The Planetary Science Journal, 4:74 (8pp), 2023 April https://doi.org/10.3847/PSJ/acc939

Toward 3D Retrieval of Exoplanet Atmospheres: Assessing Thermochemical Equilibrium Estimation Methods

Michael D. Himes[1], Joseph Harrington[2], and Atılım Güneş Baydin[3]
[1] Planetary Sciences Group, Department of Physics, University of Central Florida, USA; mhimes@knights.ucf.edu
[2] Planetary Sciences Group, Department of Physics and Florida Space Institute, University of Central Florida, USA
[3] Department of Computer Science, University of Oxford, UK

Received 2022 November 22; revised 2023 March 7; accepted 2023 March 23; published 2023 April 27

Abstract

Characterizing exoplanetary atmospheres via Bayesian retrievals requires assuming some chemistry model, such as thermochemical equilibrium or parameterized abundances. The higher-resolution data offered by upcoming telescopes enable more complex chemistry models within retrieval frameworks. Yet many chemistry codes that model more complex processes like photochemistry and 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 of <0.03 for the phase space considered. While our results are based on a 1D chemistry model, our study suggests that higher-dimensional chemistry models could be incorporated into retrieval models via this surrogate modeling approach.

Unified Astronomy Thesaurus concepts: Exoplanet atmospheric composition (2021); Exoplanet atmospheres (487); Neural networks (1933); Regression (1914); Chemical abundances (224)

Supporting material: figure set

1. Introduction

Improvements in the 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 (dis)equilibrium abundances for sub- and supersolar metallicities (e.g., Oreshenko et al. 2017; Mollière et al. 2020; Chubb & Min 2022).

Many exoplanet atmospheres are likely in disequilibrium (Moses et al. 2011, 2013b; Line & Yung 2013; Venot et al. 2015; Roudier et al. 2021). Stevenson et al. (2010) presented the first detection of disequilibrium chemistry in an exoplanet atmosphere, finding CH_4 depleted relative to thermochemical equilibrium calculations. 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 Changeat et al. (2022) found that some hot Jupiter spectra measured by the Hubble Space Telescope are not well fit by equilibrium chemistry models. 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). Furthermore, JWST's early-release science data for WASP-39b have led to the detection of SO_2, likely a product of photochemistry, providing additional motivation for more complex chemistry models within retrieval algorithms (Alderson et al. 2023; Rustamkulov et al. 2023).

Both the more detailed measurements of JWST and future telescopes and the recent results from 1D retrieval studies motivate more complex retrieval models. While the findings of Blecic 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. Community efforts are underway to develop such retrieval codes, though the computational costs associated with the increase in dimensionality require simplifying assumptions (e.g., Feng et al. 2020; Irwin et al. 2020; Changeat et al. 2021; Cubillos et al. 2021; Challener & Rauscher 2022; Chubb & Min 2022). With 1D retrievals requiring on the order of 10^5–10^6 forward model evaluations (Madhusudhan 2018), the difference of 1 s per forward model evaluation adds up to days of computing time, and higher-dimensional models multiply this further. While any chemistry model could be incorporated into a retrieval framework, the additional computational time of

Table 1
TEA Model Grid Parameters

Parameter	Minimum	Maximum	Number of Samples
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

Note. The mean abundance values correspond to solar abundance ratios as reported by Asplund et al. (2009). Temperatures are limited to 200–6000 K by the available data.

some models can be prohibitive, such as the Atmos photochemical-climate model (Arney et al. 2016) and the free molecular retrieval setup discussed in Chubb & Min (2022).

Recently, Himes et al. (2022) presented a novel machine-learning approach to exoplanet atmospheric retrieval, where the radiative transfer forward model is replaced with a neural network (NN) surrogate model. We 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 seek to determine a generalized method that can accurately approximate a given chemistry model, such that computationally expensive chemistry models could be included within a retrieval of any dimensionality. To investigate this, we compare thermochemical equilibrium estimation methods and consider each approximation's applicability to retrieval based on runtime and accuracy. In Section 2, we describe the analytical, interpolation-based, and NN-based models considered and detail our methodology. In Section 3, we present and discuss the results. Finally, we draw conclusions in Section 4.

2. 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 (using the approach of Himes et al. 2022).

