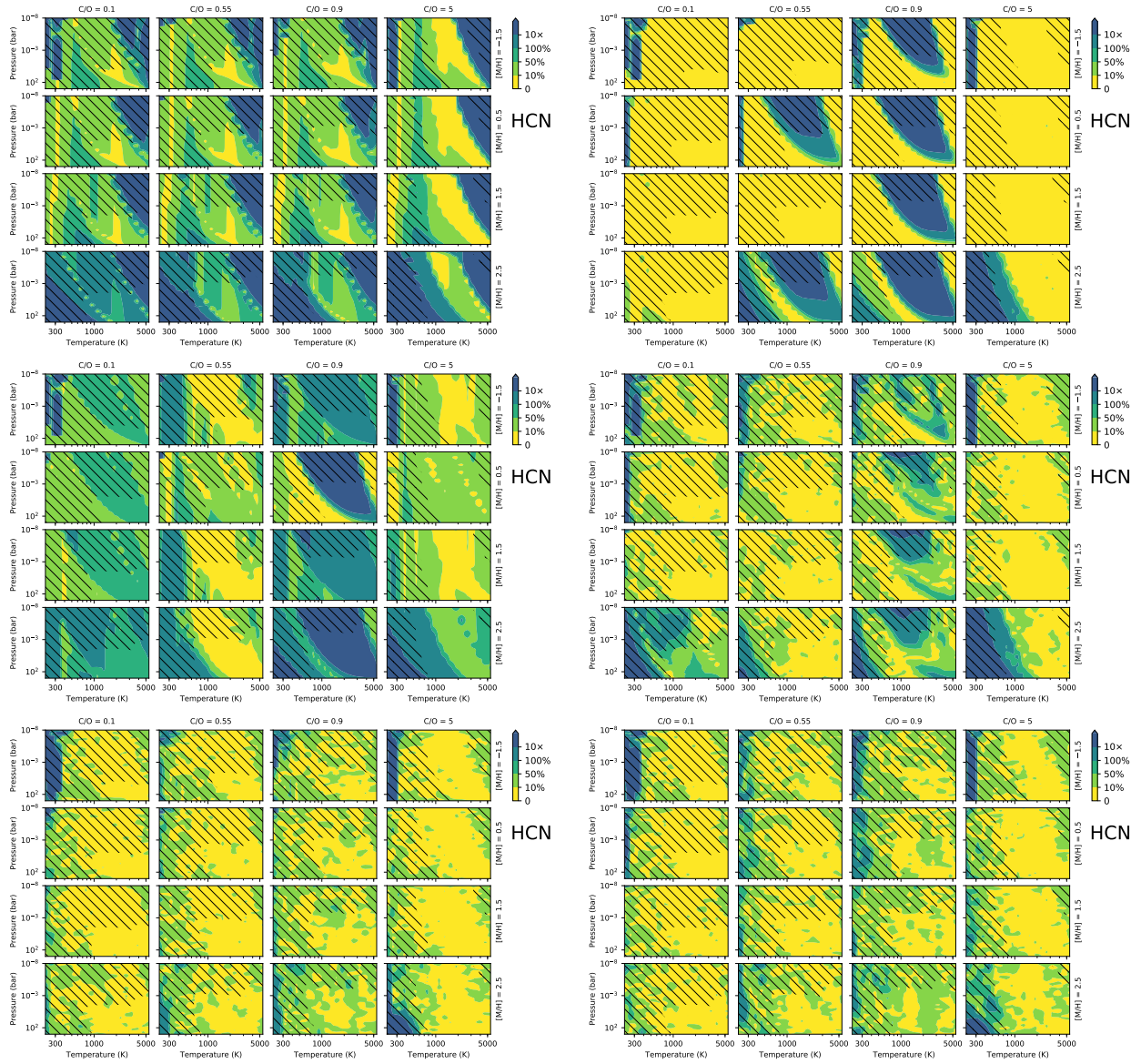


Figure 6.5: As in Figure 6.1, but for HCN.

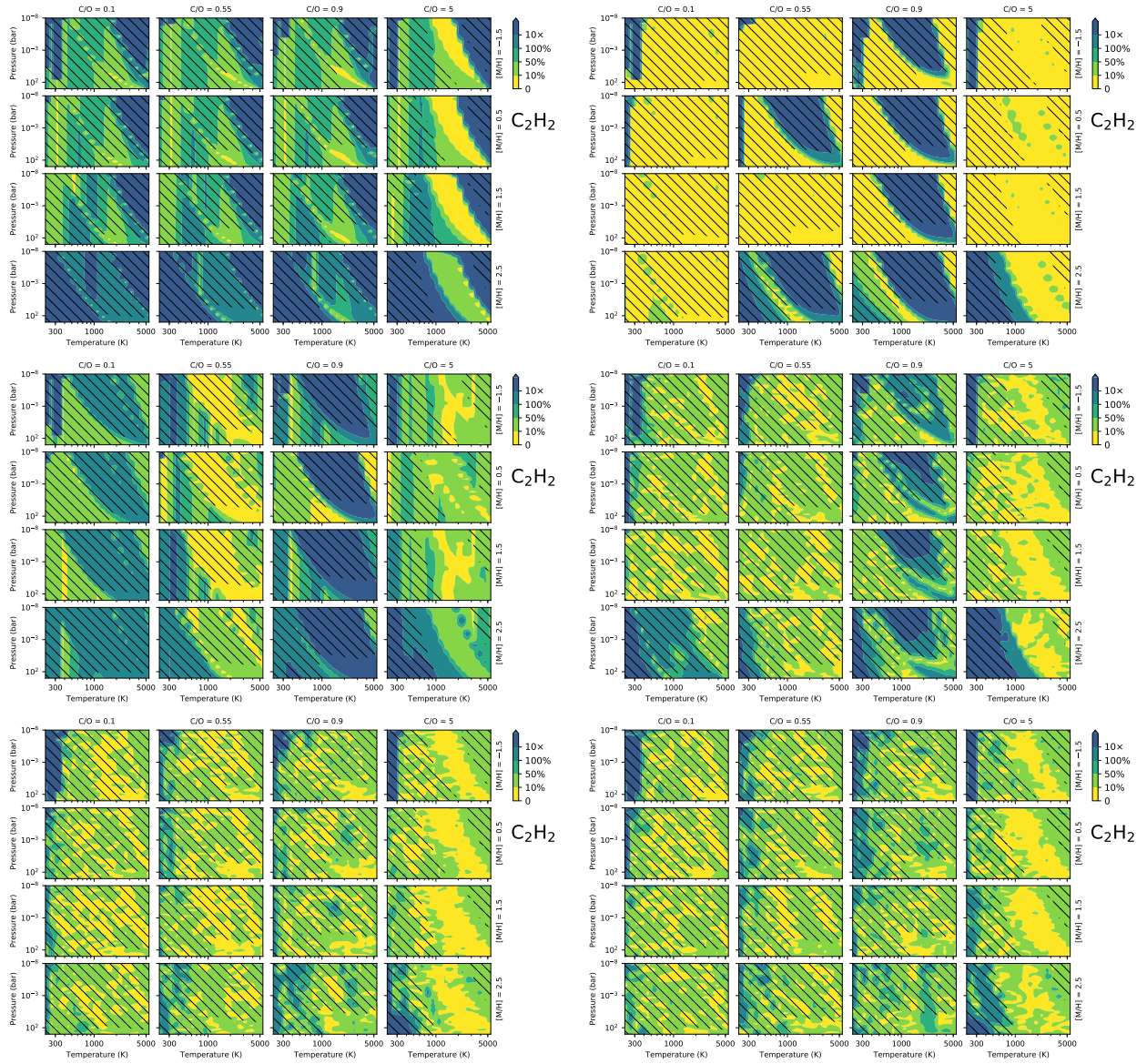


Figure 6.6: As in Figure 6.1, but for  $C_2H_2$ .

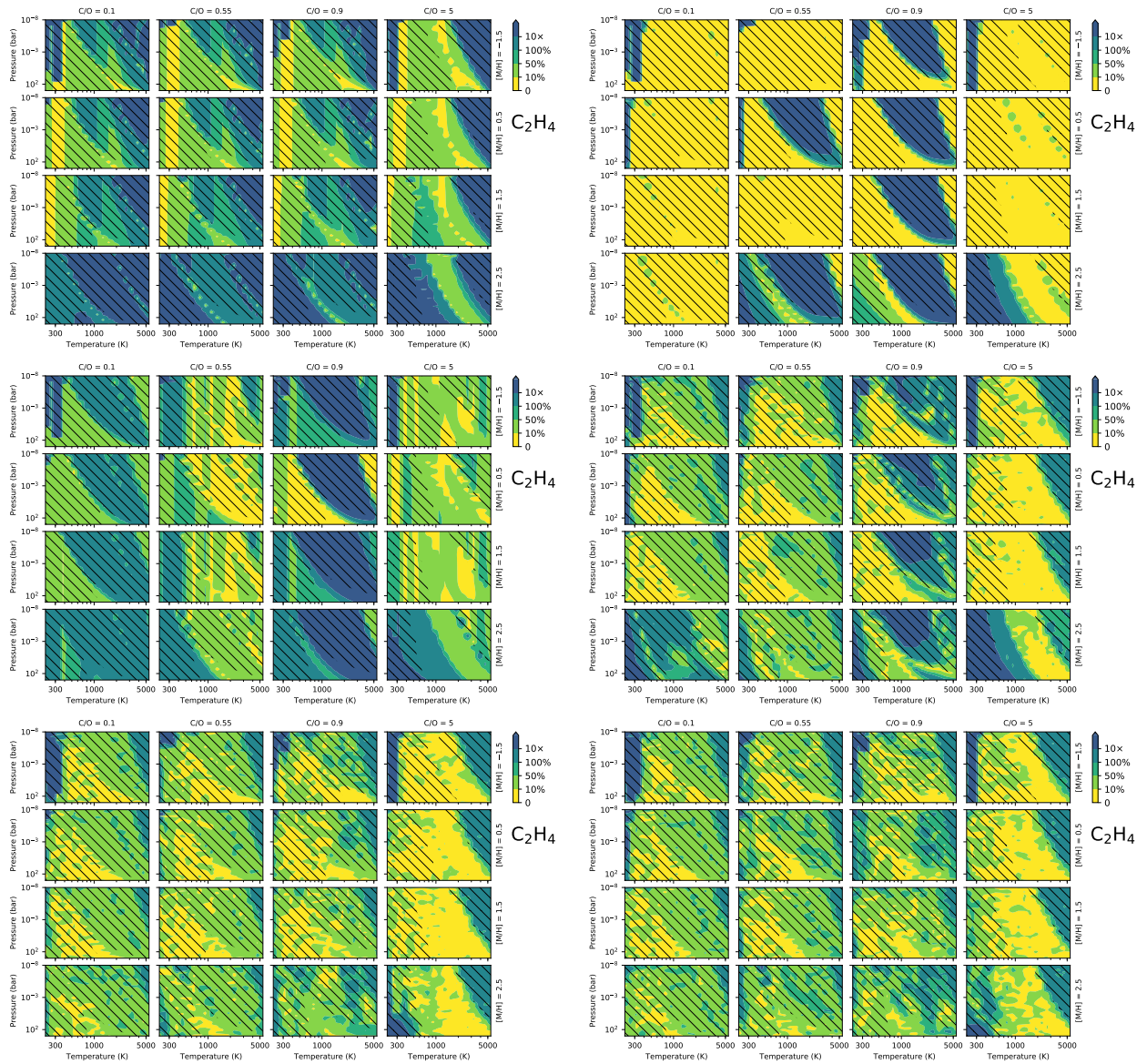


Figure 6.7: As in Figure 6.1, but for  $C_2H_4$ .

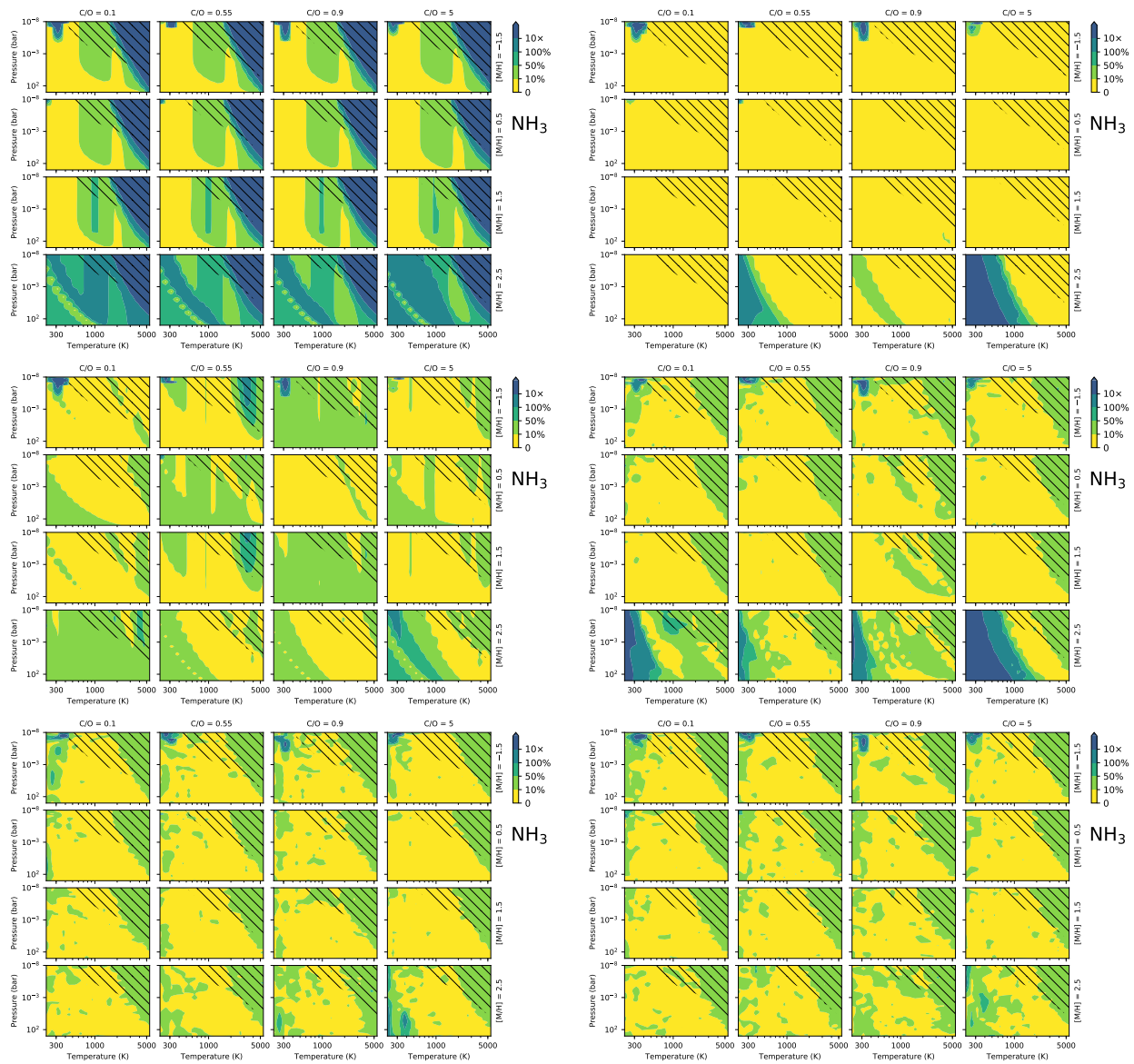


Figure 6.8: As in Figure 6.1, but for  $\text{NH}_3$ .

