The Astrophysical Journal Letters, 984:L32 (13pp), 2025 May 1 https://doi.org/10.3847/2041-8213/adc7aa

Near-instantaneous Atmospheric Retrievals and Model Comparison with FASTER

Anna Lueber[1,2], Konstantin Karchev[3], Chloe Fisher[4], Matthias Heim[1], Roberto Trotta[3,5,6,7], and Kevin Heng[1,8,9,10]
[1] Faculty of Physics, Ludwig Maximilian University, Scheinerstrasse 1, D-81679, Munich, Bavaria, Germany
[2] Center for Space and Habitability, University of Bern, Gesellschaftsstrasse 6, CH-3012 Bern, Switzerland
[3] Theoretical and Scientific Data Science, Scuola Internazionale Superiore di Studi Avanzati (SISSA), via Bonomea 265, 34136 Trieste, Italy
[4] Department of Physics, University of Oxford, Keble Road, Oxford, OX1 3RH, UK
[5] Astrophysics Group, Physics Department, Blackett Lab, Imperial College London, Prince Consort Road, London SW7 2AZ, UK
[6] INFN, National Institute for Nuclear Physics, Via Valerio 2, 34127 Trieste, Italy
[7] Italian Research Center on High-Performance Computing, Big Data and Quantum Computing, Via Magnanelli 2, 40033 Casalecchio di Reno, Italy
[8] ARTORG Center for Biomedical Engineering Research, University of Bern, Murtenstrasse 50, CH-3008, Bern, Switzerland
[9] University College London, Department of Physics & Astronomy, Gower St, London, WC1E 6BT, UK
[10] Astronomy & Astrophysics Group, Department of Physics, University of Warwick, Coventry CV 4 7AL, UK
Received 2025 February 25; revised 2025 March 27; accepted 2025 March 31; published 2025 April 28

Abstract

In the era of the James Webb Space Telescope (JWST), the dramatic improvement in the spectra of exoplanetary atmospheres demands a corresponding leap forward in our ability to analyze them: atmospheric retrievals need to be performed on thousands of spectra, applying to each large ensembles of models (that explore atmospheric chemistry, thermal profiles, and cloud models) to identify the best one(s). In this limit, traditional Bayesian inference methods such as nested sampling become prohibitively expensive. We introduce Fast Amortized Simulation-based Transiting Exoplanet Retrieval (FASTER), a neural-network-based method for performing atmospheric retrieval and Bayesian model comparison at a fraction of the computational cost of classical techniques. We demonstrate that the marginal posterior distributions of all parameters within a model and the posterior probabilities of the models we consider match those computed using nested sampling both on mock spectra and for the real NIRSpec PRISM spectrum of WASP-39b. The true power of the FASTER framework comes from its amortized nature, which allows the trained networks to perform practically instantaneous Bayesian inference and model comparison over ensembles of spectra—real or simulated—at minimal additional computational cost. This offers valuable insight into the expected results of model comparison (e.g., distinguishing cloudy from cloud-free and isothermal from nonisothermal models), as well as their dependence on the underlying parameters, which is computationally unfeasible with nested sampling. This approach will constitute as large a leap in spectral analysis as the original retrieval methods based on Markov Chain Monte Carlo have proven to be.

Unified Astronomy Thesaurus concepts: Exoplanets (498); Bayesian statistics (1900); Neural networks (1933)

1. Introduction

Measuring spectra of the atmospheres of exoplanets has become routine (e.g., A. L. Carter et al. 2024; G. Fu et al. 2025; J. Kirk et al. 2025). Encoded within these spectra is information on the physical and chemical properties of an atmosphere, which may be retrieved using Bayesian inference. Atmospheric retrieval is the approach of solving this inverse problem to extract the posterior distributions of atomic/molecular abundances, thermal profiles, cloud/haze[11] properties, and other features of an atmosphere from its measured spectrum (J. K. Barstow & K. Heng 2020). It is a crucial tool for utilizing atmospheres as chemical probes of exoplanets via remote sensing.

Since its introduction to the exoplanet literature by N. Madhusudhan & S. Seager (2009), atmospheric retrieval has been applied to transiting exoplanets (B. Benneke & S. Seager 2012; M. R. Line et al. 2013; I. P. Waldmann et al. 2015), directly imaged exoplanets (J.-M. Lee et al. 2013), and brown dwarfs (M. R. Line et al. 2015). Among likelihood-based Bayesian methods, nested sampling is one of the most widely employed techniques to obtain posterior samples (e.g., B. Benneke & S. Seager 2013), while in more recent years machine learning methods have been used to perform atmospheric retrieval as well (e.g., P. Márquez-Neila et al. 2018; T. Zingales & I. P. Waldmann 2018; A. D. Cobb et al. 2019; K. H. Yip et al. 2021; F. Ardévol Martínez et al. 2022; K. H. Yip et al. 2024). Currently used approaches face two major challenges: First, several models may be compatible with a given observed spectrum, and it would be useful to be able to compare several models at once with the data, in order to select the "best" (as measured by the Bayesian evidence), or else perform model averaging (e.g., D. J. C. MacKay 2003) to obtain more robust constraints. Second, while it is straightforward to write a Bayesian hierarchical model (e.g., A. Gelman et al. 2014) to study the parameters describing the population distribution of exoplanetary atmospheres, current Bayesian posterior samplers do not scale to the high-dimensional parameter space that such an approach would require.

Speaking to the first challenge, cogent motivation to move beyond traditional atmospheric retrieval approaches came from a recent study by A. Lueber et al. (2024), who analyzed four JWST spectra of the benchmark hot Jupiter WASP-39b. For each of these spectra, a family of models considered various combinations of seven chemical species, four treatments of temperature–pressure profiles, and both gray[12] and nongray[13]

[11] For the current study, we use the terms "cloud" and "haze" synonymously.

[12] Gray clouds have constant cross sections and consist of particles with sizes exceeding the wavelength of radiation being probed.

[13] Nongray clouds consist of particles with sizes smaller than the wavelength of radiation being probed and have wavelength-dependent cross sections.

clouds. This amounted to about 400 retrievals for analyzing these spectra of WASP-39b alone, even without considering vertically nonuniform chemical abundance profiles or more sophisticated cloud models. This amounted to a total computational budget of $\sim 10^4$ GPU-hr. Bayesian model comparison among all 400 models was performed to identify the simplest model giving a good explanation of the data (a quantitative implementation of Occam's razor)—clearly something that cannot be carried out manually for a population of hundreds of exoplanets.

In the current study, we introduce a new approach to atmospheric retrieval: simulation-based inference (SBI; for methodological reviews, see K. Cranmer et al. 2020; J.-M. Lueckmann et al. 2021). While SBI has recently gained popularity across various scientific domains,[14] it has seen only limited application in the analysis of exoplanetary atmospheres (M. Aubin et al. 2023; M. Vasist et al. 2023; F. Ardévol Martínez et al. 2024; T. D. Gebhard et al. 2025). We demonstrate that SBI can do the following:

1. Learn a probabilistic model from a large training suite of forward simulations of spectra that include physics of varying complexity (different cloud parameterizations, different treatments of temperature–pressure profiles, etc.) for any parameter combination within a prior box.
2. Once trained, perform almost instantaneous Bayesian inference on mock spectra and deliver posterior distributions of parameters that are consistent with a likelihood-based Bayesian analysis using nested sampling.
3. Perform near-instantaneous Bayesian model comparison over the entire ensemble of models being considered with minimal upfront computational cost.
4. Quantify the average performance of Bayesian model selection over the entire prior range considered, an analysis that is computationally impossible using traditional retrieval methods, as it requires thousands of retrievals per model.
5. Reproduce the WASP-39b retrieval analysis of A. Lueber et al. (2024), which was performed using nested sampling.

In Section 2, we describe our methodology, including our forward model (for computing synthetic spectra) and likelihood-based versus SBI approaches. In Section 3, we present our results. In Section 4, we discuss the implications of our findings, compare to previous work, and suggest opportunities for future work.

2. Methodology

2.1. Forward Model

A central ingredient in any atmospheric retrieval framework is the forward model that takes input parameters and computes a spectrum. A transmission spectrum is computed by determining the wavelength-dependent transit chord that has an optical depth of about unity (J. J. Fortney 2005), which is performed using ray-tracing through the atmosphere (T. M. Brown 2001).

We assume a one-dimensional, plane-parallel atmosphere with 99 discrete layers spanning pressures from $P = 10$ bars to 1 μbar equally spaced in $\log P$. The temperature–pressure profile is described using a finite-element approach, which allows smooth, continuous profiles to be constructed using a small number of parameters (see Section 2.5 of D. Kitzmann et al. 2020). In the current study, the temperature–pressure profile uses between one and four parameters, where a single-parameter description corresponds to an isothermal transit chord. Two, three, and four parameters allow for one, two, and three linear slopes in the temperature–pressure profile, respectively. This forward model and temperature–pressure profile parameterization have previously been implemented in the open-source `BeAR` retrieval code[15] (originally named `HELIOS-R2`; D. Kitzmann et al. 2020), which we use for the current study.

