A&A 672, A147 (2023)
https://doi.org/10.1051/0004-6361/202245263
© The Authors 2023

Astronomy & Astrophysics

Neural posterior estimation for exoplanetary atmospheric retrieval

Malavika Vasist[1,2,⋆], François Rozet[2,⋆], Olivier Absil[1,⋆⋆], Paul Mollière[3], Evert Nasedkin[3], and Gilles Louppe[2]

[1] STAR Institute, University of Liège, 19C Allée du Six-Août, 4000 Liège, Belgium
e-mail: mv.vasist@uliege.be
[2] Montefiore Institute, University of Liège, 10 Allée de la Découverte, 4000 Liège, Belgium
[3] Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany

Received 21 October 2022 / Accepted 10 February 2023

ABSTRACT

Context. Retrieving the physical parameters from spectroscopic observations of exoplanets is key to understanding their atmospheric properties. Exoplanetary atmospheric retrievals are usually based on approximate Bayesian inference and rely on sampling-based approaches to compute parameter posterior distributions. Accurate or repeated retrievals, however, can result in very long computation times due to the sequential nature of sampling-based algorithms.

Aims. We aim to amortize exoplanetary atmospheric retrieval using neural posterior estimation (NPE), a simulation-based inference algorithm based on variational inference and normalizing flows. In this way, we aim (i) to strongly reduce inference time, (ii) to scale inference to complex simulation models with many nuisance parameters or intractable likelihood functions, and (iii) to enable the statistical validation of the inference results.

Methods. We evaluated NPE on a radiative transfer model for exoplanet spectra (`petitRADTRANS`), including the effects of scattering and clouds. We trained a neural autoregressive flow to quickly estimate posteriors and compared against retrievals computed with `MultiNest`.

Results. We find that NPE produces accurate posterior approximations while reducing inference time down to a few seconds. We demonstrate the computational faithfulness of our posterior approximations using inference diagnostics including posterior predictive checks and coverage, taking advantage of the quasi-instantaneous inference time of NPE. Our analysis confirms the reliability of the approximate posteriors produced by NPE.

Conclusions. The inference results produced by NPE appear to be accurate and reliable, establishing this algorithm as a promising approach for atmospheric retrieval. Its main benefits come from the amortization of posterior inference: once trained, inference does not require on-the-fly simulations and can be repeated several times for many observations at a very low computational cost. This enables efficient, scalable, and testable atmospheric retrieval.

Key words. planets and satellites: atmospheres – radiative transfer – methods: numerical

1. Introduction

The characterization of exoplanet atmospheres is concerned with the identification of model parameters that best describe observed exoplanet spectra. More specifically, atmospheric retrieval aims to relate exoplanet spectra to the parameters of detailed forward models of atmospheric physico-chemical processes (Madhusudhan 2018). In this setting, Bayesian inference provides a principled framework to identify parameters that match the observations. The most widely used inference methods for exoplanet retrieval are Markov chain Monte Carlo (MCMC) algorithms (Burningham et al. 2017; Madhusudhan et al. 2011, 2014; Line et al. 2013, 2014; Wakeford et al. 2017; Evans et al. 2017; Blecic 2016; Ballard et al. 2011) and variants of nested sampling (Lavie et al. 2017; Mollière et al. 2020; Todorov et al. 2016; Benneke & Seager 2013; Waldmann et al. 2015a,b; Oreshenko et al. 2017; MacDonald & Madhusudhan 2017; Gandhi & Madhusudhan 2018). Although both families of sampling-based algorithms are asymptotically exact, their sequential nature is often an obstacle to fast, scalable, and testable inference (Cole et al. 2022). First, sampling-based algorithms can take a few hours up to a few days of computation for each single retrieval. Processing just a few observations can quickly add up to several weeks of computing time, which prevents detailed retrievals for large catalogs of observations. With the advent of the *James Webb* Space Telescope (JWST), and of future missions expected to produce a vast number of observations, this becomes largely inapplicable. Second, the necessary computational requirements to maintain accurate results often scale poorly with the number of model parameters. This issue is especially salient for simulation models that include many nuisance parameters, whose posteriors are typically not of direct interest but need to be computed anyway because sampling-based approaches require sampling the full joint posterior. Third, the reliability and statistical rigor of the approximations produced by sampling-based algorithms are difficult to assess. Statistical validation based on repeated inferences, such as simulation-based calibration (Talts et al. 2018) or expected coverage (Hermans et al. 2021), is not feasible in a reasonable time.