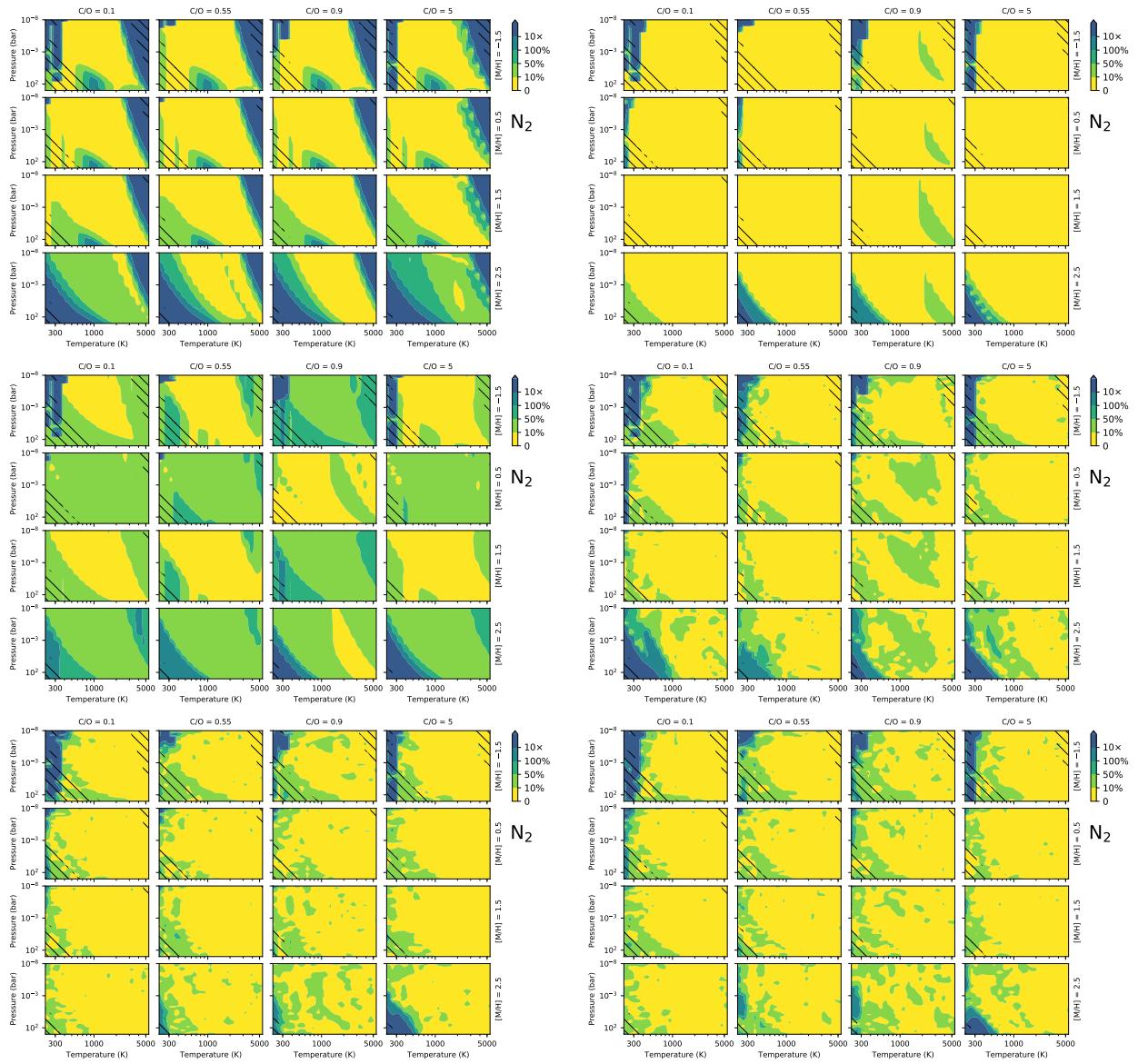


Figure 6.9: As in Figure 6.1, but for  $N_2$ .



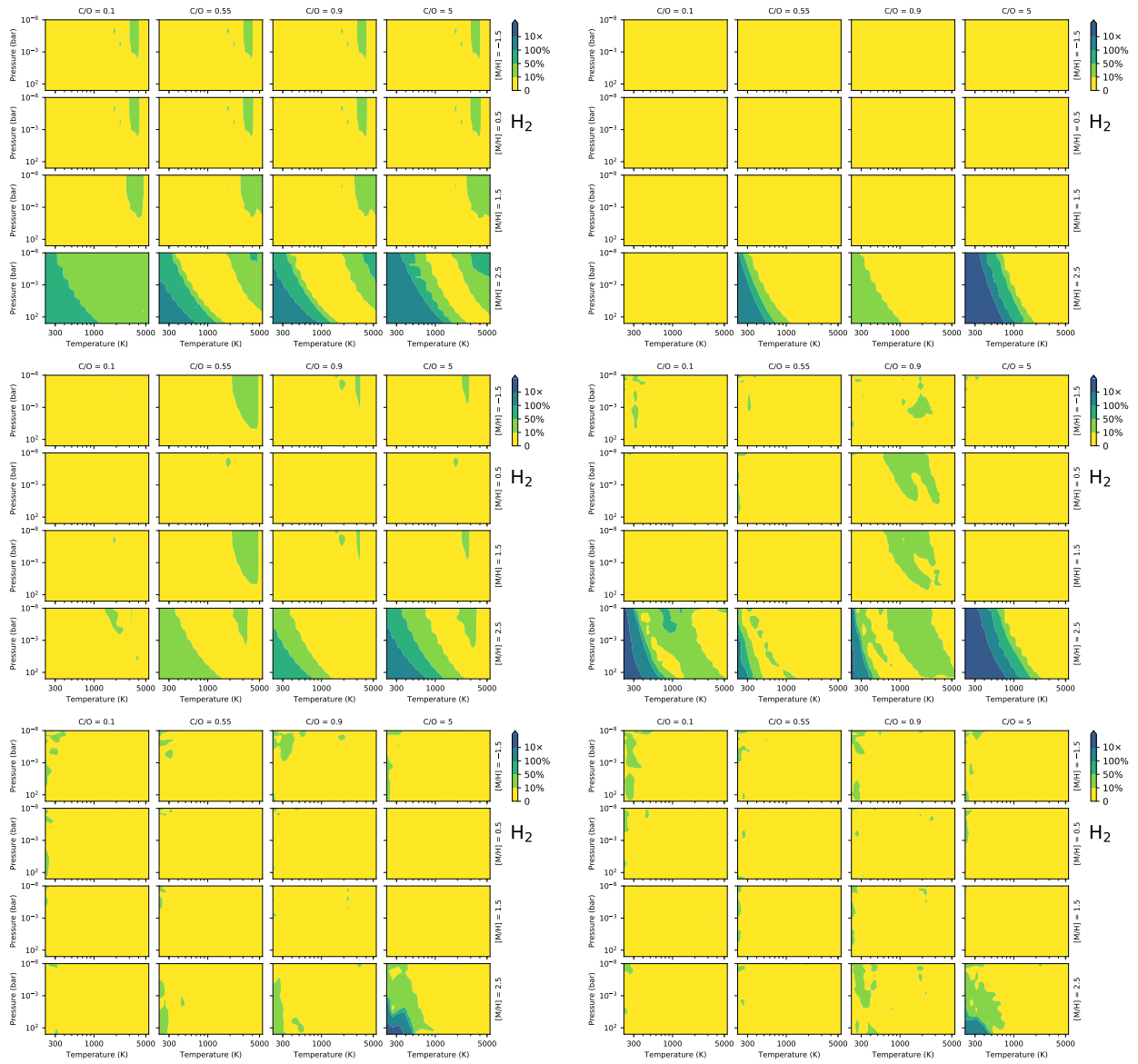


Figure 6.10: As in Figure 6.1, but for  $H_2$ .

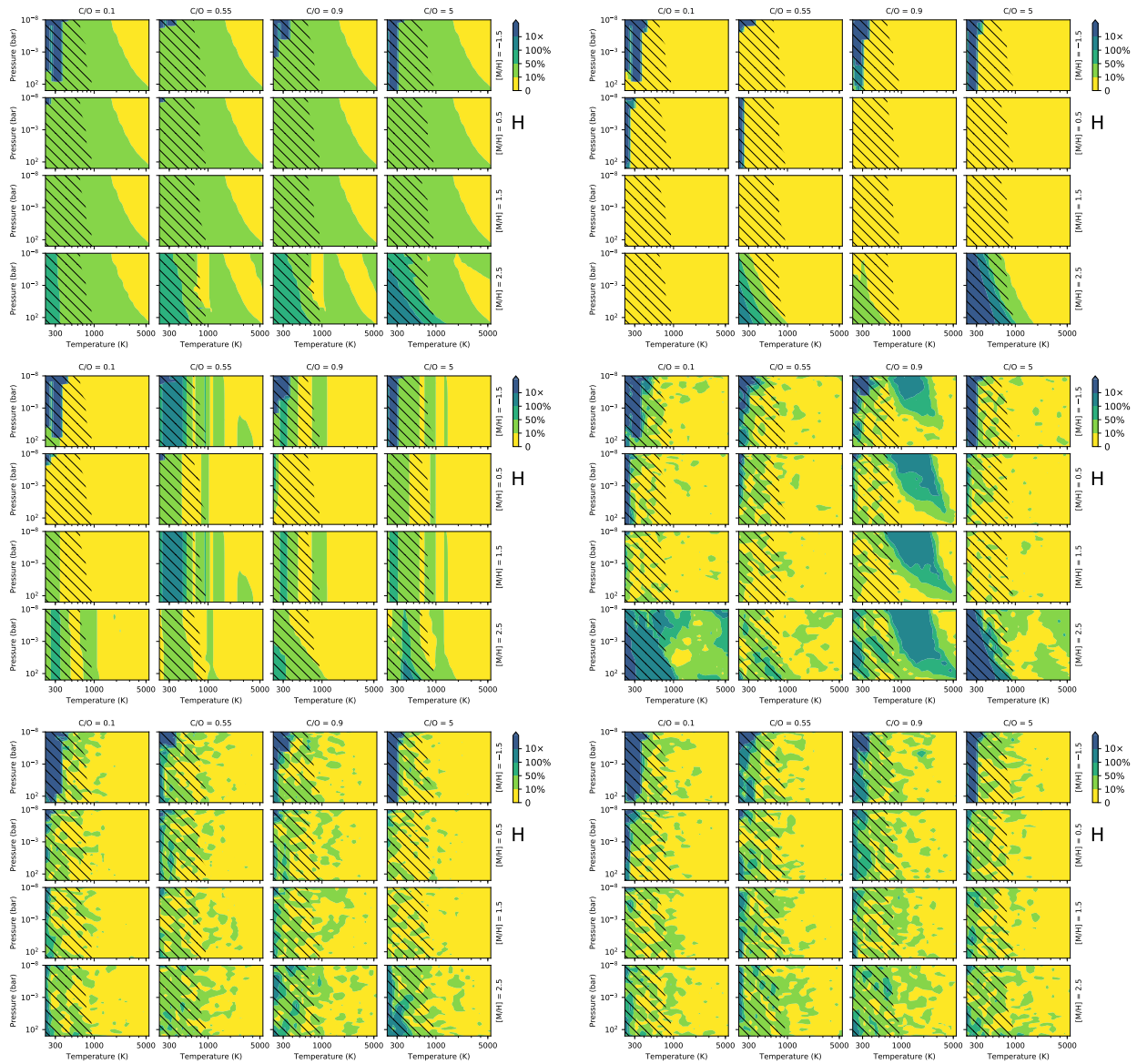


Figure 6.11: As in Figure 6.1, but for H.