Indispensable inputs are the cross sections of atoms and molecules as functions of wavelength, temperature, and pressure. These cross sections are computed using spectroscopic line lists as inputs (S. L. Grimm & K. Heng 2015; S. L. Grimm et al. 2021) and at a spectral resolution of 0.01 cm^{-1}, but they are downsampled to 1 cm^{-1} when performing atmospheric retrievals. The line lists for water (H_2O; O. L. Polyansky et al. 2018), carbon dioxide (CO_2; S. A. Tashkun & V. I. Perevalov 2011), carbon monoxide (CO; G. Li et al. 2015), sulfur dioxide (SO_2; D. S. Underwood et al. 2016), and hydrogen sulfide (H_2S; A. A. A. Azzam et al. 2016) are used. Cross sections for sodium (Na) and potassium (K) are taken from D. Kitzmann et al. (2020). Collision-induced absorption associated with hydrogen–hydrogen (H_2–H_2; M. Abel et al. 2011) and hydrogen–helium (H_2–He; M. Abel et al. 2012) pairs is also included. These cross sections are publicly available through the DACE database[16] (S. L. Grimm et al. 2021). Each cross section is multiplied by the volume mixing ratio (relative abundance by number) of the respective atom or molecule when computing the total absorption cross section.

Atmospheres are generally expected to contain condensates or aerosols, often termed "clouds" or "hazes." While their composition is connected to that of the gas in the atmosphere, this chemical relationship is typically not modeled in retrievals. Rather, the cross section of the cloud is parameterized. For gray clouds, we parameterize the cross section using a constant optical depth. For nongray clouds consisting of spherical particles with a radius r_{cloud}, the cross section is given by $Q\pi r_{\mathrm{cloud}}^2$ where the extinction efficiency is given by Equation (32) of D. Kitzmann & K. Heng (2018),

$$Q \propto \frac{\tau_{\mathrm{cloud}}}{Q_0 x^{-a_0} + x^{0.2}}. \tag{1}$$

The cloud optical depth (referenced to its value at 1 μm) is given by τ_{cloud}. The size parameter is $x = 2\pi r_{\mathrm{cloud}}/\lambda$, where λ is the wavelength of radiation being probed. The slope index a_0 describes the slope of $Q(x)$ in the small-particle regime. The parameter Q_0 is a proxy for the particle composition (D. Kitzmann & K. Heng 2018) but is typically unconstrained in retrievals (e.g., A. Lueber et al. 2024). In both gray and nongray cases, we assume the cloud to be vertically semi-infinite in extent, which is appropriate for transmission spectra (as they do not probe that deeply into the atmosphere), and parameterize the top boundary using a "cloud-top pressure" (P_{cloud}).

We consider $N_{\mathrm{mod}} = 12$ distinct simulator configurations (*models*), formed by picking one among four temperature–pressure parameterizations ("TP1," "TP2," "TP3," "TP4") and one among three cloud models (cloudfree: "CF"; gray clouds:

[14] A list of applications is automatically being compiled at https://simulation-based-inference.org/.

[15] https://newstrangeworlds.github.io/BeAR

[16] https://dace.unige.ch

Table 1
Simulator Parameters, Their Prior Distributions, and Mock Values[b] Used

Parameter	Symbol	Mock Value[b]	Prior Range	Distribution	Units
Planetary surface gravity[d]	$\log g$	2.67	2.629 ± 0.051	Gaussian	cm s^{-2}
Planetary radius[d]	r_p	1.28	1.279 ± 0.051	Gaussian	R_J
Stellar radius[d]	r_s	0.95	0.939 ± 0.030	Gaussian	$R_\odot$
Volume mixing ratios[a]	X_i	⋯	$[10^{-12}, 10^{-1}]$	Log-uniform	⋯
Temperature	T_0	1000	[500, 3000]	Uniform	K
Slope between adjacent temperature nodes[c]	$b_{i=1\ldots4}$	⋯	[0.1, 3.0]	Uniform	⋯
Additional noise[e]	ς	0	[4, 420]	Log-uniform	ppm
Gray Clouds					
Cloud-top pressure	$P_{\rm cloud}$	10^{-3}	$[10^{-6}, 10^{1}]$	Log-uniform	bar
Optical depth	$\tau_{\rm cloud}$	500	$[10^{-5}, 10^{3}]$	Log-uniform	⋯
Nongray Clouds					
Cloud-top pressure	$P_{\rm cloud}$	⋯	$10^{-6}, 10^{1}$	Log-uniform	bar
Reference optical depth	$\tau_{\rm cloud}$	⋯	$10^{-5}, 10^{3}$	Log-uniform	⋯
Composition parameter	Q_0	⋯	[1, 100]	Uniform	⋯
Slope index	a_0	⋯	[3, 6]	Uniform	⋯
Spherical cloud particle radius	$r_{\rm cloud}$	⋯	$10^{-7}, 10^{-1}$	Log-uniform	cm

Notes.
[a] Values used for the mock retrievals are $X_{H_2O} = 10^{-3}$, $X_{CO} = 10^{-3}$, $X_{CO_2} = 10^{-4}$, $X_{H_2S} = 10^{-4}$, $X_K = 10^{-7}$, $X_{Na} = 10^{-4}$, and $X_{SO_2} = 10^{-6}$.
[b] Parameter values assumed for the mock retrievals.
[c] For the nonisothermal profiles, b_i is the slope between two adjacent temperature nodes (D. Kitzmann et al. 2020).
[d] Based on measured values reported in L. Mancini et al. (2018).
[e] We consider additional variance ranging from $s^2_{\rm min}/100$ to $4s^2_{\rm max}$, where $s_{\rm min}$ and $s_{\rm max}$ are the minimum and maximum values, respectively, of the reported observational uncertainties s_i.

"G"; nongray clouds: "NG"). Each model has a different number of parameters, ranging from 12 for the simplest model (TP1 CF: $\{T_0, \log g, r_p, r_s, X_{H_2O}, X_{CO}, X_{CO_2}, X_{H_2S}, X_K, X_{Na}, X_{SO_2}, \varsigma\}$) to 20 for the most complex model (TP4 NG). Each model features the TP1 CF set of 12 parameters, plus additional parameters that describe more complex temperature–pressure profiles (up to three additional parameters) and cloud parameters (up to five additional parameters; see Table 1). In all cases, we synthesize a (noiseless) transmission spectrum $\boldsymbol{F}$ (in parts per million (ppm)) in the range 0.53–5.34 μm with 207 wavelength bins, as appropriate for the PRISM mode of JWST's NIRSpec instrument.

2.2. Statistical Model

The observable $\boldsymbol{D}$ is a noisy measurement of the noiseless spectrum $\boldsymbol{F}$ with Gaussian uncertainties ς, which we assume to be independent across wavelength bins:

$$D_i \sim \text{Normal}(F_i, \sigma_i^2). \tag{2}$$

While the data reduction procedure reports an error *estimate* **s**, we allow for an additional contribution[17] ς to account both for potential mismodeling of instrumental effects and for the fact that currently no forward model can account for all the physics and chemistry present in a real atmosphere (D. Kitzmann et al. 2020). The two components are added in quadrature to give the total variance for the likelihood of Equation (2):

$$\sigma_i^2 = s_i^2 + \varsigma^2. \tag{3}$$

Since the reported uncertainties in JWST observations are on the order of 100 ppm, we expect a similar scale for ς. Nevertheless, we treat it as a free model parameter and infer it together with the other quantities in the simulator, after placing a wide prior that encompasses both the case that $\boldsymbol{s}$ fully captures the uncertainties and the possibility that they are grossly underestimated.

This and the remaining priors we use are listed in Table 1. For the most part, they are broad (log-)uniform distributions, except for the planetary radius, surface gravity, and stellar radius, for which we adopt Gaussian prior constraints from external measurements of the hot Jupiter WASP-39b (L. Mancini et al. 2018). This allows for comparison with the previous analysis by A. Lueber et al. (2024), which we reproduce for this study using the traditional nested sampling technique (J. Skilling 2006) implemented as `MultiNest`[18] (F. Feroz et al. 2009) within the open-source `BeAR` code. We consider the resulting parameter posteriors and model evidences as the ground truth with which to validate our methodology.

2.3. Simulation-based Inference

SBI[19] is an emerging comprehensive framework for Bayesian analysis that has been rapidly gaining popularity in

[17] Previously called "error inflation" in D. Kitzmann et al. (2020; and written as $10^{\epsilon} \equiv \sigma^2$), this additional variance is an ad hoc component necessitated by the imperfections of even the best likelihood-based fits. While in principle it can be increased in complexity by considering *different* additional noise levels σ_i across the spectrum or having them be correlated (e.g., as would arise from an "incorrect" subtraction of the stellar spectrum or improper calibration), the proper way to address this is by improving the explicit model both on the side of the physics and chemistry of the atmosphere and on the side of the instrument.

[18] Everywhere, `MultiNest` is run with1000 live points and a tolerance of $\Delta \ln Z = 0.5$.

[19] For reviews and a comprehensive comparison of methods, see K. Cranmer et al. (2020) and J.-M. Lueckmann et al. (2021). Extensive up-to-date lists of references to software and applications are kept at https://github.com/smsharma/awesome-neural-sbi and https://simulation-based-inference.org.

recent years, alongside developments in machine learning and deep neural networks (NNs). SBI (also called "likelihood-free inference") uses a stochastic model as a forward simulator that samples parameter values from given priors and produces plausible mock data, which are then used to train an NN to perform inference without the need to explicitly evaluate the likelihood function. The main advantages of SBI over likelihood-based Bayesian techniques are its ability to learn a complicated (possibly intractable) likelihood directly from mock data, its scalability to a large number of free parameters (implicitly marginalized in the simulator), and—crucial for this work—its ability to perform almost instantaneous parameter inference and model comparison once appropriately trained.