2.1. 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 demonstrated 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).

2.2. Equilibrium via Analytical Formulae

Cubillos et al. (2019) presented the open-source RATE code, an analytical formalism to approximate thermochemical equilibrium that 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).

2.3. Equilibrium via Interpolation

We generate a grid of 74,088,000 points based on pressure, temperature, carbon-to-hydrogen elemental abundance ratio (C/H), nitrogen-to-hydrogen abundance ratio (N/H), and oxygen-to-hydrogen abundance ratio (O/H); the parameter minima, maxima, and number of uniformly log-spaced samples are given in Table 1. We consider the same molecules as Cubillos et al. (2019), with the addition of helium: H_2O, CO, CH_4, CO_2, HCN, C_2H_2, C_2H_4, NH_3, N_2, H_2, 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 a data set of 100,000 points—∼0.13% of the total TEA data set considered here—would require 372.5 GiB of memory to calculate the radial basis functions in double precision.

2.4. Equilibrium via NN-based Surrogate Model

Himes et al. (2022) presented an 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 as 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 log-uniform distributions spanning the ranges in Table 1. For data set (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 data set (2), we make 125,000 random draws of temperatures, C/H, N/H, and O/H; each of those cases is computed over the log-uniform grid of pressures defined in Table 1, producing a total of 12,500,000 combinations of the

Table 2
Training Differences between NN Models

Model	Data Set	Training Epochs
NN1	Grid of 53,240 TEA models	864
NN2	125,000 randomized TEA models	750
NN3	Randomized, sized to match gridded	500
NN4	Combination of gridded and randomized	500

Note. All TEA models are computed over a grid of 100 pressures.

five free parameters. We randomly split these cases such that approximately 64%, 16%, and 20% of the cases are in the training, validation, and test sets, respectively.

Each data case is comprised of the five input parameters given in Table 1 and the corresponding 12 output gas abundances. We normalize the data in two subsequent steps by (1) taking the base-10 logarithm of the inputs and outputs and (2) scaling the log data to be within the closed interval $[-1, 1]$ based on the training set extrema.

The NN model consists of an input layer with five nodes corresponding to the five 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; three dense layers with 4096 nodes, each of which uses 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 a range determined from the range test (see Appendix A of Himes et al. 2022).

Using the aforementioned data sets, we train four NN models based on the above architecture (Table 2). Model NN1 is trained for 864 epochs on the grid of TEA models. Model NN2 is trained for 750 epochs on the random data. Model NN3 is trained for 500 epochs on the random data but sized to match the gridded data set. Model NN4 is trained for 500 epochs on the combination of both the grid of TEA models and 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.

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 directly predict the molecular abundance profiles for an atmospheric model with some pressure–temperature profile and (possibly nonuniform) elemental abundance ratios. To do so, we normalize the inputs as described above, make the predictions, and denormalize the outputs.

2.5. Performance Assessment

We treat TEA as the control. We assess each non-TEA model by computing (1) the rms error (RMSE), coefficient of determination (R^2), and absolute dex errors (ADEs) over a grid of temperatures, pressures, metallicities, and C/O, and (2) the speedup factor compared to TEA. For item (1), we use a grid similar to that used in Cubillos et al. (2019) but 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.

Table 3
Model Performance Comparison

Model	RMSE	R^2	ADE	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

Note. 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 ADE for conciseness. For the full data set, see the RRC download link at the end of Section 4. 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.

Note that we compute the ADEs as

$$\mathrm{ADE} = |\log_{10}(\chi_{\mathrm{true}}) - \log_{10}(\chi_{\mathrm{pred}})|, \quad (1)$$

where χ_{true} is the molecular abundance calculated by TEA, and χ_{pred} is the molecular abundance predicted by a given model. To reduce the multidimensional matrix of ADEs to a single interpretable value, we tabulate the median ADE for each model. Additionally, we compute the speedup factors by averaging over all cases within a grid of temperatures, pressures, metallicities, and C/O.