[⋆] F.R.S.-FNRS PhD Research Fellow.
[⋆⋆] F.R.S.-FNRS Senior Research Associate.

Nested sampling and MCMC inference algorithms also pose fundamental limits to the class of possible simulation models describing the physics of exoplanet atmospheres. To operate, they require an explicit and tractable expression of the likelihood function, which generally constrains simulation models to forward processes that are mainly deterministic, or involve only a few nuisance parameters. Yet, when data quality will make it possible to account for more details in the cloud physics, more sophisticated simulation models can involve a large number of interfering stochastic processes, resulting in an implicit or intractable likelihood. Possible examples include the cloud formation mechanisms (e.g., via seeding by nucleation, Lee et al. 2018), their growth (e.g., via coagulation or surface growth, Helling & Woitke 2006), their diffusion processes and interactions with the surrounding thermodynamic conditions (e.g., by settling and mixing, Woitke et al. 2020), or their evolution with time (e.g., by ionization). Retrieval with MCMC or nested sampling becomes impossible in these scenarios, at least not without simplifying assumptions.

A possible way to speed up the inference process, and thereby allow the introduction of more complex simulation models in atmospheric retrievals, is to rely on recent advances in the field of machine learning. For instance, training a random forest (Márquez-Neila et al. 2018; Nixon & Madhusudhan 2020), a generative adversarial network (Zingales & Waldmann 2018), an ensemble of Bayesian neural networks (Cobb et al. 2019), or a convolutional neural network (Ardévol Martínez et al. 2022) to retrieve model parameters from noisy data results in quasi-instantaneous retrieval, after paying the upfront cost of generating a training dataset and of training the network. However, this comes at the expense of posterior accuracy, as the resulting parameter distributions are generally not true posteriors in the Bayesian sense, and are sometimes even enforced to follow a multivariate Gaussian distribution. Another approach is to use machine learning to generate more informed, narrower priors (Hayes et al. 2020), or even to replace the exoplanet atmosphere simulator by a surrogate model (Himes et al. 2022). These methods have the potential of providing more accurate posterior distributions, but at the expense of a more modest improvement in terms of inference time.

The rapidly developing field of simulation-based inference is now offering new tools to tackle these challenges (Cranmer et al. 2020). For instance, Yip et al. (2022) perform retrieval using the approach of variational inference with a predefined likelihood, to estimate the posterior distribution of a single spectrum, in an un-amortized fashion. Here, we propose to make use of neural posterior estimation (NPE, Papamakarios & Murray 2016; Lueckmann et al. 2017; Greenberg et al. 2019), an approach based on simulation-based inference that makes use of deep learning to amortize the retrieval procedure and bypass the evaluation of the likelihood function. With NPE, a neural network learns a surrogate posterior as an observation-parameterized conditional probability distribution, from precomputed simulations over the full prior space of interest. In this way, retrievals become fast, scalable, and testable. The rest of the paper is structured as follows. In Sect. 2, we formalize atmospheric retrieval as a Bayesian inference problem, and describe the NPE approach for approximate inference. We also describe the atmospheric radiative transfer model used in this work. Then, in Sect. 3, we describe the setup of our experimental study, present our inference results, compare them against those obtained with nested sampling, and demonstrate their validity using inference diagnostics. In Sects. 4 and 5, we discuss related work and finally conclude our study.