2.3.1. Marginal Parameter Inference with Neural Ratio Estimation

Any likelihood-based approach requires *joint* inference over the full set of model parameters, which can be computationally demanding even for parameter spaces of moderate dimension. However, in many cases, one is only interested in the *marginal* posterior of a few parameters of interest $\boldsymbol{\theta}$, with the remaining (nuisance) parameters $\boldsymbol{\nu}$ integrated out from the joint posterior:

$$p(\boldsymbol{\theta} \mid \boldsymbol{D}_0) \propto p(\boldsymbol{\theta}) \times p(\boldsymbol{D}_0 \mid \boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) \times \int p(\boldsymbol{D}_0 \mid \boldsymbol{\nu}, \boldsymbol{\theta}) \mathrm{p}(\boldsymbol{\nu} \mid \boldsymbol{\theta}) d\boldsymbol{\nu}, \quad (4)$$

where $\boldsymbol{D}_0$ are the observed data. In SBI, this integration is performed implicitly by stochastically sampling $\boldsymbol{\nu}$ while simulating training examples. While there exist several conceptually distinct variants of SBI (see, e.g., G. Papamakarios & I. Murray 2016; J.-M. Lueckmann et al. 2018; G. Papamakarios et al. 2018), here we focus on neural ratio estimation (NRE; J. Hermans et al. 2019); in this approach, the problem of inferring the posterior distribution is converted into the simpler task of binary classification, which NNs are particularly adept for. Given parameters of interest, $\boldsymbol{\theta}$, and simulated data, $\boldsymbol{D}$, the network is trained to approximate the ratio

$$r(\boldsymbol{\theta}, \boldsymbol{D}) \equiv \frac{p(\boldsymbol{\theta}, \boldsymbol{D})}{p(\boldsymbol{\theta}) p(\boldsymbol{D})} = \frac{p(\boldsymbol{D} \mid \boldsymbol{\theta})}{p(\boldsymbol{D})} = \frac{p(\boldsymbol{\theta} \mid \boldsymbol{D})}{p(\boldsymbol{\theta})}, \quad (5)$$

where the last equality follows from Bayes's theorem and shows that one can obtain the posterior in Equation (4) by simply multiplying the prior density $p(\boldsymbol{\theta})$ by $r(\boldsymbol{\theta}, \boldsymbol{D}_0)$ evaluated at the observed data. The key realization in NRE is that an NN can be trained to approximate Equation (5) through binary classification[20] of pairs $(\boldsymbol{\theta}, \boldsymbol{D})$ into two classes: one where $\boldsymbol{\theta}$ are the parameter from which $\boldsymbol{D}$ was generated ("joint" pairs from $p(\boldsymbol{\theta}, \boldsymbol{D})$), and another where $\boldsymbol{\theta}$ are randomly drawn from the prior ("marginal" pairs from $p(\boldsymbol{\theta})p(\boldsymbol{D})$).

In general, $\boldsymbol{\theta}$ can be an arbitrary group of parameters, although in the following we will specialize to a single parameter of interest, θ. This delivers the set of marginal one-dimensional posteriors, which are usually sufficient for drawing scientific conclusions. It is also easy to obtain a full "triangle plot," showing two-dimensional joint marginal posteriors among pairs of parameters, by having $\boldsymbol{\theta}$ represent each pair of parameters of interest (see, e.g., A. Cole et al. 2022, Figure 8). Moreover, the so-called auto-regressive extension of NRE (N. Anau Montel et al. 2024) can derive a full joint posterior over many ($\gtrsim$20) parameters of interest with the same simulation and training time budget. These techniques will be employed in future in-depth studies.

We simultaneously train neural ratio estimators $\hat{r}(\theta, \boldsymbol{D})$ for each individual parameter in a chosen model (e.g., $\theta \in \{T_0, \log g, r_p, r_s, X_{\mathrm{H_2O}}, X_{\mathrm{CO}}, X_{\mathrm{CO_2}}, X_{\mathrm{H_2S}}, X_{\mathrm{K}}, X_{\mathrm{Na}}, X_{\mathrm{SO_2}}, \varsigma\}$ in the TP1 CF model). After training, we evaluate each $\hat{r}(\theta, \boldsymbol{D}_0)$ at the observed data and multiply by the respective prior density, $p(\theta)$, to obtain a representation of the marginal posterior, $p(\theta \mid \boldsymbol{D}_0)$.

2.3.2. Bayesian Model Selection with SBI

While in traditional Bayesian pipelines model selection is typically understood as a separate task from parameter inference, in the context of SBI model selection can be seen as the pinnacle of marginal inference: the target is the posterior probability distribution of the model itself, $\mathcal{M}$, once all its parameters have been marginalized out:

$$p(\mathcal{M} \mid \boldsymbol{D}_0) \propto p(\mathcal{M}) \times p(\boldsymbol{D}_0 \mid \mathcal{M}) \propto p(\mathcal{M}) \times \int p(\boldsymbol{D}_0 \mid \boldsymbol{\nu}, \mathcal{M}) \mathrm{p}(\boldsymbol{\nu} \mid \mathcal{M}) d\boldsymbol{\nu}. \quad (6)$$

Here $p(\mathcal{M})$ is the prior model probability and $p(\boldsymbol{\nu} \mid \mathcal{M})$ and $p(\boldsymbol{D}_0 \mid \boldsymbol{\nu}, \mathcal{M})$ are, respectively, the prior and likelihood of *all* the parameters of the model, which are once again implemented in a stochastic simulator. Notice that if the models have entirely different parameter spaces (e.g., different number of parameters in each, or differences in the prior distributions), this is effortlessly taken care of in the simulator.

In our approach, an NN is trained to explicitly output the normalized posterior probabilities for all N_{mod} models considered. To this end, the training set combines simulations from each model in proportion to $\mathrm{p}(\mathcal{M})$ (i.e., in equal numbers, since we adopt a uniform prior in the space of models). When a mock example $\boldsymbol{D}$ is processed, the network's output that corresponds to the true model is maximized, but since *the same* $\boldsymbol{D}$ can be plausibly produced by *multiple* models (due to noise or parameter degeneracies), the final result is proportional to the posterior probabilities for the models. More formally, following L. Elsemüller et al. (2023) and K. Karchev et al. (2023a), we minimize the expected negative log-probability loss[21]

$$-\sum_{m=1}^{N_{\mathrm{mod}}} p(\mathcal{M} = M_m) \times \mathbb{E}_{p(\boldsymbol{D} \mid \mathcal{M} = M_m)}[\ln q_m(\boldsymbol{D})], \quad (7)$$

where q_m is the posterior probability that the NN assigns to M_m, obtained by normalizing the raw network outputs $[f_m]_{m=1}^{N_{\mathrm{mod}}}$ via the softmax function:

$$q_m \equiv \frac{\exp(f_m)}{\sum_{m'=1}^{N_{\mathrm{mod}}} \exp(f_{m'})}. \quad (8)$$

Notice that SBI delivers directly the posterior probabilities for all models, rather than the evidence $Z_m(\boldsymbol{D}_0) \equiv p(\boldsymbol{D}_0 \mid \mathcal{M} = M_m)$ for a single model, which is the result of a given likelihood-based analysis, e.g., with nested sampling. The set of evidences for all models can be converted to model

[20] Specifically, the network outputs a single real number $\ln \hat{r}(\boldsymbol{\theta}, \boldsymbol{D})$ and is trained using the standard binary cross-entropy loss functional.

[21] This is similar to parameter inference with neural *posterior* estimation, except that now the role of the parameter is taken by a discrete random variable that represents the model label. This simplifies the implementation of the necessary normalization, which can be performed explicitly over the discrete space.

probabilities according to Equation (6):

$$p(\mathcal{M} = M_m \mid \boldsymbol{D}_0) = \frac{Z_m(\boldsymbol{D}_0)p(\mathcal{M} = M_m)}{\sum_m Z_m(\boldsymbol{D}_0)p(\mathcal{M} = M_m)}. \quad (9)$$

Since this is of the same form as Equation (8), the network outputs f_m can be directly related to log-evidences (after training and *when the prior over models is constant*):

$$f_m(\boldsymbol{D}_0) = \ln Z_m(\boldsymbol{D}_0) + \text{constant}. \quad (10)$$

Therefore, they can also be used to compute the *Bayes factor* between any pair of models M_i, M_j:

$$\mathrm{BF}_{ij}(\boldsymbol{D}_0) \equiv \frac{Z_i(\boldsymbol{D}_0)}{Z_j(\boldsymbol{D}_0)} = \exp(f_i(\boldsymbol{D}_0) - f_j(\boldsymbol{D}_0)). \quad (11)$$

The Bayes factor balances each model's ability to fit the data (likelihood) while accounting for their model complexity (prior volume), thus including a quantitative measure of Occam's razor (R. Trotta 2008). Here we follow Jeffreys's scale for the strength of evidence as presented in R. Trotta (2008):

$$\ln \mathrm{BF}_{ij}(\boldsymbol{D}_0) \begin{cases} \in [1, 2.5] & \text{weak}, \\ \in [2.5, 5] & \text{moderate}, \\ > 5 & \text{strong}, \end{cases}$$

in favor of model M_i.

2.3.3. Advantages of Amortized Inference

A key advantage of SBI over traditional Bayesian techniques like nested sampling is the so-called *amortized* nature of its learning: this means that almost all of the computational effort is expended *only once* to train the NN for the tasks of interest (in our case, both parameter inference and model comparison), which are then carried out by querying the trained network—which requires very little computation. In this sense, the initial computational cost is recouped (i.e., amortized) when reusing the trained network multiple times later (on either real or simulated data).

In our case, the Fast Amortized Simulation-based Transiting Exoplanet Retrieval (`FASTER`) network is trained to perform inference on a large variety of possible data $\boldsymbol{D}$ rather than for only a particular observed $\boldsymbol{D}_0$. Thus, after the upfront cost of training the network (broadly comparable to a single standard nested sampling run), inferences from $\boldsymbol{D}_0$ (i.e., both marginal parameter posteriors and the posterior probabilities of all models) can be obtained at almost no computational cost via a single forward pass, which takes on the order of milliseconds. In contrast, even GPU-accelerated nested sampling retrievals take $\sim$10 hr *per model*. Importantly, SBI can be performed with similar speed on *any* other data: e.g., further observations or mock examples, provided that they are assumed to be plausible realizations from the simulator used for training, while nested sampling needs to be rerun from scratch every time.