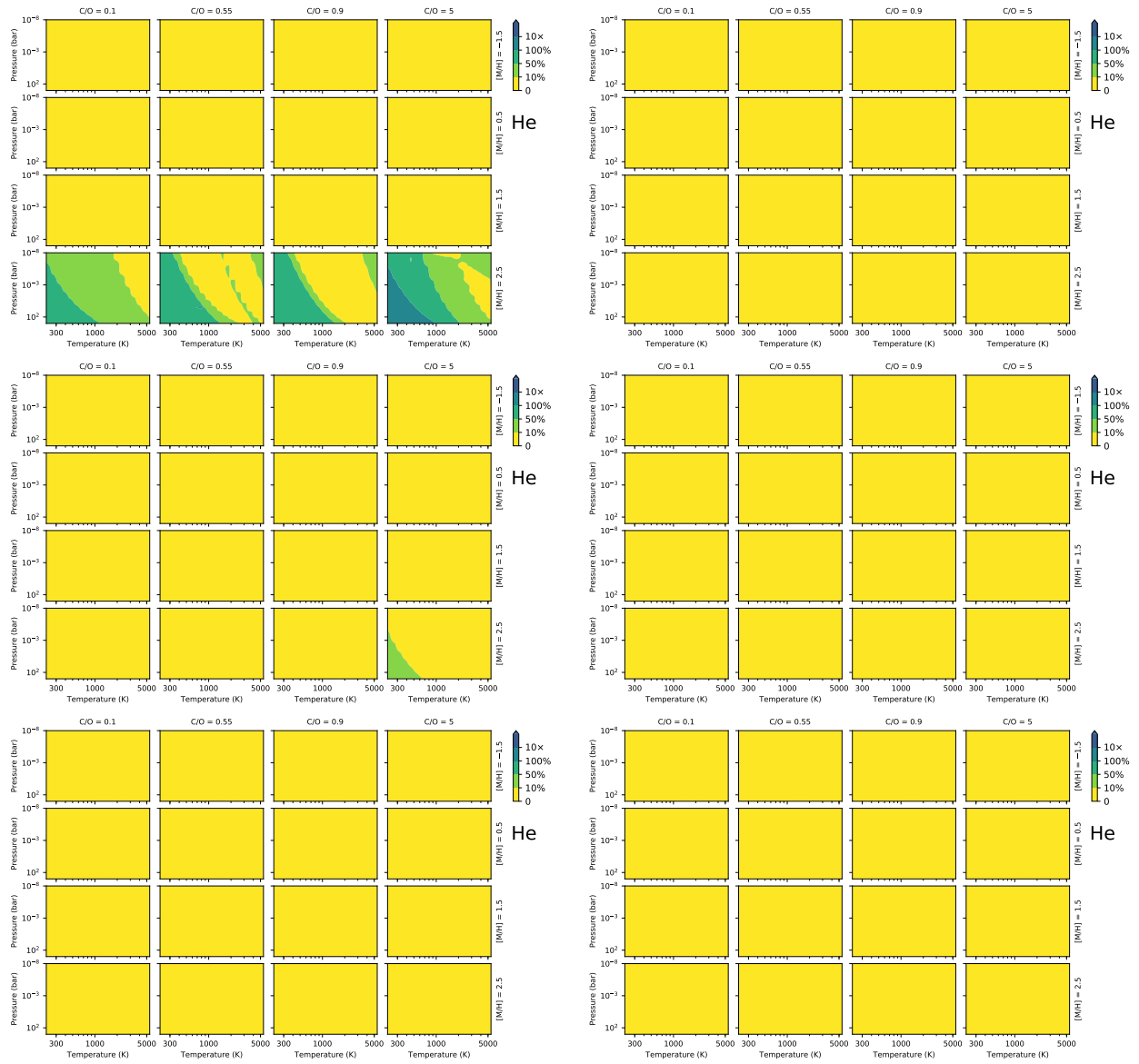


Figure 6.12: As in Figure 6.1, but for He.

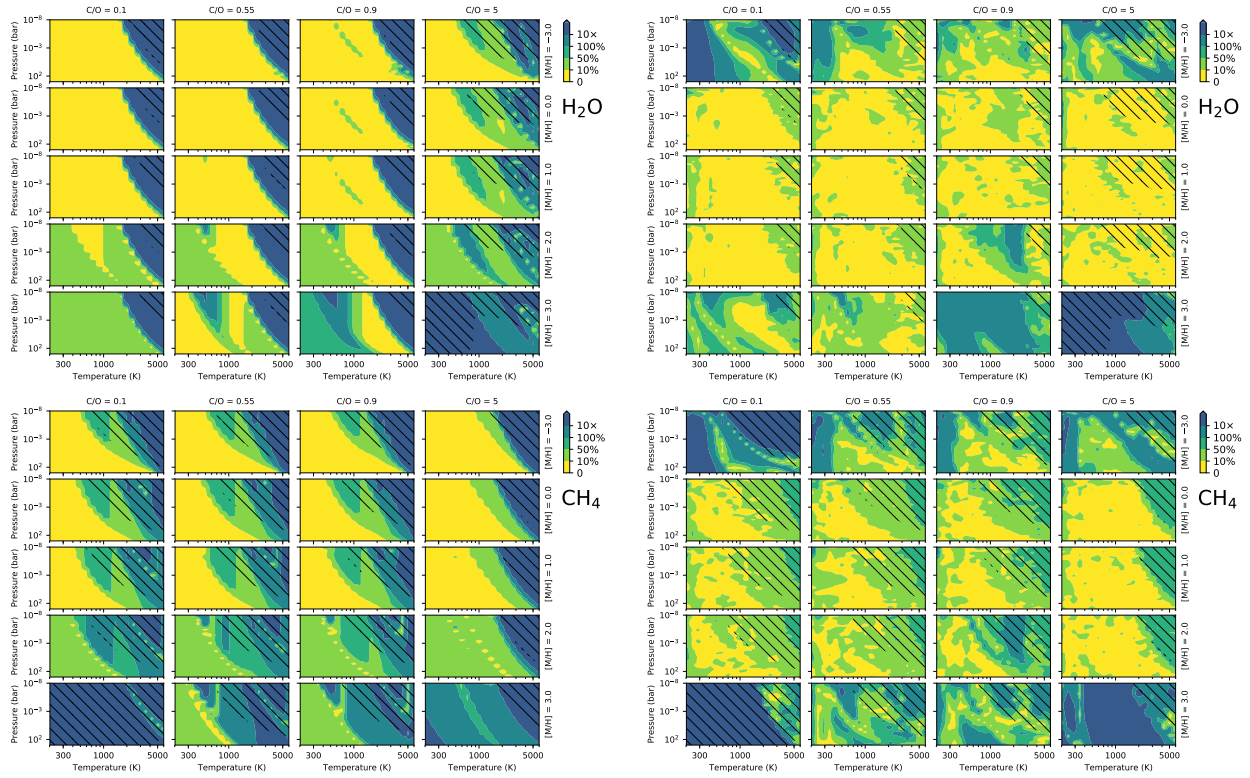


Figure 6.13: As in Figure 6.1, but using the grid of Cubillos et al. (2019) to illustrate NN inaccuracies at phase space extrema. While RATE (**left column**) performs accurately at a metallicity of  $-3$ , predictions using NN3 (**right column**) for this metallicity feature significant error in certain pressure–temperature regions. Note that the RATE plots differ from those in Cubillos et al. (2019) as we additionally consider helium in this investigation, which adjusts the relative abundances and therefore the relative errors.

## 6.5 Conclusions

In this study, we presented a comparison between thermochemical equilibrium estimation methods. We found that NN surrogate models outperform both interpolation approaches considered here as well as the analytical approximation of Cubillos et al. (2019), which is based on the formalism of Heng et al. (2016), Heng & Lyons (2016), and Heng & Tsai (2016). However, all approaches offer orders of magnitude reductions in computational cost compared to the Gibbs energy minimization implemented in TEA (Blecic et al. 2016). In the context of 2D and 3D atmospheric retrieval models, our results suggest that these thermochemical equilibrium estimation methods could run fast enough and perform accurately enough to be used in place of a Gibbs-minimization method for thermochemical equilibrium. More generally, these fast approximation approaches enable computationally expensive chemistry models to be utilized within retrieval models.

When comparing NN models, we found that the NNs trained on randomly-generated data outperformed those that were trained on a grid of models, even when the number of random data cases matched the number of cases in the model grid. Our results show that those NNs trained on random data outperform the considered interpolation methods even when the random data set size is  $\sim 7\%$  of the size of the interpolation data set. We also found that training on a combination of gridded and random data results in a less accurate NN than training on only the random data, even when trained for more iterations. While this suggests that gridded data is less effective at training NNs compared to random data, a future study that more thoroughly investigates this behavior is necessary to determine this definitively.

While the NNs performed accurately over most of the phase space, their accuracies significantly decrease near the edge of the phase space. Future investigations that seek to train comprehensive NN surrogate models should keep this in mind when generating data, as the trained surrogate model may not be valid at the extrema of the phase space. To account for this, we recommend generating data over a slightly larger phase space than required. For situations where physical limits prevent expanding the phase space, it may be helpful to force some of the random data to occur at the extrema, that is, fixing one or more parameters to extrema and randomly generating the other parameters. A future study should investigate this in detail to determine how best to handle this situation.

The Reproducible Research Compendium for this work is available for download.<sup>1</sup>

## 6.6 Acknowledgements

We thank contributors to NumPy, SciPy, Matplotlib, Tensorflow, Keras, the Python Programming Language, the free and open-source community, and the NASA Astrophysics Data System for software and services. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. This research was supported by the NASA Fellowship Activity under NASA Grant 80NSSC20K0682.

## 6.7 List of References

Baydin, A. G., Heinrich, L., Bhimji, W., et al. 2019, in *Advances in Neural Information Processing Systems* 33

Blecic, J., Dobbs-Dixon, I., & Greene, T. 2017, *ApJ*, 848, 127

---

<sup>1</sup>To be uploaded at Zenodo

- Blecic, J., Harrington, J., & Bowman, M. O. 2016, *ApJS*, 225, 4
- Blumenthal, S. D., Mandell, A. M., Hébrard, E., et al. 2018, *ApJ*, 853, 138
- Brehmer, J., Cranmer, K., Louppe, G., & Pavez, J. 2018, *Phys. Rev. D*, 98, 052004
- Caldas, A., Leconte, J., Selsis, F., et al. 2019, *A&A*, 623, A161
- Cubillos, P. E., Blecic, J., & Dobbs-Dixon, I. 2019, *ApJ*, 872, 111
- Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. 2017, in *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML17 (JMLR.org)*, 12631272
- Heng, K., & Lyons, J. R. 2016, *ApJ*, 817, 149
- Heng, K., Lyons, J. R., & Tsai, S.-M. 2016, *ApJ*, 816, 96
- Heng, K., & Tsai, S.-M. 2016, *ApJ*, 829, 104
- Himes, M. D., Harrington, J., Cobb, A. D., et al. 2022, *PSJ*, 3, 91
- Kasim, M. F., Watson-Parris, D., Deaconu, L., et al. 2021, *Machine Learning: Science and Technology*, 3, 015013
- Kawashima, Y., & Min, M. 2021, *A&A*, 656, A90
- Line, M. R., Knutson, H., Wolf, A. S., & Yung, Y. L. 2014, *ApJ*, 783, 70
- Line, M. R., & Yung, Y. L. 2013, *ApJ*, 779, 3
- Madhusudhan, N. 2012, *ApJ*, 758, 36
- . 2018, *Atmospheric Retrieval of Exoplanets (Springer International Publishing AG)*, 104

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

Mollière, P., Stolker, T., Lacour, S., et al. 2020, *A&A*, 640, A131

Moses, J. I., Madhusudhan, N., Visscher, C., & Freedman, R. S. 2013a, *ApJ*, 763, 25

Moses, J. I., Visscher, C., Fortney, J. J., et al. 2011, *ApJ*, 737, 15

Moses, J. I., Line, M. R., Visscher, C., et al. 2013b, *ApJ*, 777, 34

Munk, A., Ścibior, A., Baydin, A. G., et al. 2019, arXiv preprint arXiv:1910.11950

Oreshenko, M., Lavie, B., Grimm, S. L., et al. 2017, *ApJ*, 847, L3

Pluriel, W., Leconte, J., Parmentier, V., et al. 2022, *A&A*, 658, A42

Roudier, G. M., Swain, M. R., Gudipati, M. S., et al. 2021, *AJ*, 162, 37

Smith, L. N. 2015, arXiv e-prints, arXiv:1506.01186

Stevenson, K. B., Harrington, J., Nymeyer, S., et al. 2010, *Nature*, 464, 1161

Venot, O., Hébrard, E., Agúndez, M., Decin, L., & Bounaceur, R. 2015, *A&A*, 577, A33



**APPENDIX A: RETRIEVAL ERRORS DUE TO WAVENUMBER  
SAMPLING GRID MISMATCH**

[This appendix is attached to the published manuscript that contains Chapter 2. It is Appendix D in the official publication.]

When performing the synthetic retrievals described in Section 2.1.3, we observed a numerical effect that can bias the Bayesian sampler towards slightly incorrect answers when the forward and retrieval models use differing wavenumber grids. Here we describe this effect and how to minimize it. We note that this error most strongly manifests when the RT code that produced the synthetic spectrum matches the RT code used when retrieving on the synthetic spectrum, and the error appears to be negligible when retrieving on real data at current resolutions.

Line-by-line calculations necessitate a discrete sampling of the resulting spectrum. Yet, spectra are continuous, and this discrete sampling will therefore introduce error when, e.g., band integrating a spectrum during a retrieval. At commonly-used grid samplings (e.g.,  $1.0 \text{ cm}^{-1}$ ) for current- and next-generation telescope resolutions, these errors can drive the Bayesian sampler to an incorrect part of the phase space.