2. Methods

2.1. Simulation-based inference

In all generality, simulators are forward stochastic models or computer programs that generate synthetic observations according to input parameters. Formally, a stochastic model takes a vector of parameters of interest θ as input, samples internally a series of nuisance parameters or latent variables $z \sim p(z|\theta)$ and, finally, produces an observation $x \sim p(x|z, \theta)$ as output, thereby defining an implicit likelihood $p(x|\theta)$. The likelihood is often intractable as it corresponds to

$$p(x|\theta) = \int p(x, z|\theta)\,\mathrm{d}z = \int p(x|z, \theta)p(z|\theta)\,\mathrm{d}z, \tag{1}$$

the integral of the joint likelihood $p(x, z|\theta)$ over the latent space. Even if the likelihood is tractable, which is sometimes the case with physical simulators, the posterior

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta')p(\theta')\,\mathrm{d}\theta'} \tag{2}$$

involves an intractable integral over the parameter space, which leads to challenging Bayesian inference problems for simulators with medium to high-dimensional parameter spaces.

These computational obstacles can be bypassed using modern simulation-based inference algorithms. Instead of relying on the likelihood function to perform inference, simulation-based approaches use deep neural networks to parameterize universal density estimators and estimate the posterior. Among the simulation-based inference algorithms, neural posterior estimation consists in training a conditional normalizing flow $p_\phi(\theta|x)$ with parameters ϕ to approximate the posterior distribution $p(\theta|x)$. A normalizing flow (Papamakarios et al. 2021, see Appendix A for further details) is a composition of invertible and differentiable transformations applied to a simple distribution (e.g., a normal distribution), thereby defining a complex distribution that can be efficiently evaluated and sampled from. The transformations are parametrized by invertible neural networks, making normalizing flows universally expressive parametric distributions that can be trained to approximate other distributions. In our case, training is based on amortized variational inference and amounts to the minimization of the expected Kullback-Leibler (KL) divergence between $p(\theta|x)$ and $p_\phi(\theta|x)$, (Agakov 2004), that is

$$\begin{aligned}\phi^* &= \arg\min_\phi \mathbb{E}_{p(x)}\left[\mathrm{KL}(p(\theta|x) \parallel p_\phi(\theta|x))\right] \\ &= \arg\min_\phi \mathbb{E}_{p(x)}\,\mathbb{E}_{p(\theta|x)}\left[\log \frac{p(\theta|x)}{p_\phi(\theta|x)}\right] \\ &= \arg\min_\phi \mathbb{E}_{p(\theta,x)}[-\log p_\phi(\theta|x)].\end{aligned} \tag{3}$$

Remarkably, the amortization over $p(x)$ of the variational inference objective makes it possible to bypass the sampling or the evaluation of the unknown posterior $p(\theta|x)$ in the second line above. Indeed, the double expectation $\mathbb{E}_{p(x)}\mathbb{E}_{p(\theta|x)}$ can be rewritten as an expectation $\mathbb{E}_{p(\theta,x)}$ over the joint distribution, which we can easily sample in the forward direction as $p(\theta, x) = p(\theta)p(x|\theta)$, regardless of whether the likelihood is tractable or not. Once the normalizing flow is trained, evaluating and sampling the posterior density $p_\phi(\theta|x)$ becomes as fast as a forward pass through the network. Inference can be repeated any number of times with different observations, without having to regenerate data from the simulation model.

2.2. Atmospheric radiative transfer model

The atmospheric model used in this study consists of a deterministic atmospheric forward model implemented with petitRADTRANS, together with a noise model accounting for measurement noise. petitRADTRANS (Mollière et al. 2019) is a radiative transfer model used to generate emission and transmission spectra for exoplanets with cloudy and clear atmospheres including scattering, as described in Mollière et al. (2020). It includes a parameterized temperature structure and cloud properties. We used the disequilibrium chemistry emission model predefined in petitRADTRANS to compute an emission spectrum based on disequilibrium carbon chemistry, equilibrium clouds, and a spline temperature-pressure profile, defined by 16 parameters in total. We walk through these parameterizations briefly in the following paragraphs.