Apart from the massive speedup in inference that this entails, amortization enables several tests and calibrations that are normally impossible with traditional Bayesian methods. Analyses of a *validation set* of simulated examples (which are not used for training) can be used to examine the Bayesian (i.e., averaged over parameters sampled from the priors) coverage properties of the approximated posteriors (see also S. R. Cook et al. 2006; S. Talts et al. 2018; J. Hermans et al. 2021). Another possibility is to construct from a large number of simulated reconstructions confidence regions with *exact* and guaranteed frequentist coverage (see also N. Dalmasso et al. 2020, 2021; L. Masserano et al. 2022; K. Karchev et al. 2023b). In the case of model selection, one can test using a large number of simulated examples the *reliability* of the posterior probabilities, as well as the ability of the network—averaged over repeated draws from the parameters' priors—to identify the correct model (M. H. DeGroot & S. E. Fienberg 1983, see also K. Karchev et al. 2023a). Finally, analyses of numerous simulated data allow investigation of the dependence of model selection results on the true parameter values, which is computationally unfeasible in a traditional Bayesian setting, as it would require a prohibitive number of nested sampling retrievals.

2.3.4. Simulations, Networks, and Training

For each of the $N_{\mathrm{mod}} = 12$ models we consider, we simulate 10^5 mock *noiseless* spectra with parameters randomly sampled from the priors in Table 1. Observational noise, according to Equations (2) and (3), is then added on the fly, i.e., a different realization is used in each training epoch, as a form of simple and fast data augmentation. Each training set separately is used to train a parameter estimation network for the given model it has been generated from, while their concatenation (amounting to 1.2×10^6 examples) is used to train the model selection network.

We use simple fully connected NNs (depicted in Figure 1) for both parameter inference (NRE) and model selection. In both cases, the raw data $\boldsymbol{D}$ are first preprocessed by a multilayer perception (MLP) and embedded in 256 or 512 dimensions (for NRE and model selection, respectively). The resulting featurization is shared among the ratio estimators for all parameters (again implemented as MLPs), whereas for model selection it is directly converted into model probabilities by a linear layer of output size $N_{\mathrm{mod}} = 12$.

Generating 1.2×10^6 spectra took 6 hr on an NVIDIA RTX4090 GPU. The subsequent training[22] took additionally about 3 hr per network. We emphasize that this upfront computational effort needs to be expended only once, after which the trained networks can be deployed on any data—simulated or real, provided that they are assumed to be a plausible sample from the simulator—at almost no computational cost.

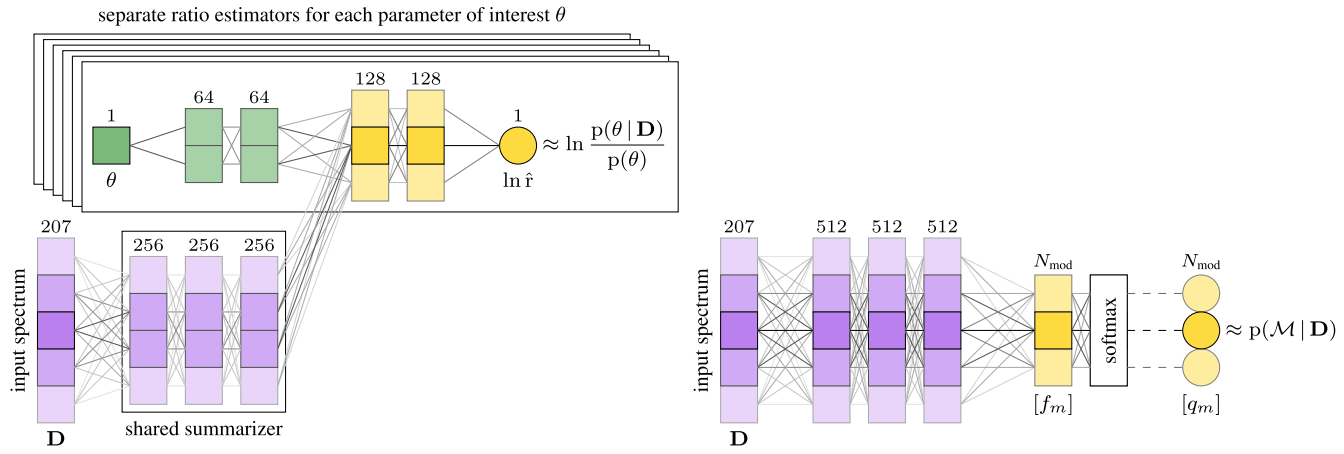
3. Results

3.1. Validation on Mock Data

First of all, we wish to validate `FASTER` against the ground-truth results from `MultiNest` in a fully controlled environment, i.e., by analyzing a mock example. For this illustration, we assume the TP1 G model (isothermal with gray clouds) and simulate a noisy spectrum with parameter values as listed in Table 1. No additional intrinsic variance is added to the measurement errors reported by JWST for WASP-39b, which corresponds to $\varsigma = 0$.

Figure 2 demonstrates that `FASTER` produces posterior distributions that are consistent with all of the input parameter values. Not only are these posteriors consistent with those

[22] We train for 100 (fixed) epochs and use the checkpoint with the best validation loss.



(a) parameter-inference (NRE) network (b) model-selection network

Figure 1. Schematic of the FASTER inference networks, which we implement as MLPs. (a) For a given choice of model, we train simultaneously ratio estimators $\hat{r}$ to approximate Equation (5) for each parameter θ within that model. This allows evaluation of the marginal posteriors when queried with the observed data and multiplied by the respective priors $p(\theta)$. (b) A model selection network is trained to estimate directly the (normalized) posterior probabilities of all $N_{\mathrm{mod}} = 12$ models considered in this Letter.

generated by our nested sampling retrieval, but for several parameters the shapes of these posteriors are also identical. This is remarkable particularly for parameters such as the chemical abundances and $\log \tau_{\mathrm{cloud}}$ that can vary over several orders of magnitude. We remind the reader that our training is fully amortized across the whole prior range—meaning that an almost instantaneous inference of the same quality can be obtained from the trained network for *any* value of the underlying parameters across that range.

In the bottom right panel of Figure 2, we also show the reconstructed spectra from the two methods, compared to the simulated data. For this purpose, we sample 1000 parameter values according to the respective NS and SBI posteriors and plot the distribution (median and ± 1 standard deviations) of the corresponding synthetic spectra. While NS gives access to the joint 13-dimensional posterior to sample from, with SBI we approximate only the one-dimensional projections. To plot the SBI reconstruction, therefore, we form an approximation to the joint by multiplying the marginals; this necessarily leads to a wider (conservative) result, which is reflected in the greater spread in the reconstructions (see, in particular, at low wavelengths).

Next, we compare the results of likelihood- and simulation-based model selection, again on mock data: in addition to the example discussed above, we simulate a realization from the simpler cloudfree model (TP1 CF) with the same parameter values (where relevant). While the inference network directly delivers all $N_{\mathrm{mod}} = 12$ posterior probabilities simultaneously, with nested sampling we need to run 12 separate retrievals, each with a different model (and then convert the evidences to posterior probabilities via Equation (9), assuming equal prior probabilities). In practical terms, this translates to approximately 100 hr of NS retrievals, compared to 24 ms for the single network evaluation.

The results are compared in Figure 3: we notice excellent agreement between the two approaches, with SBI probabilities falling generally within the estimated uncertainty from the NS runs. While there is stochasticity in the initialization of the NN, once converged this is likely a subdominant source of error in the probability estimate from FASTER. A more important role is played by the "systematics" uncertainty coming from choice of network architecture, training epochs, batch size, learning rate, number of training examples, etc., which is difficult to estimate reliably. We therefore limit our output to a point estimate of the model probability. Additionally, when performing inference on real data using the likelihood-based nested sampling algorithm or NRE, we can expect that any model misspecification (i.e., when the real data are generated in a manner that is not exactly captured by our forward model) will lead to the resulting probability estimates being different between the two methods. Concretely, in the presence of clouds (right panel: mock data from the TP1 G model), cloudfree models are decisively excluded, and moreover the type of clouds (gray/nongray) is correctly identified. On the other hand, when the true model *is* cloudfree (left panel), cloudy models cannot be excluded because their prior does include low values of τ_{cloud}, in which case the spectrum is indistinguishable from cloudfree. Moreover, whereas the temperature–pressure profile of a cloudfree atmosphere can be reliably identified, this is not the case when clouds are present. These points highlight the importance of priors on model selection results and the latter's dependence on the underlying parameter values. We now proceed to explore these two effects further, employing the unique amortization property of SBI.

3.2. Occam's Razor and Distinguishing Cloud Models

Having validated SBI on concrete examples, we turn to investigating the results of Bayesian model selection as a general methodology (i.e., irrespective of the technique used). Still, the following demonstrations are only made possible by the amortization inherent in SBI, which allows results to be rapidly derived from numerous simulated examples.

Our first setup aims to determine the values of cloud-related parameters (τ_{cloud} and P_{cloud}) that—for the given JWST noise levels—*would* lead to a positive identification of a cloudy versus cloudfree model. To this end, we simulate validation sets of 9000 mock spectra each from the gray-cloud and nongray-cloud models[23] and derive Bayes factors, via Equation (11), from all of them with respect to the cloudfree model. While with nested sampling this extreme exercise

[23] For this illustration, we focus on isothermal (TP1) models. The remaining parameters (including ς) are again sampled from the priors in Table 1.