We consider forward models with 4 different grid spacings: 0.025, 0.1, 0.25, and  $1.0 \text{ cm}^{-1}$ . For each grid spacing, we also consider a horizontal shift of all values by  $\sim 0.09 \text{ cm}^{-1}$ . Figure A.1 shows that, for the 0.25 and  $1.0 \text{ cm}^{-1}$  griddings, there is disagreement in the  $2.5 - 3.5$  and  $>4.5 \mu\text{m}$  regions. By comparison, when only considering the 0.025 and  $0.1 \text{ cm}^{-1}$  griddings, the differences tend to be comparable to the width of the plotted lines.

We further investigate this effect by simulating a spectrum at  $0.1 \text{ cm}^{-1}$ , band integrating according to some filters, and retrieving with wavenumber grids that only differ in their spacing. The retrieval models use grid spacings of 0.1, 0.25, 0.5, and  $1.0 \text{ cm}^{-1}$ . We ensure that all wavenumbers in the retrieval model grids exactly overlap with wavenumbers from the forward model grid. Figure A.2 shows that choosing coarser griddings than that of the forward model leads to a reduction in accuracy: the  $0.1 \text{ cm}^{-1}$  gridding recovers all of the parameters within  $1\sigma$ , while the 0.25 and  $0.5 \text{ cm}^{-1}$  griddings recovering most parameters at  $>1\sigma$ . The  $1.0 \text{ cm}^{-1}$  gridding entirely misses 4 out of the 5 parameters.

This effect would be expected to be most pronounced in the case where the same RT code is used for the forward and retrieval models, since the retrieval RT model can, in theory, exactly match the forward RT model. To test whether this effect arises when the forward and retrieval RT codes differ, we ran the Barstow et al. (2020) cases at two different resolutions ( $0.1$  and  $1.0 \text{ cm}^{-1}$ , Figure A.3). The inferred temperatures and radii are generally insensitive to the grid selection, though the retrieved molecular abundances only slightly overlap, with the lower/upper credible region boundaries varying by around an order of magnitude. On the other hand, when retrieving using the real data of HD 189733 b at the aforementioned grid resolutions, we find only minor differences (Figure A.4). The posterior of thermal profiles (calculated via the  $\kappa$ ,  $\gamma_1$ , and  $\beta$  parameters) are nearly identical, and  $\text{H}_2\text{O}$  and  $\text{CO}_2$  are essentially unaffected.  $\text{CO}$  and  $\text{CH}_4$  are minorly affected, with the coarser grid preferring slightly higher  $\text{CO}$  and lower  $\text{CH}_4$  abundances. These minor differences are not significant enough to change the interpretation of the results in the case of current-resolution data for HD 189733 b. However, the numerical effect described in this section may manifest in real-data retrievals as resolution improves.

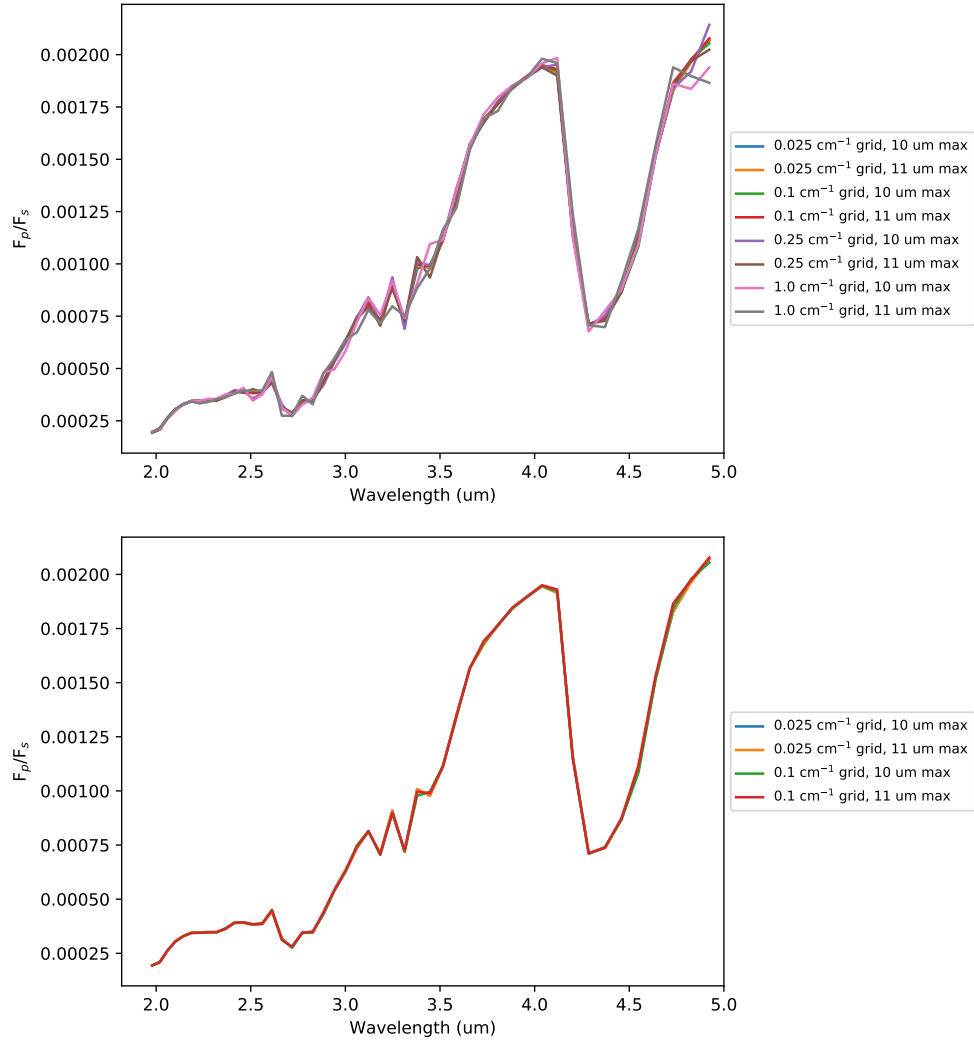


Figure A.1: Example of the effect of wavenumber griddings when band integrating spectra. **Left:** 0.025, 0.1, 0.25, and 1.0 cm<sup>-1</sup> griddings. **Right:** 0.025 and 0.1 cm<sup>-1</sup> griddings from the left plot. All spectra are computed using identical setups except for the wavenumber gridding. The gridding of the spectra computed with a maximum wavelength of 11 μm is offset by  $\sim 0.09$  cm<sup>-1</sup> compared to the spectra with a maximum wavelength of 10 μm. Note the model spread in the 2.5 – 3.5 and  $> 4.5$  μm regions of the left plot. By comparison, the right plot shows little spread in the band-integrated spectra.

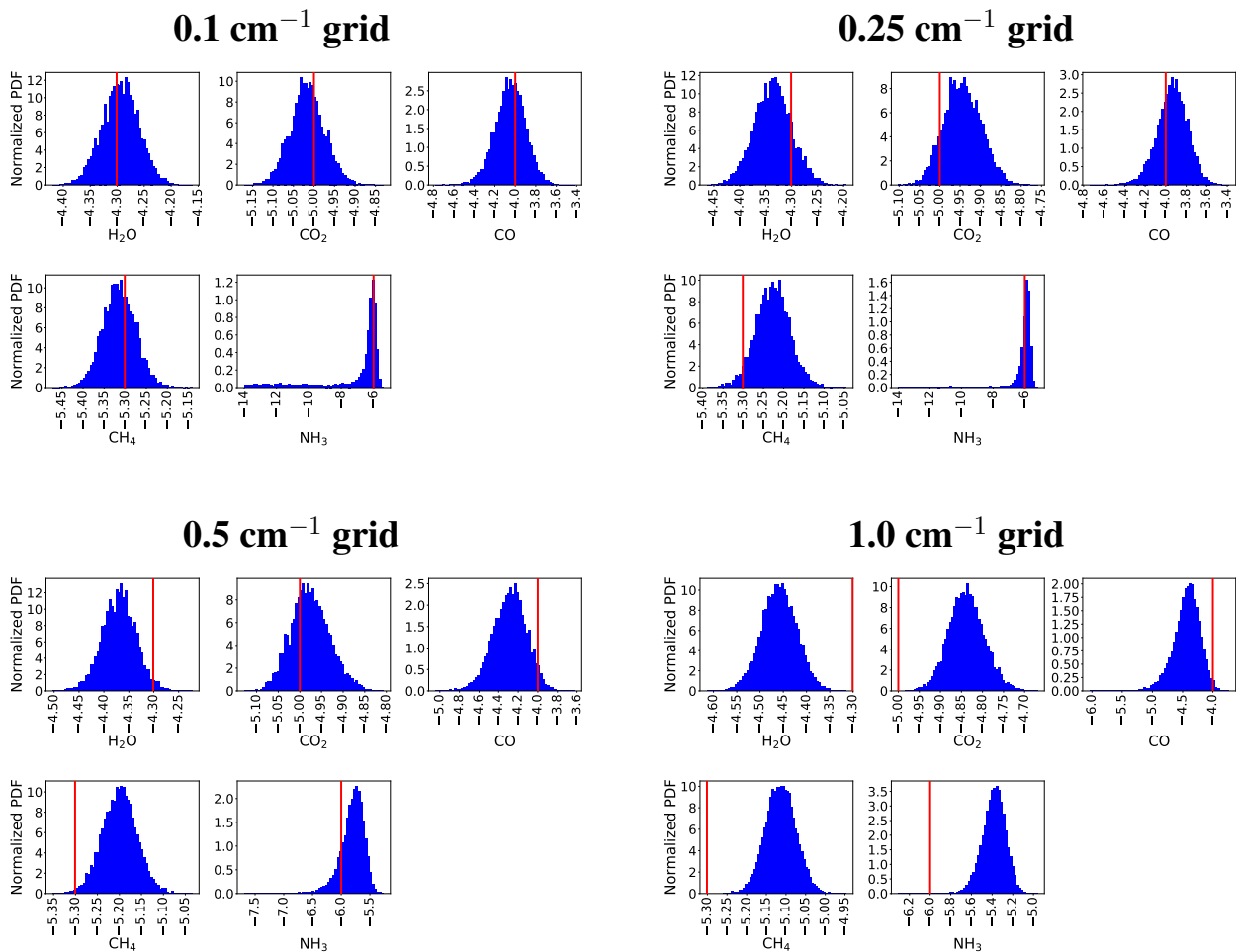


Figure A.2: Retrieved 1D marginalized posterior distributions for each molecule considered in the synthetic retrieval. The forward model was computed using a  $0.1 \text{ cm}^{-1}$  grid, while the retrieval models were computed using  $0.1$  (**top left**),  $0.25$  (**top right**),  $0.5$  (**bottom left**), and  $1.0$  (**bottom right**)  $\text{cm}^{-1}$  grids. Note that only the  $0.1 \text{ cm}^{-1}$  grid correctly retrieves the underlying parameters within  $1\sigma$ ; as the gridding becomes coarser, the accuracy decreases, with the  $1.0 \text{ cm}^{-1}$  gridding entirely missing 4 out of 5 of the parameters.

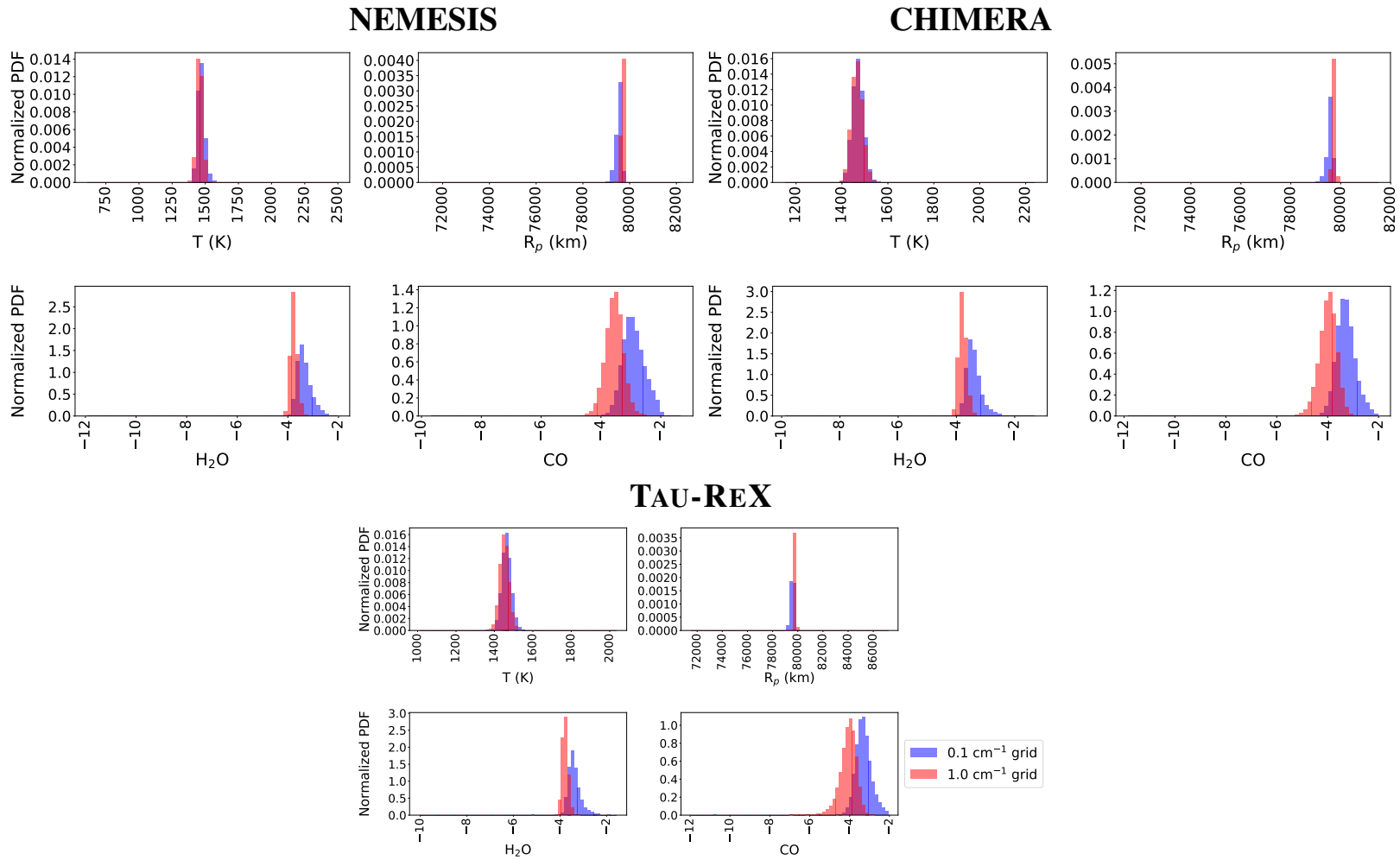


Figure A.3: Comparison of retrieved posteriors for the three indicated Barstow et al. (2020) model 0 spectra at two different wavenumber grid resolutions. While the retrieved temperatures and radii are generally unaffected, the molecular abundances are more sensitive to the grid resolution.