The temperature structure uses both freely variable and physically motivated parameterizations based on atmospheric altitudes. The optical depth defined as $\tau = \delta P^{\alpha}$ is parameterized as a function of the pressure P while keeping δ and α as model parameters. The temperature structure is split into three parts. The mid altitude (photosphere), with an optical depth $\tau > 0.1$, models the temperature according to the Eddington approximation (Eq. (2) of Mollière et al. 2020) with T_{int} as a model parameter. In the upper altitude with an optical depth $\tau < 0.1$, the structure is computed by a cubic spline interpolation between T_1, T_2, and T_3 considered as model parameters. In low altitudes (troposphere), wherever the atmospheric temperature gradient of the Eddington approximation is greater than the moist adiabatic gradient (i.e, $\nabla_{\text{edd}} > \nabla_{\text{ad}}$), convection ensues. The ∇_{ad} is interpolated from a T-P-[Fe/H]-C/O space of a chemical equilibrium table. Here the metallicity [Fe/H], and the carbon-to-oxygen number ratio C/O, are also model parameters.

Once the $P - T$ profile is constructed, equilibrium cloud abundances are calculated in the form of their mass fractions, where they are modified from solar abundances based on the model parameters [Fe/H] and C/O. The cloud mass fractions are further scaled with the scaling parameters $\log \tilde{X}_{\text{Fe}}$ and $\log \tilde{X}_{\text{MgSiO}_3}$, where $\tilde{X}_i = X_0^i / X_{\text{eq}}^i$ is the ratio of the cloud mass fraction X_0^i at the cloud base (i.e., at pressure P_{base}) to the mass fraction X_{eq}^i predicted at the same location for the cloud species when assuming equilibrium condensation. The cloud mass fraction decays with altitude based on the settling parameter f_{sed}:

$$X(P) = X_0 \left(\frac{P}{P_{\text{base}}} \right)^{f_{\text{sed}}}. \tag{4}$$

For $P > P_{\text{base}}$ the cloud mass fraction is zero. The other cloud parameters include the vertical eddy diffusion coefficient K_{zz} and the width of the log normal size distribution σ_g defined in the Ackerman & Marley (2001) cloud model, called Cloud Model 1 in Mollière et al. (2020). The chemical abundances for species H_2O, CO, CH_4, NH_3, CO_2, H_2S, VO, TiO, PH_3, Na, and K are interpolated from the chemical equilibrium table calculated with easyCHEM (Mollière et al. 2017) as a function of T-P-[Fe/H]-C/O. The model parameter $\log P_{\text{quench}}$ is used to account for disequilibrium chemistry through atmospheric mixing. For pressures below P_{quench}, the mass fractions of CH_4, H_2O, and CO are held constant at their values at $P = P_{\text{quench}}$. The gas opacities required for the radiative transfer solution are obtained by combining the correlated k (opacity) tables of individual atmospheric absorbers in the resort-rebin fashion (e.g., Mollière et al. 2015; Amundsen et al. 2017). The surface gravity ($\log g$) and radius (R_{p}) of the planet are considered as model parameters to calculate the emission flux. The radiative transfer equations are then solved using the Feautrier method (Feautrier 1964) as in the self-consistent petitCODE (Mollière et al. 2015, 2017), which also includes isotropic scattering. Following Mollière et al. (2020), we rebinned down the default wavelength spacing $\lambda/\Delta\lambda = 1000$ to a spacing of 400 between 0.95 and 2.45 µm. This was done by generating the binned correlated-k opacities in petitRADTRANS, and using them instead of the original opacities to generate linearly binned spectra within the same wavelength range, resulting in vectors of 379 elements. We denote the output spectrum produced by this first simulation stage as $f(\theta)$, where θ contains all 16 model parameters.