3. Results and Discussion

Table 3 summarizes the median RMSE, median R^2, median ADE, 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 ± 1 s 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. Figure 1 provides two comparisons between the outputs of one NN model and the corresponding TEA case. The left panel is representative of well-modeled situations, showing close agreement between the NN and TEA for a temperature around 1000 K with solar metallicity and C/O; the right panel is representative of difficult situations for the NN to model, with a temperature around 2000 K with 2.5× solar metallicity and $\mathrm{C/O} = 0.9$. Figure 2 (and its online figure set) visualizes the model errors over the considered phase space, 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 the figures from NN4, though they are available in the online compendium (see Section 4).

Though the best linear interpolation model achieves the lowest median ADE, its RMSE and R^2 indicate it to be less accurate than the best IDW interpolation model, as well as most

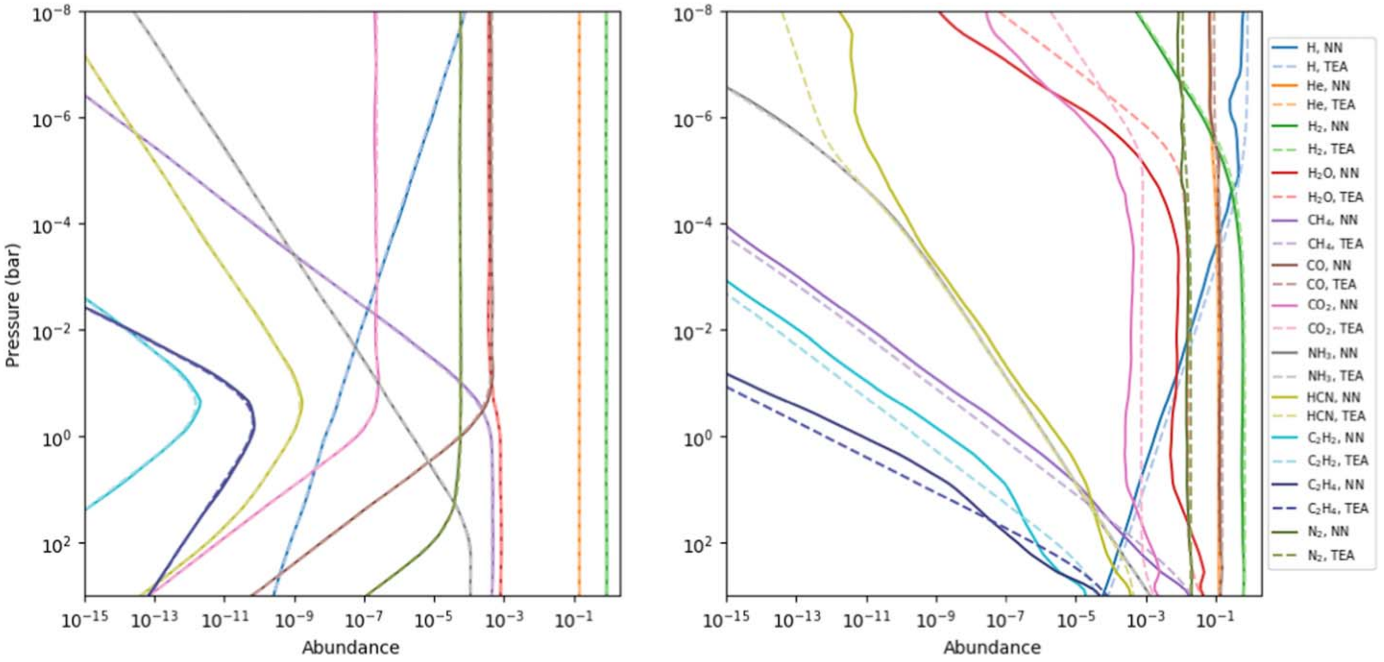

Figure 1. Two example cases comparing the outputs of one of the NN models with the corresponding TEA outputs. Left: case where the NN's predictions closely match the outputs of TEA, at a temperature of ∼1053 K, log C/H of −3.57, log N/H of −4.17, and log O/H of −3.31 over the full range of pressures considered in Table 1. Right: case where the NN performs less accurately, at a temperature of ∼2112 K, M/H of 2.5, and C/O of 0.9 (log C/H of −0.956, log N/H of −1.556, and log O/H of −0.910). Despite the less accurate performance, the NN's predictions are within an order of magnitude of the true values, except for H_2O, CO_2, and HCN at the lowest pressures. Only abundances $\geqslant 10^{-15}$ are shown for clarity.

of the NNs. This is confirmed via the mean ADEs; 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 result in significant relative errors (as indicated by the greater average dex error). This is confirmed by inspecting the error plots, as the model generally performs accurately but features significant errors in certain regions, e.g., for carbon- and oxygen-bearing species for $C/O = 0.9$. The interpolation error in nonlinear regimes could be minimized by more finely sampling in that region, though this may not be feasible for more computationally expensive chemistry models than considered here.