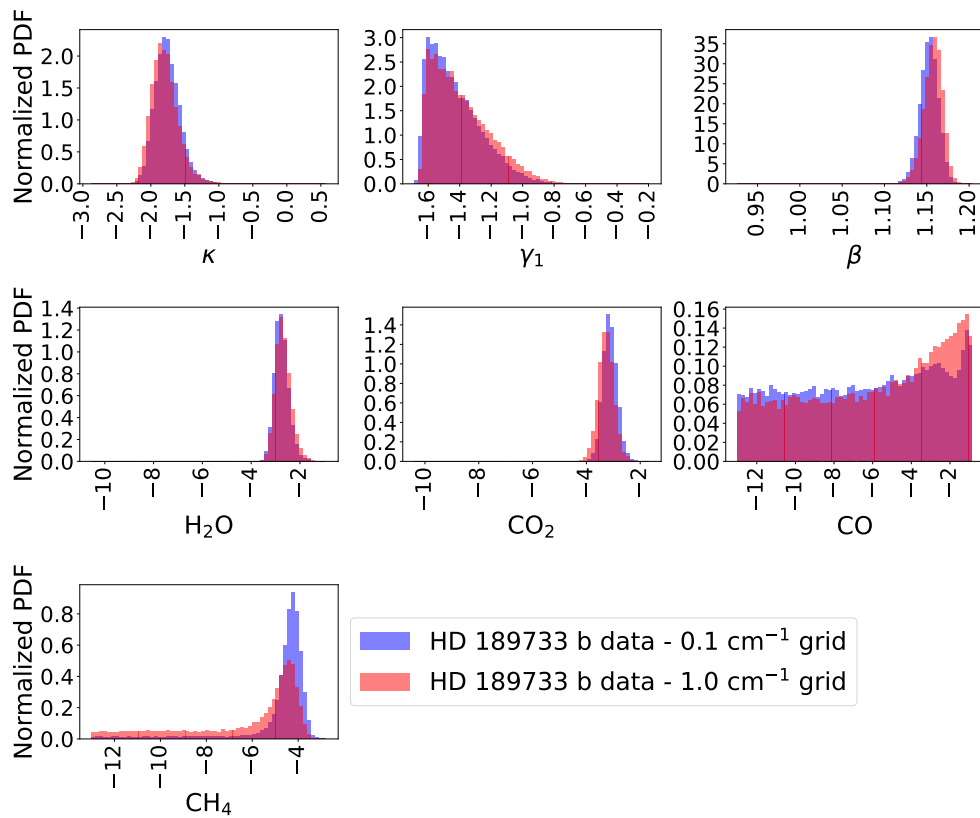


Figure A.4: Comparison of retrieved posteriors for HD 189733 b at two different wavenumber grid resolutions. In general, the posteriors match; only CO and CH<sub>4</sub> show minor differences, though they are not significant enough to affect conclusions drawn from the posterior.

## A.1 List of References

Barstow, J. K., Changeat, Q., Garland, R., Line, M. R., Rocchetto, M., & Waldmann, I. P. 2020, MNRAS, 493, 4884



## **APPENDIX B: DETERMINING MODEL ARCHITECTURE**

[This appendix is attached to the published manuscript that contains Chapter 5. It is Appendix A in the official publication.]

To select an ideal neural-network architecture, a grid search must be carried out. This includes varying the types of hidden layers, number of hidden layers, number of nodes per layer, activation functions for each hidden layer, the parameter(s) for each activation function, and the learning rate.

We carried out a grid search by training each model on a subset of the total data set (171,456 training, 66,432 validation) for 20 epochs using a batch size of 64. We considered 3–5 dense and convolutional+pooling hidden layers; 64–4096 nodes; rectified linear unit (ReLU), leaky ReLU, exponential linear unit (ELU), hyperbolic tangent (tanh), and sigmoid activation functions. The convolutional layers use a kernel size of 3, and pooling layers use a size of 2. We consider four learning rate policies: (1) a cyclical rate ranging from  $8 \times 10^{-6}$  –  $5 \times 10^{-3}$  where the maximum is reduced by half of the difference with the minimum every 8 epochs, (2) as before but ranging from  $10^{-5}$  –  $10^{-3}$ , (3) a constant learning rate of  $10^{-5}$ , and (4) a constant  $10^{-3}$ . Policies 3 and 4 are only considered for models that do not include tanh or sigmoid activations.

Table B.1 presents the minimum validation loss for each architecture considered. There is some randomness to the minimum validation loss due to the shuffling of the training data, so models with comparable minimum validation losses can be considered equivalent in performance. We chose to perform a more exhaustive grid search than is typical to emphasize certain points that can guide future investigations.

In general, we find that models with 4+ hidden layers with ReLU, leaky ReLU, and ELU activations achieve the lowest validation loss for this problem. The best-performing models all have a 1D convolutional layer as the first hidden layer, while the worst-performing models use tanh or sigmoid activations. Additional layers generally lead to reductions in the loss. Cases where this does not occur can be attributed to the learning rate policy (e.g., models 25–27, LR1 vs. LR2),

highlighting the importance of properly selecting the policy (described below). Minor variations (e.g., models 28–30 LR1) can be attributed to training randomness. ReLU and leaky ReLU activations tend to have similar performance; leaky ReLU with a small parameter tends to perform equivalently or better than ReLU (e.g., models 32 and 33). While these results point to deep architectures as optimal configurations for this application of ML RT, tests varying the spectral resolution, wavelength range, etc. are necessary to definitively confirm if such variations change the optimal architecture(s). A future investigation should consider this in more detail.

Based on this grid search, we selected model 40. While a similar architecture with ELU activations performed equivalently (model 37), it took longer to train per epoch. Additionally, we found that the retrieval accuracy did not significantly change below some threshold validation loss (see Appendix C), so training time is a more important consideration than minor differences in minimum validation loss.

Our results show that, when the learning rate range is properly chosen, cyclical learning rates outperform constant learning rates, confirming the findings of Smith (2015) for this particular problem. Select models do not follow this trend (e.g., models 31, 35, 38), which is likely attributable to the small number of epochs considered in this grid search.

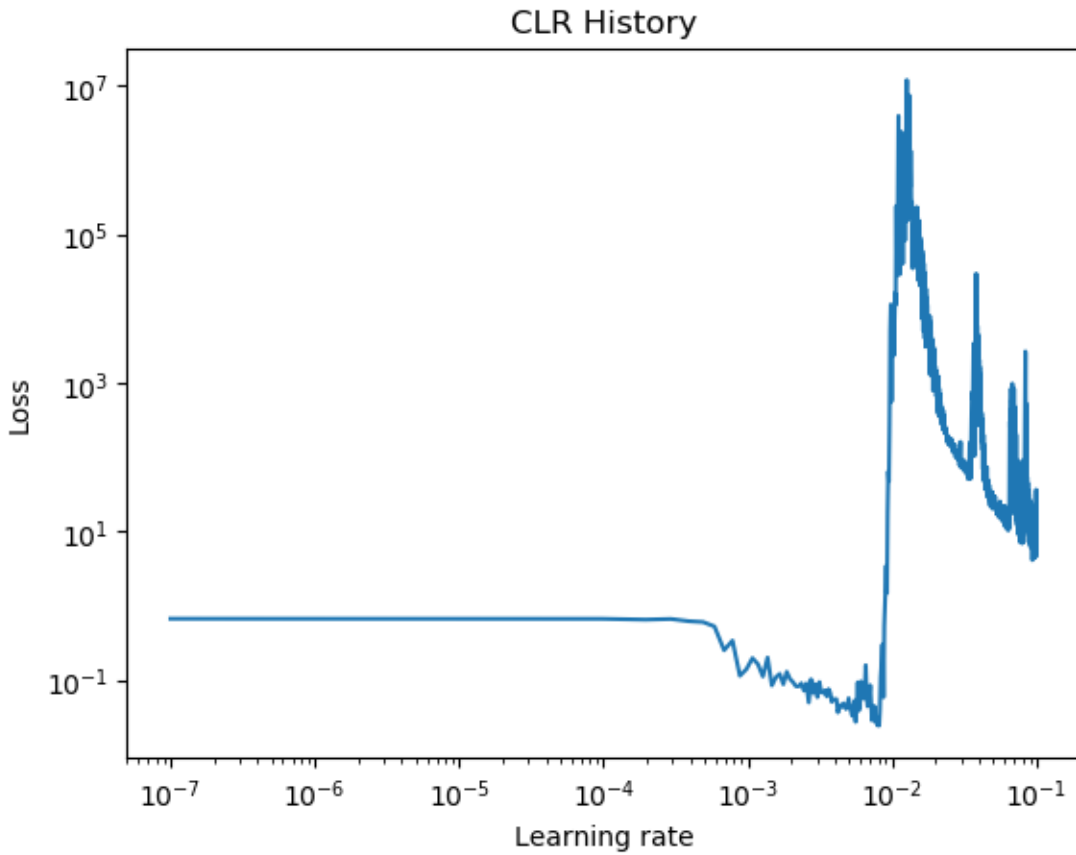


Figure B.1: Example of a range test. The learning rate begins at a value too small to make noticeable changes to the weights of the model. At a learning rate of  $\sim 4 \times 10^{-4}$ , the loss begins to decrease, indicating that the model has begun learning. However, at a learning rate  $\sim 5 \times 10^{-3}$ , the loss begins to behave erratically, and it becomes very large at a learning rate of  $10^{-2}$ . From this, a learning rate policy varying between  $6 \times 10^{-4}$  and  $4 \times 10^{-3}$  would likely perform well for this architecture.

Table B.1: Model Grid Search, 20 Epochs

#	Hidden Layers	Min. Val. Loss ( $\times 10^5$ )			
		LR1*	LR2†	LR3‡	LR4§
1	D(512) <sup>a</sup> R <sup>b</sup> –D(512)R–D(512)R	17.4	43.8	491	19.2
2	D(1024)R–D(1024)R–D(1024)R	9.61	24.6	291	12.3
3	D(2048)R–D(2048)R–D(2048)R	7.60	13.3	179	8.28
4	D(4096)R–D(4096)R–D(4096)R	8.56	7.17	106	6.54
5	D(512)R–D(512)R–D(512)R–D(512)R	13.0	31.3	382	16.2
6	D(512)S <sup>c</sup> –D(512)S–D(512)S–D(512)S	68.0	951	—	—
7	D(512)T <sup>d</sup> –D(512)T–D(512)T–D(512)T	487	47.0	—	—
8	D(512)S–D(1024)S–D(2048)S–D(4096)S	48.5	932	—	—
9	D(512)T–D(1024)T–D(2048)T–D(4096)T	1390	68.8	—	—
10	D(1024)R–D(1024)R–D(1024)R–D(1024)R	8.43	16.8	238	10.1
11	D(2048)R–D(2048)R–D(2048)R–D(2048)R	7.46	9.13	125	8.01
12	D(4096)R–D(4096)R–D(4096)R–D(4096)R	8.23	5.05	62.5	6.14
13	D(4096)S–D(4096)S–D(4096)S–D(4096)S	1350	197	—	—
14	D(4096)T–D(4096)T–D(4096)T–D(4096)T	1740	46.0	—	—
15	D(4096)E(0.05) <sup>e</sup> –D(4096)E(0.05)–D(4096)E(0.05)–D(4096)E(0.05)	220	5.10	64.6	5.55
16	D(4096)E(0.1)–D(4096)E(0.1)–D(4096)E(0.1)–D(4096)E(0.1)	227	5.03	70.3	5.93
17	D(4096)E(0.15)–D(4096)E(0.15)–D(4096)E(0.15)–D(4096)E(0.15)	206	5.13	74.5	7.13
18	D(4096)E(0.2)–D(4096)E(0.2)–D(4096)E(0.2)–D(4096)E(0.2)	238	5.39	81.7	6.18
19	D(4096)L(0.05) <sup>f</sup> –D(4096)L(0.05)–D(4096)L(0.05)–D(4096)R	7.32	5.21	67.2	7.33
20	D(4096)L(0.05)–D(4096)L(0.05)–D(4096)L(0.05)–D(4096)L(0.05)	232	4.96	65.5	6.82
21	D(4096)L(0.1)–D(4096)L(0.1)–D(4096)L(0.1)–D(4096)L(0.1)	270	5.09	72.2	6.13
22	C(64) <sup>g</sup> E(0.05)–M <sup>h</sup> (2)–D(4096)E(0.05)–D(4096)E(0.05)	5.89	9.69	123	6.34
23	C(64)E(0.05)–M(2)–D(4096)E(0.05)–D(4096)E(0.05)–D(4096)E(0.05)	4.71	5.67	70.6	4.43
24	C(64)E(0.05)–M(2)–D(4096)E(0.05)–D(4096)E(0.05)–D(4096)E(0.05)–D(4096)E(0.05)	4.76	4.58	51.7	5.79
25	C(64)L(0.05)–M(2)–D(4096)L(0.05)–D(4096)L(0.05)	6.06	9.61	118	6.41
26	C(64)L(0.05)–M(2)–D(4096)L(0.05)–D(4096)L(0.05)–D(4096)L(0.05)	4.55	5.56	71.5	4.71
27	C(64)L(0.05)–M(2)–D(4096)L(0.05)–D(4096)L(0.05)–D(4096)L(0.05)–D(4096)L(0.05)	149	4.57	49.3	4.24
28	C(128)E(0.05)–M(2)–D(4096)E(0.05)–D(4096)E(0.05)	5.94	8.49	105	5.04