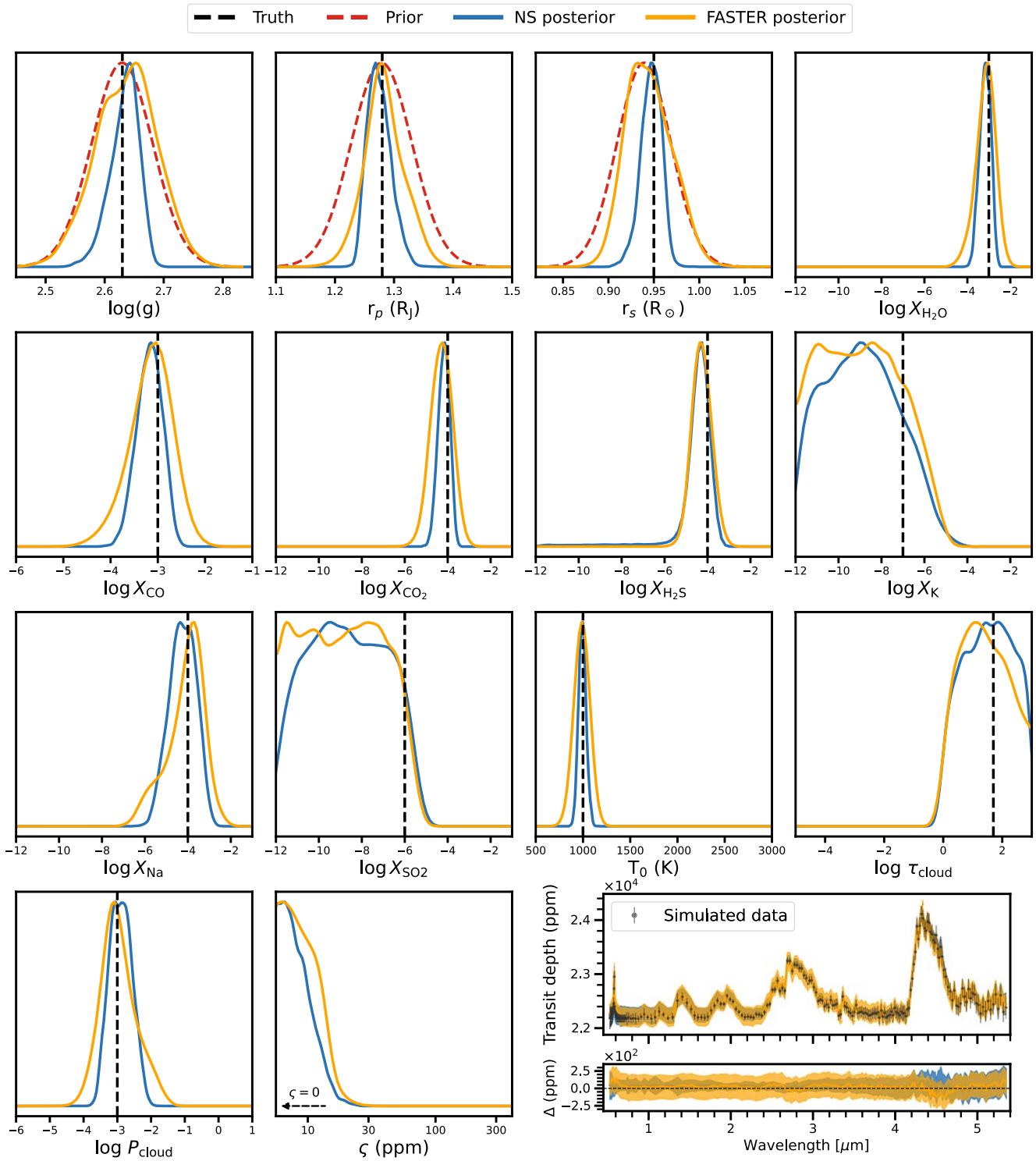


Figure 2. Retrievals assuming the TP1 G model (isothermal profile and gray clouds) performed on a simulated spectrum from the same model and with parameter values indicated by black vertical dashed lines. The results (one-dimensional marginal posteriors, normalized to their peak) from nested sampling and SBI (blue and orange) are plotted as solid lines, while priors are depicted as dashed red lines, where not uniform across the plotted range. The bottom right panel shows the simulated data (error bars, containing only the reported instrumental noise, i.e., $\varsigma = 0$ as simulated) and the reconstructions from the two methods, derived by simulating spectra from the posterior samples.

would have cost $\sim 10^5$ GPU-hr, it took us a mere 0.6 s in parallel on a single GPU.

The outcomes are plotted in the top panels of Figure 4, where the axes correspond to the *true* parameters from which the mock data were generated, and the color scale reflects the NN-derived Bayes factors (equivalent to posterior odds since we take equal prior $p(\mathcal{M})$). The revealed pattern aligns well with our physical intuition about transmission spectra: when the cloud optical depth exceeds unity, Bayesian model comparison strongly ($\ln \mathrm{BF} > 5$) favors a cloudy model. (For the nongray-cloud model, recall that the optical depth is wavelength dependent and referenced to 1 μm, which is within the observed range.) But this also depends on the

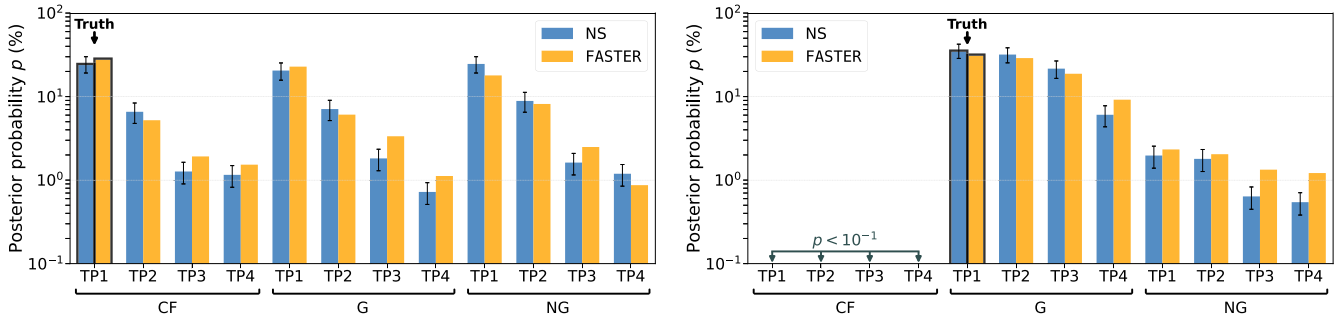


Figure 3. Comparison of model posterior probabilities derived with SBI (orange) and nested sampling (blue) from mock data generated from the TP1 CF (isothermal, cloudfree: left panel) or TP1 G (isothermal, gray clouds: right panel) models.

cloud not residing too deeply in the atmosphere: if the cloud-top pressure resides at pressures greater than 0.1 bars, the cloud has a vanishingly small effect on the spectrum, and therefore the cloudfree model is preferred by Bayesian model comparison in virtue of Occam's razor (yellow and green regions in the top panels).

In the bottom panel of Figure 4, visual inspection of particular spectra (labeled "A" and "C") confirms that the cloudy models are indistinguishable from one another and from the cloudfree case in the optically thin limit. The latter—simpler—model is both a sufficient and physically correct description of the data. This failure to reject the cloudfree model, when it is indeed false, would be a "type II" error in the frequentist sense. However, from a Bayesian model comparison perspective this is indeed the correct conclusion: if the data do not allow us to distinguish between two models (cloudfree vs. cloudy), selecting the simpler one is in accord with the Occam's razor principle. We investigate this point on average across parameter values in the next section.

3.3. Average Performance of Bayesian Model Comparison

In the previous section, we scrutinized Bayesian model comparison for individual examples as a function of the underlying parameters. Here we instead turn to the results *averaged* across parameters sampled from each model's prior distributions and examine the "expected model posterior probability" (EMPP) matrix.[24] Its construction again relies on the amortized nature of SBI and is otherwise impossible using traditional nested sampling retrievals.

In detail, we simulate 9000 mock realizations from a given model (with parameters drawn from the priors as above (including ς) and evaluate the trained model selection network on them to obtain the corresponding posterior probability distributions $[q_m]$. Averaging them across the 9000 examples produces a single row of the EMPP matrix; repeating this with each of the 12 models gives the full 12×12 matrix depicted in Figure A1. For clarity, we further group models according to temperature–pressure profile (TP1–4) or cloud type (CF, G, NG) and obtain the matrices shown in Figure 5.