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 ADE, IDW achieves a lower median RMSE and slightly higher median R^2. Despite this, both models achieve similar average ADEs of ∼0.38. Table 4 details the performance of the other linear interpolation models considered in this investigation, while Table 5 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 anticorrelated with the ADE. 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 ADE is typically minimized when considering the two nearest neighbors along each axis. While considering only the nearest neighbor along each axis performs equivalently to or only slightly worse than the $n = 2$ case, considering three 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 ∼4× 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, a multidimensional retrieval model, or a global circulation 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. Model 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 2, 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 ∼7% the size of the grid used for interpolation. These similar

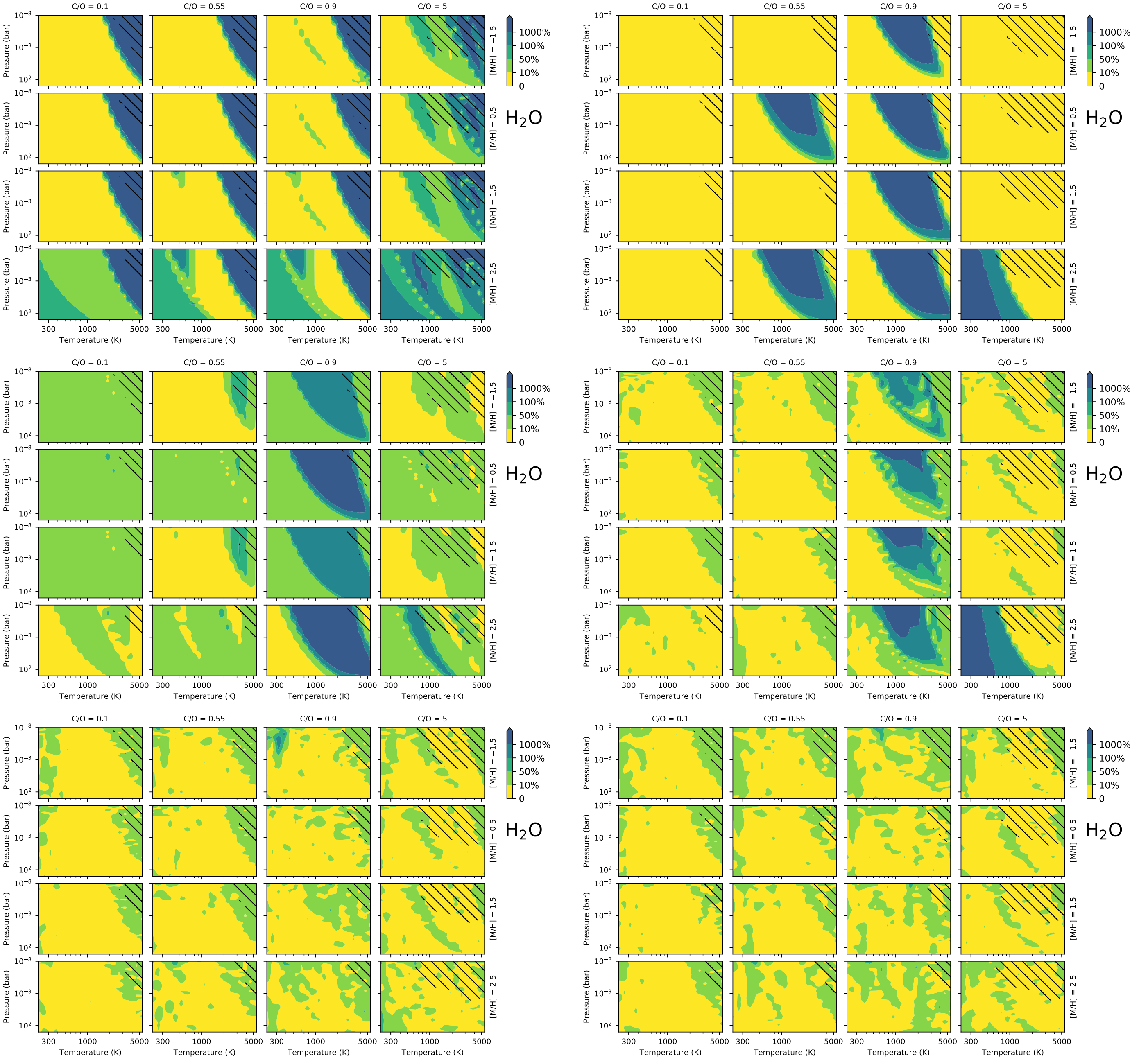

Figure 2. Absolute percent differences between the abundances predicted by the various approximation models and TEA for H_2O. Hatched regions indicate where the molecular abundance is less than 10^{-10}. Top left: RATE. Top right: best-performing linear interpolation model. Middle left: best-performing IDW interpolation model. Middle right: NN1. Bottom left: NN2. Bottom right: NN3. The complete figure set (12 images) is available in the online journal.

(The complete figure set (12 images) is available.)

Table 4
Linear Interpolation Model Performance Comparison

Number of Temperatures	Number of Elemental Abundances	RMSE	R^2	ADE	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

Note. For each model listed, we present only the median RMSE, R^2, and ADE for conciseness. For the full data set, see the RRC download link at the end of Section 4.

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 (the last five figures in the online figure set associated with Figure 2), while all carbon- and oxygen-bearing species bear this topography at varying magnitudes (the first seven figures in the online figure set associated with

Table 5
IDW Model Performance Comparison

Exponent	Neighbors per Axis	RMSE	R^2	ADE	Speedup Factor
1	1	1.1341	0.9830	0.0706	15,190
2	1	1.0780	0.9799	0.0619	15,250
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	15,000
5	1	0.8495	0.9840	0.0780	14,920
5	2	0.8502	0.9839	0.0722	2716
5	3	5.6693	0.3243	0.0767	841