#	Hidden Layers	Min. Val. Loss ( $\times 10^5$ )			
		LR1*	LR2†	LR3‡	LR4§
29	C(128)E(0.05)–M(2)–D(4096)E(0.05)–D(4096)E(0.05)–D(4096)E(0.05)	4.77	5.19	63.5	4.40
30	C(128)E(0.05)–M(2)–D(4096)E(0.05)–D(4096)E(0.05)–D(4096)E(0.05)–D(4096)E(0.05)	5.46	4.29	45.6	4.88
31	C(128)L(0.05)–M(2)–D(4096)L(0.05)–D(4096)L(0.05)	5.99	8.62	105	5.18
32	C(128)L(0.05)–M(2)–D(4096)L(0.05)–D(4096)L(0.05)–D(4096)R	5.34	5.18	61.1	5.86
33	C(128)L(0.05)–M(2)–D(4096)L(0.05)–D(4096)L(0.05)–D(4096)L(0.05)	4.48	5.18	61.1	4.61
34	C(128)L(0.05)–M(2)–D(4096)L(0.05)–D(4096)L(0.05)–D(4096)L(0.05)–D(4096)L(0.05)	4.23	4.28	44.9	4.86
35	C(256)E(0.05)–M(2)–D(4096)E(0.05)–D(4096)E(0.05)	6.46	8.18	93.0	5.05
36	C(256)E(0.05)–M(2)–D(4096)E(0.05)–D(4096)E(0.05)–D(4096)E(0.05)	5.40	4.98	56.5	5.12
37	C(256)E(0.05)–M(2)–D(4096)E(0.05)–D(4096)E(0.05)–D(4096)E(0.05)–D(4096)E(0.05)	5.98	4.08	41.4	5.28
38	C(256)L(0.05)–M(2)–D(4096)L(0.05)–D(4096)L(0.05)	6.14	7.97	93.6	5.29
39	C(256)L(0.05)–M(2)–D(4096)L(0.05)–D(4096)L(0.05)–D(4096)L(0.05)	4.68	5.00	56.0	5.60
40	C(256)L(0.05)–M(2)–D(4096)L(0.05)–D(4096)L(0.05)–D(4096)L(0.05)–D(4096)L(0.05)	4.32	4.10	41.7	4.94
41	C(256)S–M(2)–D(4096)S–D(4096)S–D(4096)S–D(4096)S	11700	1400	—	—
42	C(256)T–M(2)–D(4096)T–D(4096)T–D(4096)T–D(4096)T	11800	49.5	—	—

**Notes.** Models trained for 20 epochs in batches of 64.

\* Triangular2 learning rate policy ranging from  $8 \times 10^{-6}$  –  $5 \times 10^{-3}$  with a complete cycle spanning 8 epochs.

† Like LR1, but ranging from  $10^{-5}$  –  $10^{-3}$ .

‡ Constant learning rate of  $10^{-5}$ .

§ Constant learning rate of  $10^{-3}$ .

<sup>a</sup> Dense layer with n nodes, D(n)

<sup>b</sup> ReLU activation

<sup>c</sup> Sigmoid activation

<sup>d</sup> tanh activation

<sup>e</sup> ELU activation  $E(\alpha)$ , with scaling parameter  $\alpha$

<sup>f</sup> Leaky ReLU activation  $L(m)$ , with a slope of  $m$  for  $x < 0$

<sup>g</sup> Convolution1D layer with a kernel size of 3 and n nodes, C(n)

<sup>h</sup> MaxPooling1D layer  $M(s)$ , with a pooling size  $s$

In Section 5.3, we note that we make the learning rate policy selection as described in Smith (2015), except based on the loss instead of the accuracy. Our selection process is to perform a ‘range test’ by training the model over a few epochs using a learning rate policy that constantly increases from a very small rate (e.g.,  $10^{-7}$ ) to a large rate (e.g.,  $10^{-1}$ ). Looking at a plot of loss vs. learning rate (e.g., Figure B.1), the learning rate range can be deduced based on when the loss begins decreasing (minimum learning rate) and when the loss begins increasing (maximum learning rate). In practice, we find more efficient training using a range that is slightly interior to the extrema determined via the plot. This is analogous to the method described in Smith (2015), except that it is more straightforward to determine the learning rate boundaries.

## B.1 List of References

Smith, L. N. 2015, arXiv e-prints, arXiv:1506.01186

## **APPENDIX C: DATA SET SIZE CONSIDERATIONS**



[This appendix is attached to the published manuscript that contains Chapter 5. It is Appendix B in the official publication.]

To briefly investigate the effect of data set size for our problem, we consider three models in addition to that presented in Section 5.3 (“Main”). The additional models are trained on around 25% of the total data set (614,208 training, 171,584 validation, 78,848 testing). Two models were trained according to the same learning rate policy as described in Section 5.3, for 187 and 500 epochs (“Sub1a” and “Sub1b”, respectively). The third model was trained according to the LR2 learning rate policy described in Appendix B for 500 epochs (“Sub2”). All other setup parameters (e.g., data normalization) match those described in Section 5.3.

Table C.1 compares the normalized RMSE and denormalized  $R^2$  test-set metrics over the high-resolution spectra, as well as the Bhattacharyya coefficients for the retrieved 1D marginalized posteriors. Based on the differences between models Sub1a and Sub1b (which only differ in the number of epochs trained), it can be concluded that manually stopping training once the loss begins to negligibly change does not have a major effect on the model performance. Models Sub1b and Sub2 (which only differ in learning rate policies) illustrate the importance of selecting the learning rate policy. However, both of these effects are negligible compared to those of the data set size: the differences among models Sub1a, Sub1b, and Sub2 are smaller than the differences between model Main and Sub2 (the best-performing Sub model). While Main and Sub2 underwent similar numbers of total training steps, Main outperforms Sub2. These results motivate the generation of large, comprehensive data sets of spectra to train surrogate RT models, though further research into how data set size, number of inputs/outputs, and architecture complexity influence model performance is needed to inform optimal the data set sizes for future investigations.

Table C.1: Model Comparison

Metric	Model	Min.	Median	Mean	Max.
Normalized RMSE	Main	0.00153	0.00224	0.00247	0.01040
	Sub1a	0.00271	0.00373	0.00407	0.01846
	Sub1b	0.00264	0.00365	0.00398	0.01679
	Sub2	0.00224	0.00331	0.00366	0.01721
Denormalized $R^2$	Main	0.99885	0.99993	0.99990	0.99997
	Sub1a	0.99646	0.99980	0.99974	0.99991
	Sub1b	0.99696	0.99981	0.99975	0.99992
	Sub2	0.99740	0.99983	0.99977	0.99994
Bhattacharyya coeff.	Main	0.9843	0.9948	0.9925	0.9972
	Sub1a	0.9585	0.9858	0.9853	0.9991
	Sub1b	0.8919	0.9933	0.9655	0.9976
	Sub2	0.9783	0.9940	0.9918	0.9984

**Notes.** See text for model descriptions.

**APPENDIX D: ASSESSING BAYESIAN FRAMEWORKS FOR  
EXOPLANET ATMOSPHERIC RETRIEVAL**

[This appendix is part of a manuscript in preparation for submission.]

## D.1 Introduction

Atmospheric retrieval is the inverse modeling process by which an object’s atmospheric properties are inferred from measured spectra (see review by Madhusudhan 2018). Due to the lack of spatial resolution as well as low signal-to-noise ratios of exoplanet observations, characterizing exoplanetary atmospheres requires a Bayesian statistical framework to properly infer the collection of atmospheric models that can explain the data (the posterior distribution). For 1D retrieval models, this process requires on the order of hours to days of runtime, typically on a multiprocessor (hours to weeks of compute cost, depending on model complexity).

Markov chain Monte Carlo (MCMC; see review by Hogg & Foreman-Mackey 2018) and nested sampling (NS; Skilling 2006) are the two most commonly used approaches to Bayesian retrievals of exoplanet atmospheres (e.g., Benneke & Seager 2013, Line et al. 2014, Lavie et al. 2017, Wakeford et al. 2017, Gandhi & Madhusudhan 2018, Barstow et al. 2020, Al-Refaie et al. 2021, Harrington et al. 2022, Ardevol Martinez et al. 2022, Himes et al. 2022). Theoretically, each Bayesian framework will retrieve the same posterior as the number of iterations approaches infinity. While this is not possible in practice, frameworks achieve an accurate approximation to the posterior after “enough” iterations have been performed (related to the effective sample size [ESS]; see Harrington et al. 2022 for more details). However, for computationally expensive forward models, “enough” iterations may require an unreasonable amount of compute time. Alternatively, if the Bayesian sampler is configured poorly, the inference may terminate before the posterior approximation is sufficiently accurate. Both situations produce numerically incorrect results, yet the end user may be unaware. A recent study by Ardevol Martinez et al. (2022) reported that  $\sim 8\%$  of their retrievals using MULTINEST (Feroz et al. 2009) yielded results that were  $> 3\sigma$  from the

truth for one or more parameters. While it is unclear whether this finding is due to the MULTINEST algorithm itself or if it is specific to how they set up MULTINEST (they do not describe in detail how they configured it), the study more importantly illustrates that seemingly legitimate results from a Bayesian inference could be incorrect. In exoplanet retrievals on real data, there is no way to know the ground-truth posterior, and thus it is imperative to characterize and understand any shortcomings of the Bayesian algorithm of choice.

In this study, we compare 9 Bayesian frameworks using toy problems and real-data retrievals to determine how their reported ESS relate to the yielded posteriors' accuracies as well as measure their relative runtimes to discern which algorithms are most efficient. Section D.2 describes each of the test problems, each Bayesian framework we consider, and how we apply each framework to each problem. Section D.3 presents and discusses the frameworks' results for each problem. Finally, in Section D.4, we draw conclusions from our results and provide recommendations for future investigations utilizing Bayesian frameworks.

## D.2 Methods

We utilize 3 test problems, described below, to compare the performance of 9 Bayesian frameworks, summarized in Table D.1. While not an exhaustive list of MCMC and NS implementations, it includes algorithms commonly used in exoplanet retrievals as well as those used in other areas of astronomy (e.g., Feroz et al. 2009, Buchner et al. 2014, Handley et al. 2015, Brewer et al. 2016, Cubillos et al. 2017, Buchner 2019, Barstow et al. 2020, Speagle 2020, Harrington et al. 2022, Ardevol Martinez et al. 2022, Himes et al. 2022). For tests with known posteriors, we consider the runtime, whether the algorithm finds all modes, and the mean root-mean-square error of the

differences between the 1D marginalized posteriors and the corresponding true distributions. For retrieval tests with unknown posteriors, we consider the runtime and how an algorithm’s yielded posterior compares to those of the other algorithms considered. Since our focus is on posterior accuracy, we configure the NS algorithms to prioritize deriving the posterior over deriving the evidence where possible.

We use the “eggbox” and “log gamma” toy problems described in Buchner (2016) to quantify each algorithm’s ability to capture multi-modal and long-tailed distributions. While not exoplanet retrievals, these problems provide insight into regimes where these posterior sampling algorithms may begin to fail. Additionally, we use a real-data retrieval on HD 189733 b as described in Harrington et al. (2022) to consider if differences seen in the toy problems similarly show up when retrieving on real data. We use the TRANSIT radiative transfer (RT) code (Rojo 2006, Cubillos et al. 2022) for the forward model in these retrieval tests. This enables direct comparisons with the results presented in Harrington et al. (2022), eliminating the forward model as a source of differences. Rather than repeat the detailed test descriptions here, we direct readers to the respective references as well as this work’s Reproducible Research Compendium (see link at end of Section D.4).

For each test, we strive for similar ESS (see Harrington et al. 2022 for MCMC calculation method and Speagle 2020 for NS calculation method) among the algorithms considered to more easily compare their results. In some cases, we additionally consider runs with higher ESS, either to properly solve the problem or to demonstrate the minimal compute cost needed to significantly improve the ESS. We run the algorithms in parallel where possible, utilizing 10 processors unless the algorithm requires more than that (DEMC requires twice as many parallel walkers as there are free parameters). Some algorithms are not considered for all problems due to the required runtimes.

Table D.1: Bayesian Frameworks Considered

Code	Type	References
DNest4	NS	Brewer et al. (2011), Brewer & Foreman-Mackey (2018)
dynesty	NS	Higson et al. (2019), Speagle (2020)
emcee	MCMC	Goodman & Weare (2010), Foreman-Mackey et al. (2013)
MCcubed's DEMC	MCMC	ter Braak (2006), Cubillos et al. (2017)
MCcubed's DEMC <sub>(ZS)</sub>	MCMC	ter Braak & Vrugt (2008), Cubillos et al. (2017)
PolyChordLite	NS	Handley et al. (2015), Handley et al. (2015)
PyDREAM	MCMC	Laloy & Vrugt (2012), Shockley et al. (2017)
PyMultiNest	NS	Feroz et al. (2009), Buchner et al. (2014)
UltraNest	NS	Buchner (2016, 2019, 2021)

**Notes:** Some of the provided references describe the algorithm/technique that is implemented in the code, though they are not necessarily involved in the code's development.