The interpretation of an EMPP matrix is as follows. The values on the diagonal represent the probabilities assigned to the "correct" model, i.e., the one that data were generated from, and so it is in general desirable that these entries are close to unity. However, since a given data set may have been plausibly

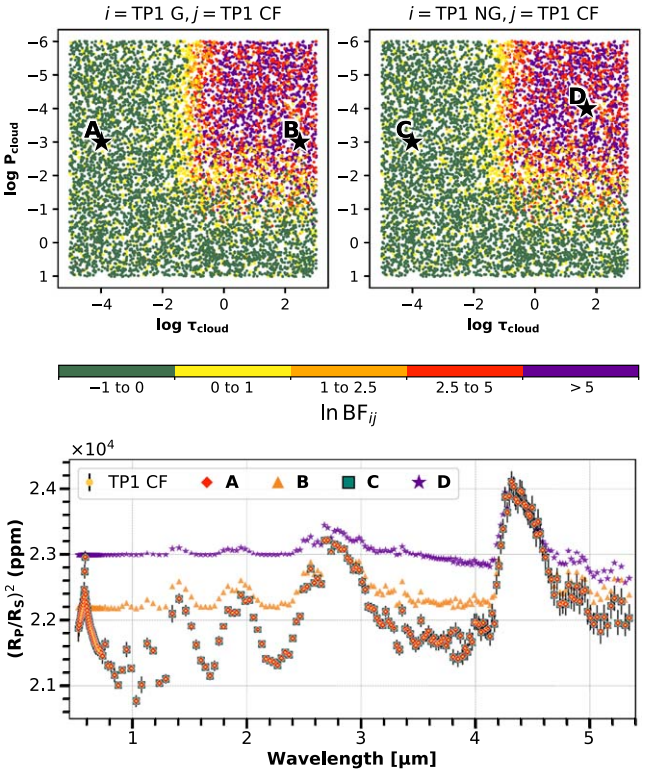


Figure 4. Top panels: Bayes factor (color scale) comparing an isothermal cloudy (left: gray; right: nongray) vs. cloudfree model as a function of the cloud-top pressure (P_{cloud}) and cloud optical depth (τ_{cloud}), for 9000 simulated data realizations. Parameters not shown are randomly sampled from their prior distributions for each realization. Bottom panel: synthetic spectra corresponding to the four realizations labeled in the top panels (and remaining parameters as in Table 1). Also shown is a cloudfree spectrum with measurement errors reported by JWST for WASP-39b as illustration of the measurement uncertainty.

produced by several models, due to either noise fluctuations or parameter degeneracies, nonzero off-diagonal entries are an expected feature of Bayesian model comparison and arise under two different circumstances.

First, for nested models (e.g., CF $\subset$ G $\subset$ NG), when the strength of the "additional" effect (e.g., the optical depth τ_{cloud} of the cloud layer) is negligible in comparison with the noise, Bayesian model comparison includes the principle of Occam's razor and favors the simpler (nested) model, as demonstrated in Figure 4. Here it manifests in the prominent entries below the diagonal[25] of the top panel of Figure 5, which compares all

[24] While this appears similar to the standard confusion matrix used in multiclass classification, it is a different construct in that it shows average posterior probabilities across its rows rather than classification frequencies. In K. Karchev et al. (2023a), following M. H. DeGroot & S. E. Fienberg (1983), it is referred to as the "refinedness" matrix of the probabilistic classifier.

[25] Notice that models in the EMPP matrix are arranged from top to bottom and left to right in order of increasing complexity.

three cloud models on average, and similarly for the temperature–pressure profiles in the bottom panel.

Second, it is also possible that significant posterior probability is assigned to models that are more complicated than the true one: for example, TP3 → TP4 in the bottom panel of Figure 5, when the data are not sufficiently informative to distinguish between them. Consider the extreme case in which the models cannot be distinguished at all: then, the posterior probabilities revert to the prior, which in our case is equal among all models. This can happen either because the data are *uninformative* about the two models (e.g., the noise level is much larger than the differences in the spectra predicted by the two models) or because the network is not optimally trained. In general, it is difficult to distinguish the source of such probability leakage, but given that the SBI results we examined in Figure 3 are very close to the ground truth, we tend to exclude the network's undertraining as the main culprit. Regardless, the EMPP matrix quantifies the model comparison average performance using SBI irrespective of the source of the probability leakage to the upper right off-diagonal elements.

Finally, we remark that the EMPP matrix is a useful tool to evaluate the effectiveness of model comparison. It results from a complex interplay among quality of the data (the observational noise), completeness and accuracy of the models being compared, residual intrinsic dispersion, and prior ranges for the models' parameters (which control the strength of the Occam's razor effect). To our knowledge, no approach other than amortized SBI can deliver this kind of insight with reasonable computational effort.

3.4. Application to JWST Data of WASP-39b

Finally, we test the FASTER framework on real JWST spectra. Previously, A. Lueber et al. (2024) performed nested sampling retrievals on the NIRSpec PRISM spectrum (0.53–5.34 μm) of the hot Jupiter WASP-39b. They determined that the isothermal gray-cloud model (TP1 G in the terminology of the current study) has the highest Bayesian evidence. FASTER reproduces this conclusion with the following ranking of posterior probabilities: TP1 G (26.7%), TP2 G (21.1%), TP4 G (18.1%), TP3 G (17.6%); nongray clouds and cloudfree models have posterior probabilities of at most 4.2% and 2.4%, respectively.

Figure 6 displays the marginal posterior distributions of parameters obtained using both nested sampling and SBI, restricting the analysis to the TP1 G model that is preferred by Bayesian model comparison. There is good agreement between the median values and shapes of the posteriors, but the differences here are more pronounced than in the case of simulated data (Figure 2). Curiously, the residuals between data and model are most pronounced around 3 and 4.4 μm, where the spectral features of water and carbon dioxide are expected to dominate the fluxes.

Moreover, we find—as A. Lueber et al. (2024) did with nested sampling—evidence for additional scatter present in the data at the $\varsigma \approx 100$ ppm level. This implies that the physical and/or instrumental modeling (implemented in the simulator *and* the nested sampling likelihood) may be incomplete and therefore the real data may not be a plausible sample from the model, i.e., it may lie outside the distribution of training examples; therefore, we should not expect a perfect match between SBI and likelihood-based results in this case, as model misspecification has in general different failure modes in likelihood-based inference and SBI. Verifying that the observed data are in distribution with respect to the training set is a key open issue that is the subject of ongoing research in the SBI community (see, e.g., A. Wehenkel et al. 2024).

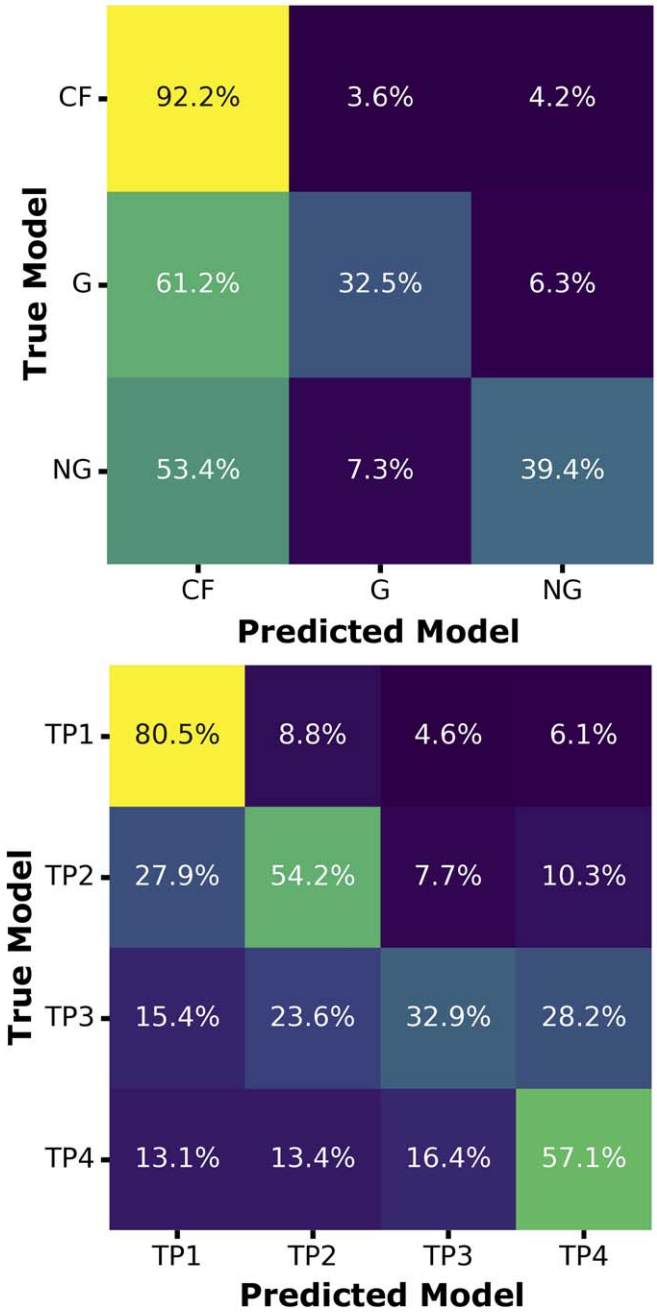


Figure 5. EMPP matrix for cloud models (top panel; cloudfree, gray clouds, nongray clouds from top to bottom) and temperature–pressure profile models (bottom panel, increasing the number of slope changes in the profile from top to bottom). Each row is the average posterior probability distribution (over the models as labeled on the bottom) when evaluated for random examples from the model given on the left.

Despite the above, this exercise demonstrates that our FASTER approach produces essentially equivalent scientific conclusions to what can be obtained via nested sampling.

4. Discussion

4.1. Comparison to Previous Work

A few recent studies in the exoplanet literature use alternative SBI approaches to atmospheric retrieval.