5	4	7.5990	−0.0859	0.099 3	299
10	1	0.8301	0.9884	0.0781	14,930
10	2	0.8194	0.9886	0.0721	2188
10	3	5.6713	0.3345	0.0741	835
10	4	7.6006	−0.0631	0.099 6	299
20	1	0.8405	0.9877	0.0761	15,160
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	15,020
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	13,680
40	2	0.8316	0.9878	0.0719	2689

Note. For each model listed, we present only the median RMSE, R^2, and ADE for conciseness. For the full data set, see the RRC download link at the end of Section 4.

Figure 2). 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. Model NN3 was trained on a data set with ∼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 are randomly generated, rather than being generated on a fixed gridding. Model NN4 utilized both NN1's gridded data and NN3's random data—the most data out of any of the NN models. 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. Model NN4's performance falls between NN1 and NN2, which, together with NN1's performance, suggests that the gridded data force the NN toward a solution that, while optimal for those 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 3 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 making random draws from the phase space, few samples will have multiple parameters at the extrema (e.g., both $Z = 3$ and $C/O = 0.1$), resulting in reduced accuracy in that regime. While this could be addressed by preferentially drawing samples that are in these regions, doing so would bias the training data and thus would need to be handled carefully to ensure it does not bias the trained model. A more straightforward solution would be to generate data over a larger phase space than needed for the target application, thus ensuring the model is more accurate at the extrema of the target phase space.

4. Conclusions

In this study, we presented a comparison between thermochemical equilibrium estimation methods. We found that neural network (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). Our results suggest that these thermochemical equilibrium estimation methods can run fast enough and perform accurately enough to be used in place of a Gibbs minimization method for thermochemical equilibrium during retrievals. More importantly, these fast approximation approaches enable computationally expensive chemistry models, such as those considering disequilibrium processes, to be utilized within models requiring thousands of model calculations, such as Bayesian atmospheric parameter retrieval from spectra, global circulation models, etc.