### *D.2.1 Software*

For this investigation, we developed the Large-selection Interface for Sampling Algorithms<sup>1</sup> (LISA), an open-source Python package released under the Reproducible Research Software License<sup>2</sup>. While not a Bayesian framework itself, LISA functions as a unified interface for the 9 Bayesian frameworks considered here, enabling users to easily switch between Bayesian frameworks for the problem of their choice. It offers reasonable defaults if users do not wish to finely tune the inference parameters, and it includes checks to ensure the inference is set up properly before beginning. We encourage users to contribute wrappers for additional Bayesian frameworks via pull requests on Github.

## D.3 Results and Discussion

We summarize the results for the eggbox problem in Table D.2; the 2D log gamma problem in Table D.3; the 10D log gamma problem in Table D.4; and the HD 189733 b retrieval case in Figure D.1, Figure D.2, and Table D.5.

---

<sup>1</sup><https://github.com/exosports/lisa>

<sup>2</sup><https://planets.ucf.edu/resources/reproducible-research/software-license/>



### *D.3.1 Eggbox Problem*

For the eggbox problem, the algorithms can be categorized into 3 groups: (1) those that find all modes and somewhat characterize the wings of each mode, (2) those that find all modes (or all except a few at the edge of the phase space) and do not characterize the wings of each mode, and (3) those that do not solve this problem. We find that MCCubed’s implementations of DEMC and  $\text{DEMC}_{(ZS)}$  fall into group 1 and achieve the most accurate posteriors for this problem. DNest4, PyDREAM, UltraNest, dynesty, PolyChordLite, and PyMultiNest fall into group 2, with the latter 3 achieving the fastest execution. Group 3 only contains emcee for this problem.

Note that this problem represents an extreme case of multimodality; while not typically seen in 1D exoplanet retrievals, it provides insight into the performance trends and limitations of these algorithms. In terms of performance, DNest4, dynesty, and PolyChordLite feature around a factor of 2 difference in runtime between the  $\text{ESS} \sim 1000$  and  $\text{ESS} \sim 2000$  cases, while PyMultiNest and UltraNest do not see a significant change in runtime between those ESS cases. Despite PyMultiNest’s fast execution time, it misses at least 1 mode at the edge of the phase space regardless of the ESS. Finding at most 3 of the 18 modes, emcee struggles to solve this problem, whether using its default “stretch move” sampling strategy of Goodman & Weare (2010) or the same sampling strategy as MCCubed’s  $\text{DEMC}_{(ZS)}$  (ter Braak & Vrugt 2008). When using emcee’s default “stretch move” sampling strategy of Goodman & Weare (2010), we find that it yields a significantly higher SPEIS compared to the DE and snooker updates. PyDREAM requires significantly more runtime than all other algorithms considered except emcee, likely attributed to its multi-try approach.

Table D.2: Performance Comparison: Eggbox Problem

Code	ESS	Relative runtime factor	Finds all 18 modes?	Mean RMSE
DNest4	500	4.0	Y	1.095
DNest4	1000	8.6	Y	1.042
DNest4	2000	18.2	Y	1.045
dynesty	500	14.8	Misses 1 mode at PS corner	1.124
dynesty	1000	1.0	Y	1.065
dynesty	2000	2.2	Y	1.042
emcee	500	182	Only finds 2 modes	5.691
emcee	1000	187	Only finds 3 modes	5.036
MCcubed's DEMC	500	1.6	Y	0.530
MCcubed's DEMC	1000	2.8	Y	0.514
MCcubed's DEMC <sub>(ZS)</sub>	500	9.0	Y	0.529
MCcubed's DEMC <sub>(ZS)</sub>	1000	16.8	Y	0.504
PolyChordLite	500	1.2	Misses 2 modes	1.152
PolyChordLite	1000	1.0	Y	1.147
PolyChordLite	2000	2.2	Y	1.120
PyDREAM	500	25.0	Y	1.028
PyDREAM	1000	41.6	Y	1.021
PyMultiNest	500	1.0	Misses 3 modes at PS edges	1.422
PyMultiNest	1000	1.6	Misses 1 mode at PS corner	1.113
PyMultiNest	2000	1.2	Misses 2 modes at PS edges	1.157
UltraNest	500	6.8	Y	1.064
UltraNest	1000	3.0	Y	1.028
UltraNest	2000	3.2	Y	1.025

**Notes.** The relative runtime factor is computed with respect to the minimum runtime among all runs for this problem. For the accuracy metric, we calculate the mean of the RMSE between each 1D marginalized posterior distribution and the known true distribution over the phase space (PS) considered. For emcee, we use the same sampling strategy as MCcubed's DEMC<sub>(ZS)</sub> (90% differential evolution updates, 10% snooker updates) because emcee's default sampling strategy (the "stretch move" of Goodman & Weare 2010) takes much longer to converge for this problem.

### *D.3.2 Log Gamma Problems*

For the log gamma problems, all considered algorithms correctly identify the 4 modes. We find that for  $\text{ESS} \sim 500$ , NS algorithms' posteriors are less accurate than MCMC algorithms. However, this is not a shortcoming of NS algorithms; in order to achieve this ESS value, it is necessary to limit the accuracy of NS algorithms, as typical runs yield  $\geq 2000$  ESS and more accurate posteriors.

In the 2D case, UltraNest and both of MCcubed's MCMC algorithms tie for the fastest algorithms, though in those specific cases the MCMC approaches are more accurate by roughly a factor of 2. All 4 MCMC approaches are able to achieve a lower mean RMSE at an ESS of 500 than almost every NS algorithm at any ESS considered. We note that dynesty appears to undersample the wings in the 3 runs considered; for the highest ESS run, it also undersamples the cores of the distributions. Utilizing a Gaussian filter as in dynesty's plotting routines can smooth out the errors and improve the accuracy, but we present the raw posteriors to more clearly compare the performances of the algorithms.

In the 10D case, we find that PyMultiNest achieves the best overall performance, requiring minimal runtime while still yielding an accurate result. While UltraNest represents the fastest overall execution time, that run features significant error; runs which yield a more accurate posterior require significantly more compute time. Both dynesty and PolyChordLite feature similar runtimes for a variety of ESS, highlighting the non-linear compute cost associated with some NS implementations. Among the MCMC algorithms, PyDREAM yields the least runtime, though both DEMC and  $\text{DEMC}_{(zS)}$  achieve more accurate posteriors at both ESS values considered.

Table D.3: Performance Comparison: 2D Log Gamma Problem

Code	ESS	Relative runtime factor	Mean RMSE
DNest4	500	25.5	0.458
DNest4	1000	52.6	0.301
DNest4	2000	98.2	0.251
dynesty	500	7.0	0.718
dynesty	1000	13.3	0.452
dynesty	2000	15.8	0.658
emcee	500	21.8	0.239
emcee	1000	36.5	0.174
MCcubed’s DEMC	500	1.0	0.275
MCcubed’s DEMC	1000	1.2	0.177
MCcubed’s DEMC <sub>(ZS)</sub>	500	1.0	0.229
MCcubed’s DEMC <sub>(ZS)</sub>	1000	1.8	0.178
PolyChordLite	500	3.2	0.915
PolyChordLite	1000	4.7	0.554
PolyChordLite	2000	7.7	0.391
PyDREAM	500	9.7	0.236
PyDREAM	1000	11.8	0.147
PyMultiNest	500	3.5	0.643
PyMultiNest	1000	3.8	0.417
PyMultiNest	2000	4.3	0.280
UltraNest	500	1.0	0.516
UltraNest	1000	1.2	0.455
UltraNest	2000	1.8	0.257

**Notes.** All algorithms find all modes. The relative runtime factor is computed with respect to the minimum runtime among all runs for this problem. For the accuracy metric, we calculate the mean of the RMSE between each 1D marginalized posterior distribution and the known true distribution over the phase space (PS) considered. For emcee, we use the same sampling strategy as MCcubed’s DEMC<sub>(ZS)</sub> (90% differential evolution updates, 10% snooker updates) because emcee’s default sampling strategy (the “stretch move” of Goodman & Weare 2010) takes much longer to converge for this problem.

Table D.4: Performance Comparison: 10D Log Gamma Problem

Code	ESS	Relative runtime factor	Mean RMSE
DNest4	500	4.48	0.991
DNest4	1000	6.43	0.662
DNest4	2000	11.1	0.407
dynesty	500	6.14	1.172
dynesty	1000	6.33	0.679
dynesty	2000	7.56	0.300
MCcubed’s DEMC	500	8.86	0.140
MCcubed’s DEMC	1000	14.2	0.102
MCcubed’s DEMC <sub>(ZS)</sub>	500	7.62	0.258
MCcubed’s DEMC <sub>(ZS)</sub>	1000	8.30	0.176
PolyChordLite	500	25.2	3.738
PolyChordLite	1000	31.8	3.138
PolyChordLite	4000	29.7	0.562
PolyChordLite	16000	118	0.247
PyDREAM	500	2.02	0.445
PyDREAM	1000	3.5	0.379
PyMultiNest	500	1.07	1.228
PyMultiNest	1000	1.35	0.687
PyMultiNest	2000	1.24	0.494
UltraNest	500	1.0	1.997
UltraNest	1000	1.47	1.044
UltraNest	3000	77.9	0.325
UltraNest	16000	105	0.129

**Notes.** All algorithms find all modes. The relative runtime factor is computed with respect to the minimum runtime among all runs for this problem. For the accuracy metric, we calculate the mean of the RMSE between each 1D marginalized posterior distribution and the known true distribution over the phase space (PS) considered. For emcee, we use the same sampling strategy as MCcubed’s DEMC<sub>(ZS)</sub> (90% differential evolution updates, 10% snooker updates) because emcee’s default sampling strategy (the “stretch move” of Goodman & Weare 2010) takes much longer to converge for this problem.

### D.3.3 HD 189733 b

Figures D.1 and D.2 show comparisons between the 1D marginalized posteriors for HD 189733 b for 7 algorithms. Table D.5 summarizes the ESS and execution times of each algorithm. Note that we do not consider DNest4 or PolyChordLite as they do not easily parallelize TRANSIT’s execution method. We also note that the relative runtime factors are not necessarily representative of these algorithms for all exoplanet retrievals but rather only represent their performances when used to parallelize the TRANSIT RT code. Other RT codes may interface better with some of these algorithms.

Overall, the algorithms agree. The notable exceptions are DEMC, which has two rogue chains that did not converge to the same solution and thus contains features not seen in the other posteriors; PyMultiNest, which undersamples the high-probability region for the log CO parameter; and PyDREAM, which oversamples the high-probability region for log CO. We emphasize that, even though PyMultiNest’s result is more physically plausible than the other algorithms, its disagreement with the other Bayesian frameworks identifies an algorithmic shortcoming. PyMultiNest’s behavior is not unexpected given the eggbox results, though further research is necessary to determine if MULTINEST generally struggles with characterizing high-probability regions that are on the edge of the phase space. While DEMC’s behavior is not observed in the earlier tests, it is not unusual when applied to exoplanet retrievals, based on our prior experiences. In theory, these rogue chains will converge if the chains continue for more iterations; however, in practice that can be significantly more iterations than the other MCMC algorithms considered here. The presented DEMC run features  $4\times$  the number of total iterations as the DEMC<sub>(ZS)</sub> run, thus highlighting the convergence and efficiency improvements of the DEMC<sub>(ZS)</sub> algorithm of ter Braak & Vrugt (2008) over the earlier DEMC algorithm of ter Braak (2006).

PyDREAM features the longest runtime, attributable at least in part to its multi-try approach. For computationally expensive forward models, as in this case, multi-try adds significant overhead per iteration. If a MT-DREAM algorithm were written in a way to better parallelize a program like TRANSIT, we expect that this overhead could be minimized. Regarding its agreement with the other algorithms, PyDREAM only features disagreement in the high-probability region for the log CO marginalized posterior, where it finds a much higher probability for the planet to contain a significant fraction of CO ( $> 10\%$ ). However, except for PyMultiNest, each algorithm finds this to be the most probable abundance for CO despite that this is not plausible for a hot Jupiter. Evidently, this unusually high abundance of CO provides a better numerical fit to the data, though the long, non-negligible tails of the distributions indicate a lack of a strong constraint, consistent with some previous studies (see Harrington et al. 2022 for additional discussion).

By using an even proportion of stretch moves and DE snooker updates, emcee achieves a higher ESS and lesser runtime compared to MCcubed's  $\text{DEMC}_{(ZS)}$ . When emcee's sampling strategy exactly matches MCcubed's  $\text{DEMC}_{(ZS)}$ , the ESS drops by around a factor of 3 for the same number of total iterations. This behavior was not observed in the toy problems. While this seems to suggest that the stretch moves help to better characterize the posterior, runs that only use stretch moves take longer to converge, suggesting that they are less efficient than the  $\text{DEMC}_{(ZS)}$  sampling strategy. Further investigation is necessary to determine the optimal MCMC sampling strategy for atmospheric retrieval as well as why allowing for both sampling methods improves convergence compared to using either independently.

UltraNest finds similar results to the other algorithms but at additional compute cost. Reducing the tolerance or number of live points accordingly reduces the compute cost, however that resulting posterior is not in close agreement with the other algorithms presented here. UltraNest's additional compute cost may be attributed to differences in how the algorithms parallelize likelihood evaluations.