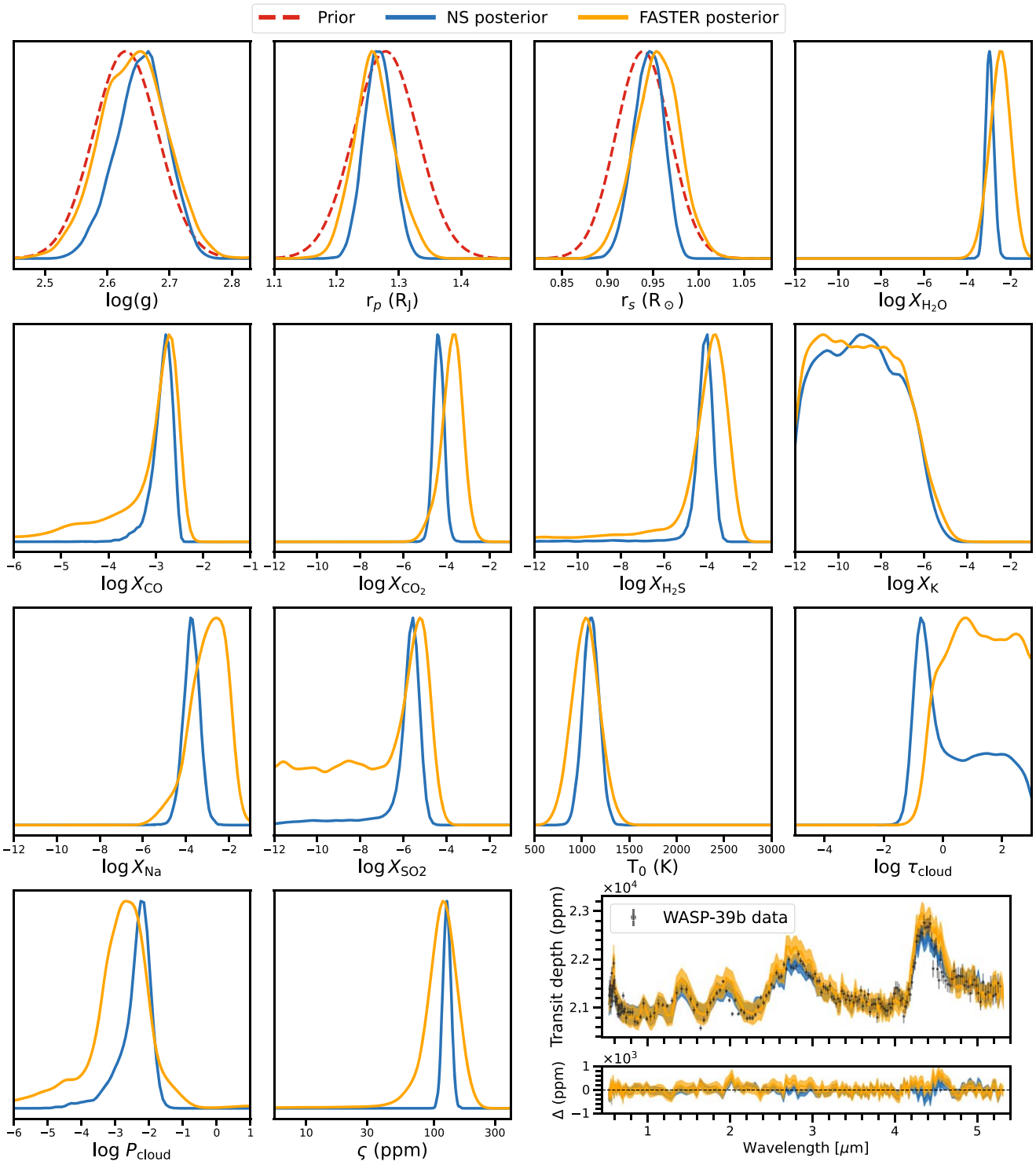


Figure 6. Atmospheric retrieval analysis of the JWST NIRSpec PRISM spectrum of the hot Jupiter WASP-39b using both FASTER and nested sampling for the TP1 G model preferred by Bayesian model comparison. The nested sampling analysis was previously published by A. Lueber et al. (2024) and reproduced here for convenience of comparison. The marginal posterior distributions computed by SBI and nested sampling are shown in orange and blue, respectively, normalized to their peak. Priors (dashed red) are shown only where not uniform across the range. In the spectrum panel (bottom right), only the measurement uncertainty of the data is shown (without the additional intrinsic dispersion, which is, however, accounted for in the fit). The SBI and nested sampling retrievals required ∼1 s (post-training) and ∼8 hr of GPU time, respectively.

M. Vasist et al. (2023) apply neural posterior estimation (NPE) for computationally efficient retrievals, using normalizing flows to estimate the posterior in an amortized way. NPE is also used by M. Aubin et al. (2023), the winner of the ARIEL 2023 data challenge,[26] and T. D. Gebhard et al. (2025), who compare discrete and continuous normalizing flows and refine their results with (likelihood-based) importance sampling for improved posteriors and Bayesian evidence estimates. Lastly, F. Ardévol Martínez et al. (2024) estimate posteriors sequentially, reusing initial approximate results to simulate training data targeted to each observation. While this improves precision, it sacrifices amortization.

[26] https://www.ariel-datachallenge.space/adc2023/

In contrast to these studies, we have employed NRE for parameter inference, which relies on a very simple feed-forward NN instead of a normalizing flow. While we have only estimated one-dimensional marginals—which are usually sufficient for drawing scientific conclusions—ratio estimation can be extended to derive the joint posterior over many ($\gtrsim$20) parameters with essentially the same simulation and training time requirements (N. Anau Montel et al. 2024). Furthermore, the likelihood-to-evidence ratio is independent of the prior (modulo parameter-independent normalization), which allows posterior evaluation under different prior assumptions and fully amortized sequential training, which we will demonstrate in future work. Moreover, we have presented the first simulation-based fully Bayesian and amortized model comparison of models of exoplanetary atmospheres, which bypasses explicit evidence calculations and directly compares (and ranks) numerous alternatives. This technique, while distinct from the simulation-based approaches to parameter estimation discussed above, can fully reuse existing simulations and still relies on a simple classifier network architecture. Finally, we have applied both methods to the real JWST PRISM spectrum of WASP-39b and obtained results in good agreement with traditional nested sampling retrievals.

4.2. Implications and Opportunities for Future Work

While this study validated the SBI approach for a specific exoplanet, with minimal additional training effort it is possible to generalize our networks to be applicable for inference and model selection to a broad range of planetary radii, stellar radii, and planetary surface gravity. On top of one-dimensional marginal posteriors, the same inference strategy can be trivially extended to two-dimensional joint marginal posteriors for all pairs of parameters of interest. Exploiting the amortized nature of the network, posterior distributions can be calibrated to obtain exact frequentist intervals with guaranteed coverage, following the procedure described in K. Karchev et al. (2023b).

With such a tool in hand, it becomes possible to use it for near-instantaneous retrieval over a large ($\gtrsim$100) sample of spectra and to extend it to an ensemble ($\gtrsim$1000) of models. For each exoplanet, one could trivially perform Bayesian model averaging, where the final retrieval becomes a weighted average of every model (according to its respective posterior probability) in the ensemble. Additionally, population-level studies become feasible without compromising on any aspect of the analysis—including selection effects, which are key to obtaining accurate inferences for the population parameters of exoplanets.

The computational streamlining and greatly increased efficiency of the SBI framework will also allow us to finally incorporate the uncertainties associated with the cross sections or opacities of atoms and molecules into atmospheric retrievals. Using traditional methods such as nested sampling, this improvement would be computationally prohibitive, as each retrieval would incur order-of-magnitude increases in computational cost. Using SBI, we would merely require that the empirical uncertainties associated with these cross sections be stochastically sampled in the simulator.

We believe that SBI approaches of the kind presented here will constitute as large a jump forward in spectral analysis of exoplanetary atmospheres as the original retrieval methods based on Markov Chain Monte Carlo have proven to be.

Acknowledgments

R.T. acknowledges cofunding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE1—Project FAIR "Future Artificial Intelligence Research." This resource was cofinanced by the Next Generation EU [DM 1555 del 11.10.22]. R.T. is partially supported by the Fondazione ICSC, Spoke 3 "Astrophysics and Cosmos Observations," Piano Nazionale di Ripresa e Resilienza Project ID CN00000013 "Italian Research Center on High-Performance Computing, Big Data and Quantum Computing" funded by MUR Missione 4 Componente 2 Investimento 1.4: Potenziamento strutture di ricerca e creazione di "campioni nazionali di R&S (M4C2-19)"—Next Generation EU (NGEU). A.L. acknowledges partial financial support from the Swiss National Science Foundation (via grant No. 192022 awarded to K.H.). K.H. and M.H. acknowledge partial financial support from the European Research Council (ERC) Geoastronomy Synergy Grant (grant No. 101166936). C.F. acknowledges financial support from the ERC under the European Union's Horizon 2020 research and innovation program (grant agreement No. 805445).

A.L. and K.K. performed all numerical calculations, created the figures, and led all aspects of the technical implementation. C.F., R.T., and K.H. initiated the model design of the project and provided scientific direction. M.H. assisted in the calculations, as part of his Master project training, and checked for reproducibility. All coauthors participated in a long series of Zoom discussions and contributed to the writing of the manuscript. No generative AI was used in the production of this manuscript.

Appendix
Full EMPP Matrix

For completeness, we provide the full 12 × 12 EMPP matrix in Figure A1. From this matrix, the two marginalized matrices in Figure 5 can be derived by summing over the temperature profile model (thus obtaining the top matrix in Figure 5) or over the cloud type model (thus obtaining the lower matrix). We notice in this matrix the appearance of diagonal substructure in the lower-diagonal block submatrices spanning the same cloud class (e.g., TP1 G to TP4 G submatrix); similarly to what was discussed in the main text, those are instances of Occam's razor preferring a simpler model (e.g., TP2 CF when the true model is TP2 G, as the data are insufficiently constraining).