For our application, our results show that NNs trained on random data outperform the considered interpolation methods even when the random data set size is ∼7% of the size of the uniform data set used for interpolation. 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. This finding is consistent with previous studies in the literature that showed that random sampling outperforms uniformly sampled grids when trying to approximate some forward model (Loyola et al. 2016; Fisher & Heng 2022). While this suggests that gridded data are less effective at training NNs compared to random data, a future study that more thoroughly investigates this behavior is necessary to determine this definitively.

When considering interpolation within a fixed grid, we found that interpolation error significantly increases in regimes of nonlinear behavior, such as $C/O \sim 1$. We thus recommend that investigators utilizing a grid-based interpolation approach verify that the interpolation is sufficiently accurate across the phase space of application. While a more finely sampled grid will reduce this error, the increased number of models would similarly benefit the NN approach. Based on our finding that the NN outperformed grid interpolation when considering a smaller gridded data set than the interpolation approaches, it is likely that the NN approach would continue to outperform the grid approach except at the highest resolutions, where the nonlinear behavior becomes nearly linear between grid points.

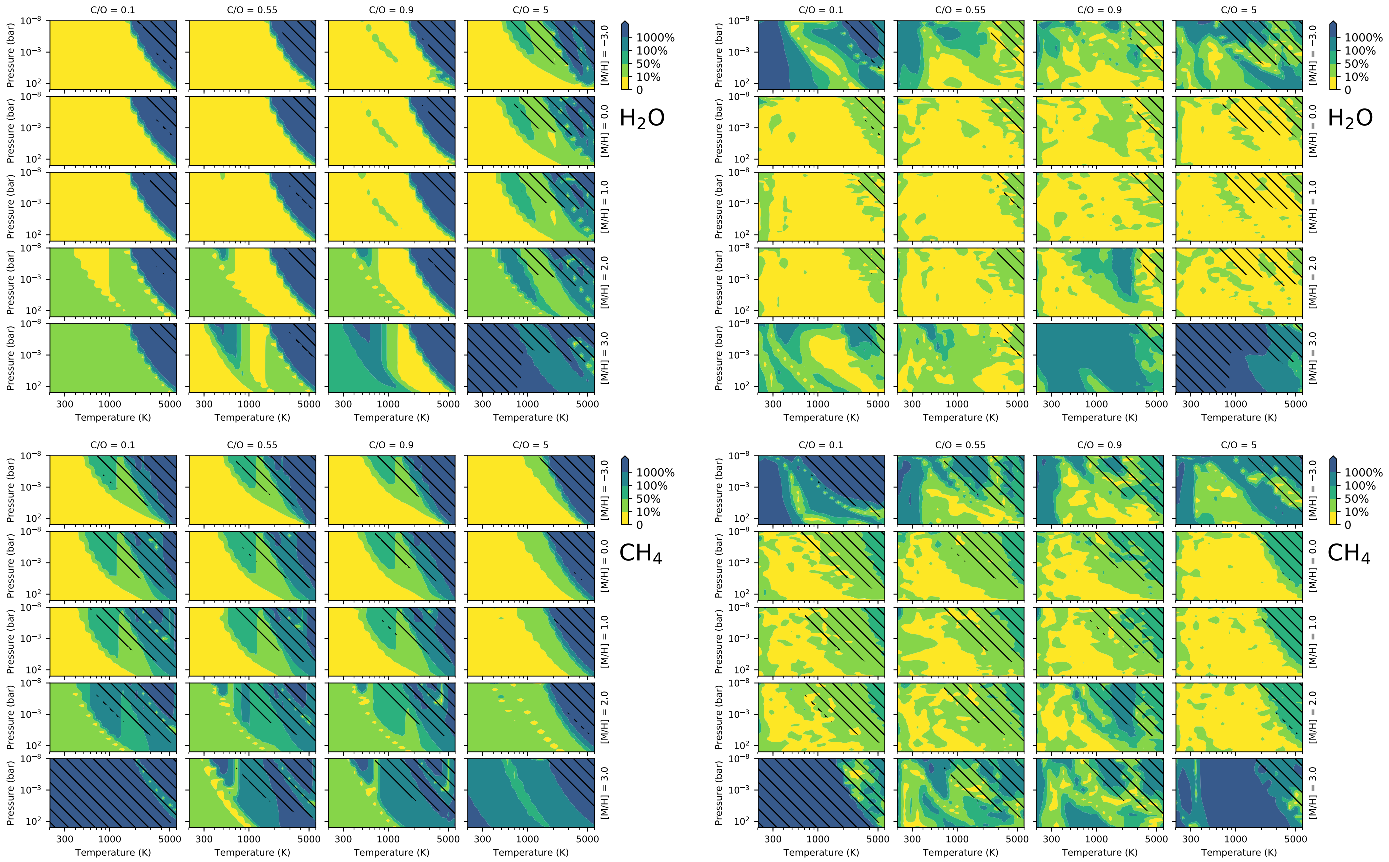

Figure 3. Same as Figure 2 but using the grid of Cubillos et al. (2019) to illustrate NN inaccuracies at phase space extrema. While RATE (left column) performs accurately at a metallicity of −3, predictions using NN3 (right column) for this metallicity feature significant error in certain pressure–temperature 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.

Future work should determine how finely sampled a grid must be for accurate interpolation across the phase space of interest.

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 the 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 (RRC) for this work is available for download at 10.5281/zenodo.7783281.

We graciously thank the anonymous reviewers for comments that improved the quality of this manuscript. 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.