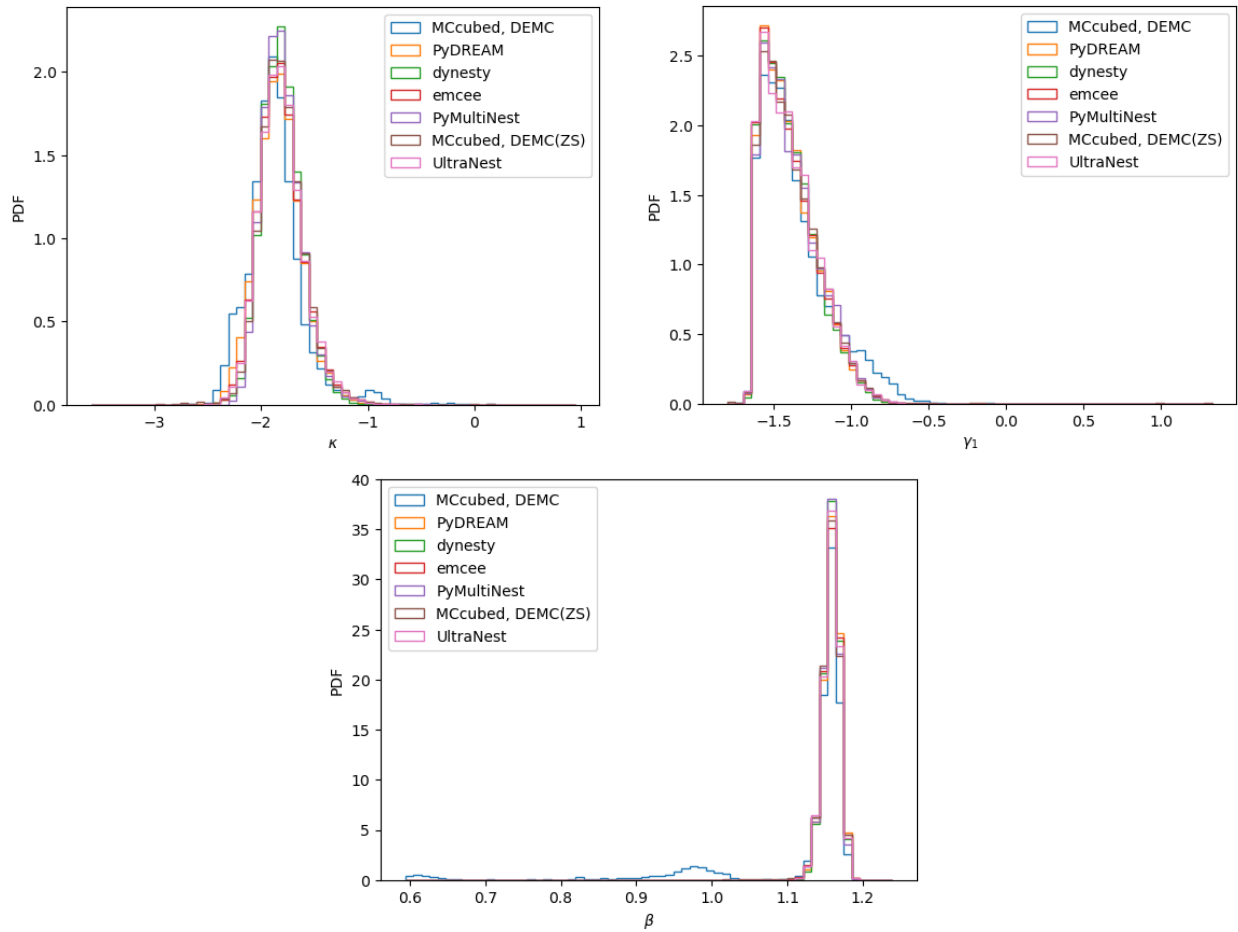


Figure D.1: Comparison of the thermal profile 1D marginalized posteriors for HD 189733 b between the 7 Bayesian frameworks listed.



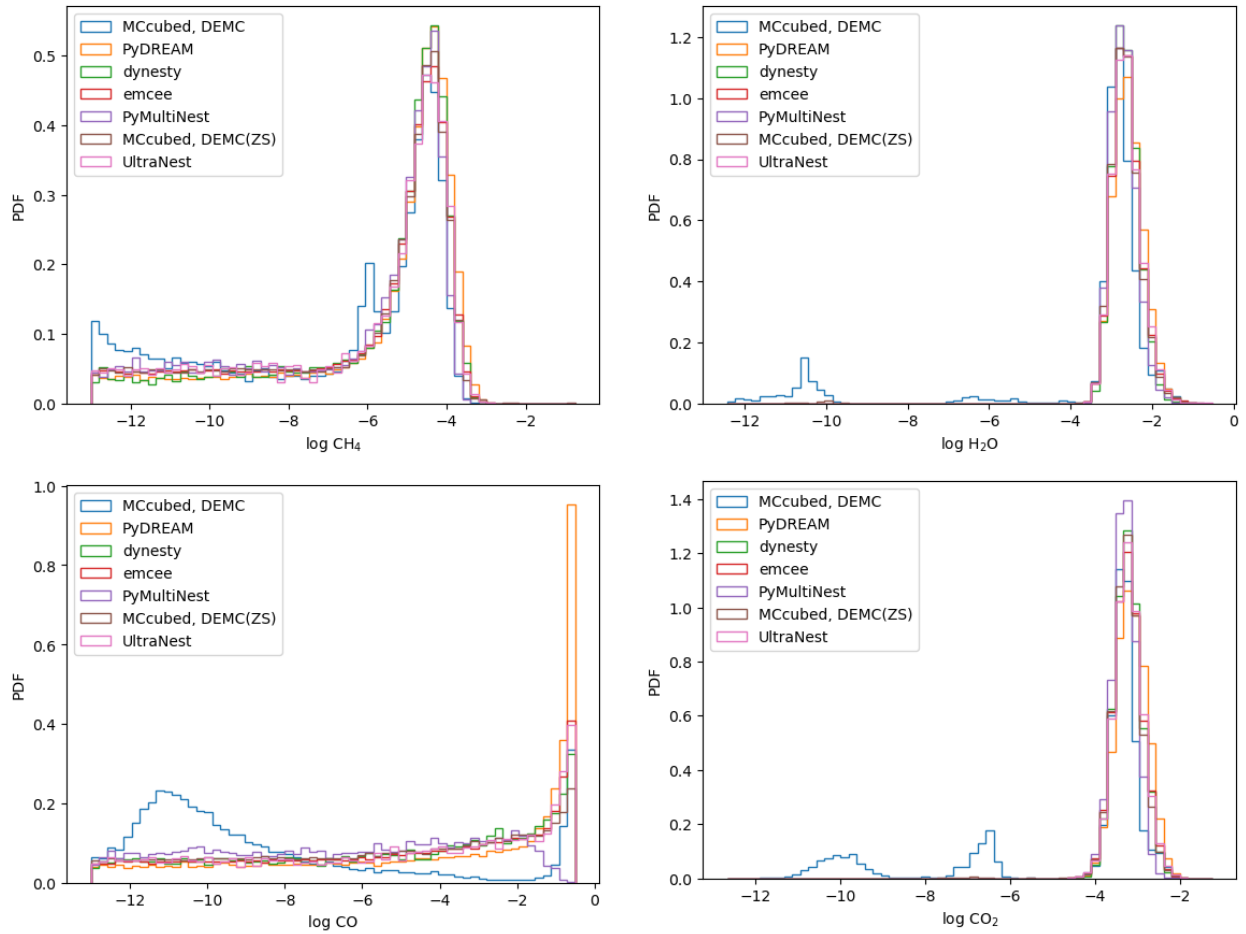


Figure D.2: Comparison of the gas abundance 1D marginalized posteriors for HD 189733 b between the 7 Bayesian frameworks listed.

Table D.5: Performance Comparison: HD 189733 b

Code	ESS	Relative runtime factor
dynesty	6317	2.0
emcee	723	2.3
MCcubed’s DEMC	65	2.7
MCcubed’s DEMC <sub>(ZS)</sub>	500	2.6
PyDREAM	1258	8.3
PyMultiNest	7476	1.0
UltraNest	3630	4.2

**Notes.** For emcee, we set the sampling strategy to an even mixture of its default “stretch move” and MCcubed’s DEMC<sub>(ZS)</sub> sampling strategy (90% differential evolution updates, 10% snooker updates), as we find it converges faster than using either approach on its own.

Overall, dynesty appears to best capture the posterior while minimizing compute cost. Despite that its compute cost is around double that of PyMultiNest, its 1D marginalized posterior for log CO more closely matches the other algorithms considered. Additionally, these results do not appear to exhibit the wing undersampling observed in the eggbox and 2D log gamma problems.

#### D.4 Conclusions

In this paper, we presented comparisons between 9 Bayesian frameworks using toy problems and a real-data retrieval case on HD 189733 b. In general, we find agreement between the algorithms considered. We also introduced an open-source Python package called LISA, which provides a unified interface for these 9 algorithms.

While all Bayesian algorithms should in theory agree with one another, in practice we find that some algorithms yield the wrong posteriors in certain regimes. Specifically, PyMultiNest appears to have difficulty characterizing high-probability regions at phase space edges; dynesty appears to undersample the wings of low-dimensional, multi-modal problems; and DEMC can have one or more rogue chains that delay convergence for exoplanet retrievals. We also find that, while emcee’s default stretch move delays convergence in the multi-modal toy problems considered, its inclusion improves the resulting ESS compared to only using DE and snooker updates. Further research is necessary to determine the optimal ratio of these update methods.

Specifically in the case of exoplanet retrievals, we find that dynesty, emcee, and MCCubed’s  $\text{DEMC}_{(ZS)}$  minimize compute cost while achieving posteriors that agree with the other algorithms. For retrievals where the high-probability region is not at the edge of the phase space, PyMultiNest may also be an efficient option. We emphasize that the other algorithms are in theory acceptable choices for exoplanet retrievals, but in our setup, the parallelization approaches of the other al-

gorithms are less efficient for our RT forward model. We recommend that future investigations utilize 2 or more Bayesian frameworks when reporting retrieval results to confirm that their results are independent of the posterior sampling algorithm. Looking ahead to the next generation of exoplanet modeling, minimizing overhead from the Bayesian posterior sampling algorithm without reducing posterior accuracy will be imperative as 2D and 3D exoplanet retrieval models become the standard.

## D.5 Acknowledgements

We thank contributors to NumPy, SciPy, Matplotlib, Tensorflow, Keras, the Python Programming Language, the free and open-source community, and the NASA Astrophysics Data System for software and services. This research was supported by the NASA Fellowship Activity under NASA Grant 80NSSC20K0682.

## D.6 List of References

Al-Refaie, A. F., Changeat, Q., Waldmann, I. P., & Tinetti, G. 2021, *ApJ*, 917, 37

Ardevol Martinez, F., Min, M., Kamp, I., & Palmer, P. I. 2022, arXiv e-prints, arXiv:2203.01236

Barstow, J. K., Changeat, Q., Garland, R., et al. 2020, *MNRAS*, 493, 4884

Benneke, B., & Seager, S. 2013, *ApJ*, 778, 153

Brewer, B. J., & Foreman-Mackey, D. 2018, *Journal of Statistical Software*, 86, 133

Brewer, B. J., Huijser, D., & Lewis, G. F. 2016, *MNRAS*, 455, 1819

Brewer, B. J., Pártay, L. B., & Csányi, G. 2011, *Statistics and Computing*, 21, 649

- Buchner, J. 2016, *Statistics and Computing*, 26, 383
- . 2019, *PASP*, 131, 108005
- . 2021, *The Journal of Open Source Software*, 6, 3001
- Buchner, J., Georgakakis, A., Nandra, K., et al. 2014, *A&A*, 564, A125
- Cubillos, P., Harrington, J., Loredó, T. J., et al. 2017, *AJ*, 153, 3
- Cubillos, P. E., Harrington, J., Blečić, J., et al. 2022, *The Planetary Science Journal*, 3, 81
- Feroz, F., Hobson, M. P., & Bridges, M. 2009, *MNRAS*, 398, 1601
- Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, *PASP*, 125, 306
- Gandhi, S., & Madhusudhan, N. 2018, *MNRAS*, 474, 271
- Goodman, J., & Weare, J. 2010, *Communications in Applied Mathematics and Computational Science*, 5, 65
- Handley, W. J., Hobson, M. P., & Lasenby, A. N. 2015, *MNRAS*, 450, L61
- Handley, W. J., Hobson, M. P., & Lasenby, A. N. 2015, *Monthly Notices of the Royal Astronomical Society*, 453, 4384
- Harrington, J., Himes, M. D., Cubillos, P. E., et al. 2022, *The Planetary Science Journal*, 3, 80
- Higson, E., Handley, W., Hobson, M., & Lasenby, A. 2019, *Statistics and Computing*, 29, 891
- Himes, M. D., Harrington, J., Cobb, A. D., et al. 2022, *PSJ*, 3, 91
- Hogg, D. W., & Foreman-Mackey, D. 2018, *ApJS*, 236, 11
- Laloy, E., & Vrugt, J. A. 2012, *Water Resources Research*, 48, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2011WR010608>

Lavie, B., Mendonça, J. M., Mordasini, C., et al. 2017, *AJ*, 154, 91

Line, M. R., Knutson, H., Wolf, A. S., & Yung, Y. L. 2014, *ApJ*, 783, 70

Madhusudhan, N. 2018, *Atmospheric Retrieval of Exoplanets* (Springer International Publishing AG), 104

Rojo, P. M. 2006, PhD thesis, Cornell University

Shockley, E. M., Vrugt, J. A., & Lopez, C. F. 2017, *Bioinformatics*, 34, 695

Skilling, J. 2006, *Bayesian Analysis*, 1, 833

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

ter Braak, C. 2006, *Statistics and Computing*, 16, 239

ter Braak, C. J. F., & Vrugt, J. A. 2008, *Statistics and Computing*, 18, 435

Wakeford, H. R., Sing, D. K., Kataria, T., et al. 2017, *Science*, 356, 628