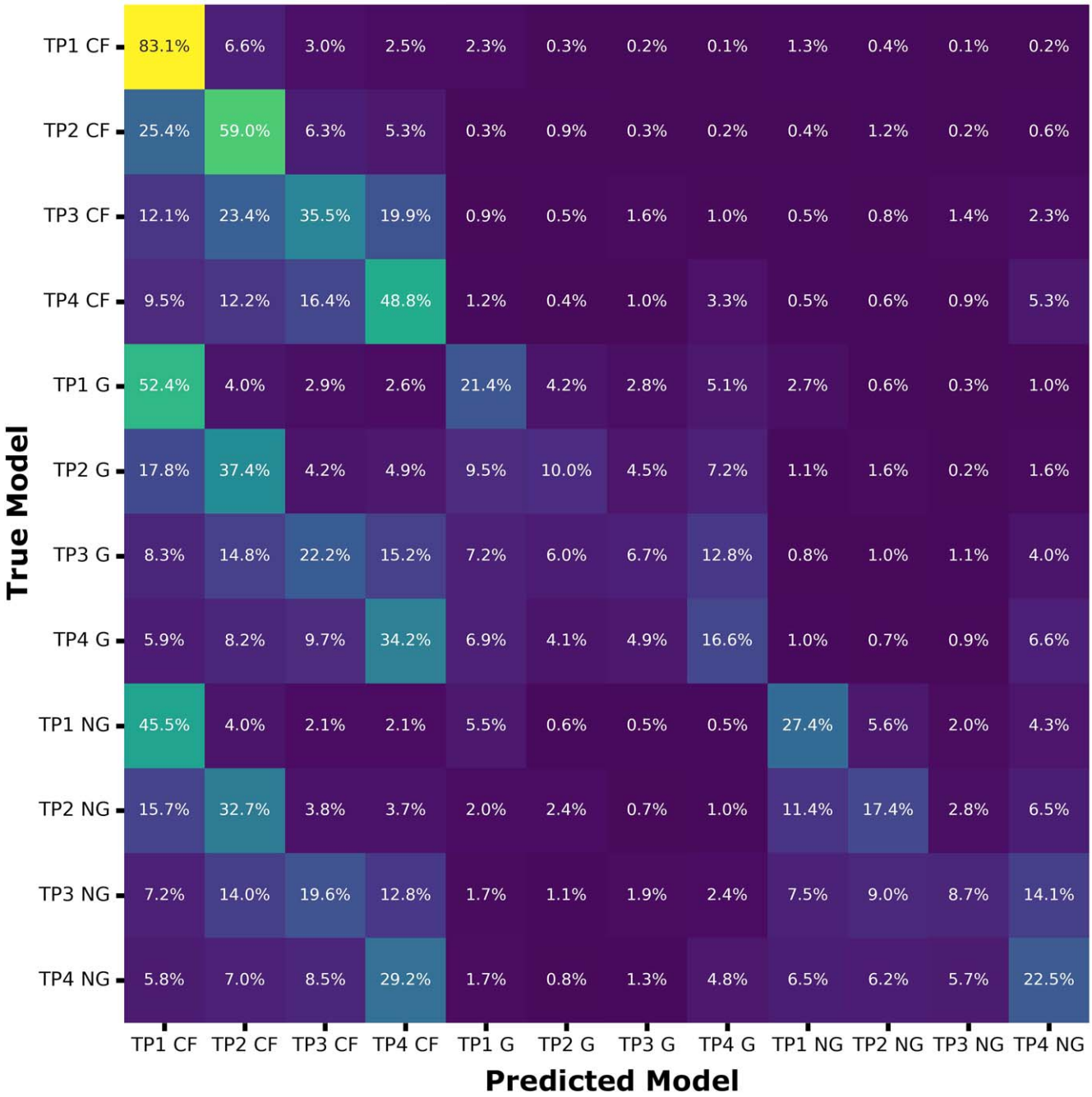


Figure A1. The full EMPP matrix. See Section 3.3 for construction and interpretation.

ORCID iDs

Anna Lueber https://orcid.org/0000-0001-6960-0256
Konstantin Karchev https://orcid.org/0000-0001-9344-736X
Chloe Fisher https://orcid.org/0000-0003-0652-2902
Matthias Heim https://orcid.org/0009-0005-9020-0827
Kevin Heng https://orcid.org/0000-0003-1907-5910

References

Abel, M., Frommhold, L., Li, X., et al. 2011, JPCA, 115, 6805
Abel, M., Frommhold, L., Li, X., et al. 2012, JChPh, 136, 044319
Anau Montel, N., Alvey, J., & Weniger, C. 2024, MNRAS, 530, 4107
Ardévol Martínez, F., Min, M., Huppenkothen, D., et al. 2024, A&A, 681, L14
Ardévol Martínez, F., Min, M., Kamp, I., & Palmer, P. I. 2022, A&A, 662, A108
Aubin, M., Cuesta-Lazaro, C., Tregidga, E., et al. 2023, arXiv:2309.09337
Azzam, A. A. A., Tennyson, J., Yurchenko, S. N., et al. 2016, MNRAS, 460, 4063
Barstow, J. K., & Heng, K. 2020, SSRv, 216, 82
Benneke, B., & Seager, S. 2012, ApJ, 753, 100
Benneke, B., & Seager, S. 2013, ApJ, 778, 153
Brown, T. M. 2001, ApJ, 553, 1006
Carter, A. L., May, E. M., Espinoza, N., et al. 2024, NatAs, 8, 1008
Cobb, A. D., Himes, M. D., Soboczenski, F., et al. 2019, AJ, 158, 33
Cole, A., Miller, B. K., Witte, S. J., et al. 2022, JCAP, 2022, 004
Cook, S. R., Gelman, A., & Rubin, D. B. 2006, Journal of Computational and Graphical Statistics, 15, 675
Cranmer, K., Brehmer, J., & Louppe, G. 2020, PNAS, 117, 30055
Dalmasso, N., Izbicki, R., & Lee, A. B. 2020, arXiv:2002.10399
Dalmasso, N., Masserano, L., Zhao, D., et al. 2021, arXiv:2107.03920

DeGroot, M. H., & Fienberg, S. E. 1983, Journal of the Royal Statistical Society. Series D (The Statistician), 32, 12
Elsemüller, L., Schnuerch, M., Bürkner, P.-C., et al. 2023, arXiv:2301.11873
Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601
Fortney, J. J. 2005, MNRAS, 364, 649
Fu, G., Stevenson, K.B., Sing, D.K., et al. 2025, arXiv:2501.02081
Gebhard, T. D., Wildberger, J., Dax, M., et al. 2025, A&A, 693, A42
Gelman, A., Carlin, J. B., Stern, H. S., et al. 2014, Bayesian Data Analysis (3rd ed.; Boca Raton, FL: CRC Press)
Grimm, S. L., & Heng, K. 2015, ApJ, 808, 182
Grimm, S. L., Malik, M., Kitzmann, D., et al. 2021, ApJS, 253, 30
Hermans, J., Begy, V., & Louppe, G. 2019, arXiv:1903.04057
Hermans, J., Delaunoy, A., Rozet, F., et al. 2021, arXiv:2110.06581
Karchev, K., Trotta, R., & Weniger, C. 2023a, arXiv:2311.15650
Karchev, K., Trotta, R., & Weniger, C. 2023b, MNRAS, 520, 1056
Kirk, J., Ahrer, E.-M., Claringbold, A.B., et al. 2025, MNRAS, 537, 3027
Kitzmann, D., & Heng, K. 2018, MNRAS, 475, 94
Kitzmann, D., Heng, K., Oreshenko, M., et al. 2020, ApJ, 890, 174
Lee, J.-M., Heng, K., & Irwin, P. G. J. 2013, ApJ, 778, 97
Li, G., Gordon, I. E., Rothman, L. S., et al. 2015, ApJS, 216, 15
Line, M. R., Teske, J., Burningham, B., et al. 2015, ApJ, 807, 183
Line, M. R., Wolf, A. S., Zhang, X., et al. 2013, ApJ, 775, 137
Lueber, A., Novais, A., Fisher, C., et al. 2024, A&A, 687, A110
Lueckmann, J.-M., Bassetto, G., Karaletsos, T., et al. 2018, arXiv:1805.09294
Lueckmann, J.-M., Boelts, J., Greenberg, D. S., et al. 2021, arXiv:2101.04653
MacKay, D. J. C. 2003, Information Theory, Inference, and Learning Algorithms (Cambridge: Cambridge Univ. Press)
Madhusudhan, N., & Seager, S. 2009, ApJ, 707, 24
Mancini, L., Esposito, M., Covino, E., et al. 2018, A&A, 613, A41
Márquez-Neila, P., Fisher, C., Sznitman, R., et al. 2018, NatAs, 2, 719
Masserano, L., Dorigo, T., Izbicki, R., et al. 2022, arXiv:2205.15680
Papamakarios, G., & Murray, I. 2016, arXiv:1605.06376
Papamakarios, G., Sterratt, D. C., & Murray, I. 2018, arXiv:1805.07226
Polyansky, O. L., Kyuberis, A. A., Zobov, N. F., et al. 2018, MNRAS, 480, 2597
Skilling, J. 2006, BayAn, 1, 833
Talts, S., Betancourt, M., Simpson, D., et al. 2018, arXiv:1804.06788
Tashkun, S. A., & Perevalov, V. I. 2011, JQSRT, 112, 1403
Trotta, R. 2008, ConPh, 49, 71
Underwood, D. S., Tennyson, J., Yurchenko, S. N., et al. 2016, MNRAS, 459, 3890
Vasist, M., Rozet, F., Absil, O., et al. 2023, A&A, 672, A147
Waldmann, I. P., Tinetti, G., Rocchetto, M., et al. 2015, ApJ, 802, 107
Wehenkel, A., Gamella, J. L., Sener, O., et al. 2024, arXiv:2405.08719
Yip, K. H., Changeat, Q., Al-Refaie, A., et al. 2024, ApJ, 961, 30
Yip, K. H., Changeat, Q., Nikolaou, N., et al. 2021, AJ, 162, 195
Zingales, T., & Waldmann, I. P. 2018, AJ, 156, 268