ORCID iDs

Michael D. Himes https://orcid.org/0000-0002-9338-8600
Joseph Harrington https://orcid.org/0000-0002-8955-8531
Atılım Güneş Baydin https://orcid.org/0000-0001-9854-8100

References

Alderson, L., Wakeford, H. R., Alam, M. K., et al. 2023, Natur, 614, 664
Arney, G., Domagal-Goldman, S. D., Meadows, V. S., et al. 2016, AsBio, 16, 873
Asplund, M., Grevesse, N., Sauval, A. J., & Scott, P. 2009, ARA&A, 47, 481
Baydin, A. G., Heinrich, L., Bhimji, W., et al. 2019, in 33rd Conf. on Neural Information Processing Systems (NeurIPS 2019)
Blecic, J., Dobbs-Dixon, I., & Greene, T. 2017, ApJ, 848, 127
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, PhRvD, 98, 052004
Caldas, A., Leconte, J., Selsis, F., et al. 2019, A&A, 623, A161
Challener, R. C., & Rauscher, E. 2022, AJ, 163, 117
Changeat, Q., Al-Refaie, A. F., Edwards, B., Waldmann, I. P., & Tinetti, G. 2021, ApJ, 913, 73
Changeat, Q., Edwards, B., Al-Refaie, A. F., et al. 2022, ApJS, 260, 3
Chubb, K. L., & Min, M. 2022, A&A, 665, A2
Cubillos, P. E., Blecic, J., & Dobbs-Dixon, I. 2019, ApJ, 872, 111
Cubillos, P. E., Keating, D., Cowan, N. B., et al. 2021, ApJ, 915, 45
Feng, Y. K., Line, M. R., & Fortney, J. J. 2020, AJ, 160, 137
Fisher, C., & Heng, K. 2022, ApJ, 934, 31

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. 2017, in Proc. 34th Int. Conf. Machine Learning, 70, 1263
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
Irwin, P. G. J., Parmentier, V., Taylor, J., et al. 2020, MNRAS, 493, 106
Kasim, M. F., Watson-Parris, D., Deaconu, L., et al. 2021, MLS&T, 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
Loyola, R. D. G., Pedergnana, M., & Gimeno García, S. 2016, NN, 78, 75
Madhusudhan, N. 2012, ApJ, 758, 36
Madhusudhan, N. 2018, Handbook of Exoplanets (Cham: Springer), 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., Line, M. R., Visscher, C., et al. 2013b, ApJ, 777, 34
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
Munk, A., Ścibior, A., Baydin, A. G., et al. 2019, arXiv:1910.11950
Oreshenko, M., Lavie, B., Grimm, S. L., et al. 2017, ApJL, 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
Rustamkulov, Z., Sing, D. K., Mukherjee, S., et al. 2023, Natur, 614, 659
Smith, L. N. 2015, arXiv:1506.01186
Stevenson, K. B., Harrington, J., Nymeyer, S., et al. 2010, Natur, 464, 1161
Venot, O., Hébrard, E., Agúndez, M., Decin, L., & Bounaceur, R. 2015, A&A, 577, A33