To account for measurement noise and make the simulation model similar to instrumental data, we considered a Gaussian noise model with a standard deviation σ. The spectra $f(\theta)$ generated by petitRADTRANS were randomly perturbed with additive noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, where $\epsilon \in \mathbb{R}^{379}$ is a vector of random noise instances in each wavelength bin. Here we assumed the same noise variance in each wavelength bin for the sake of simplicity, but more complex noise models (including noise covariance) could be used in our simulator. The final simulator output is given by $x = f(\theta) + \epsilon$.

3. Atmospheric retrieval

We start our empirical evaluation of NPE-based atmospheric retrieval by describing the creation of the training data in Sect. 3.1, together with the training protocol and a description of the architecture of the neural posterior estimator $p_\phi(\theta|x)$. We demonstrate and discuss in Sect. 3.2 an example of atmospheric retrieval using NPE, and then validate the statistical quality of the posterior estimation in Sect. 3.3. Finally, in Sect. 3.4, we report and compare computational times against the MultiNest algorithm for nested sampling. The inference pipeline is summarized in Fig. 1.

3.1. Setup

As a starting point to atmospheric retrieval with neural posterior estimation, we first defined in Table 1 a 16-dimensional multivariate uniform prior distribution, with physically motivated ranges for each parameter θ. This prior distribution is the same as the one used by Mollière et al. (2020). Our training data set is composed of 12 million parameters-spectrum pairs $(\theta, f(\theta))$. It is created by drawing parameters $\theta \sim p(\theta)$ from the prior and passing them through the simulator as shown in Fig. 1. We split this dataset into 90 %, 9 %, and 1 % for training, validation, and testing respectively.

We implemented the posterior estimator $p_\phi(\theta|x)$ as a neural autoregressive flow (Huang et al. 2018) composed of three transformations. Each transformation was parameterized by a multilayer perceptron (MLP) with five hidden layers of size 512 and ELU activation functions (Clevert et al. 2015). A second network, called the embedding, was used to compress the 379-dimensional spectrum x into a vector of 64 features, which was then used to condition the flow with respect to x. The rationale behind this compression is that it forces the posterior estimator to extract informative features from the spectra instead of memorizing the training data. The embedding network was implemented as a ResidualMLP (or ResMLP), composed of 10 residual blocks (He et al. 2016) of decreasing size (two times 512, three times 256 and five times 128) and also uses ELU activation functions.

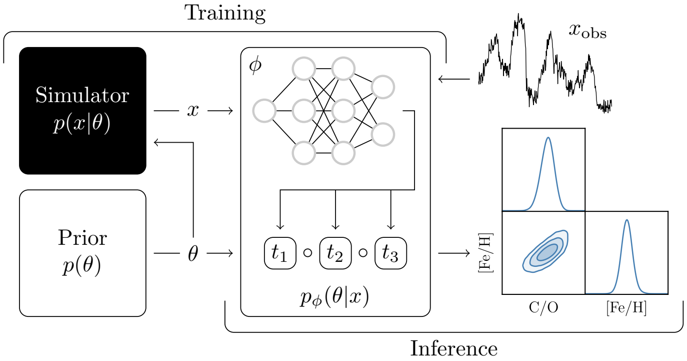

Fig. 1. Inference pipeline using amortized neural posterior estimation. The joint simulation model $p(x, \theta) = p(\theta)p(x|\theta)$ is used to generate a training set $\{(\theta, x)\}$ of model parameters θ and exoplanet spectra observations x. A conditional normalizing flow $p_\phi(\theta|x)$ composed of an embedding network and three invertible transformations t_i is trained to estimate the posterior density $p(\theta|x)$. Once trained, sampling from the posterior estimator is as fast as a forward pass through the normalizing flow. Inference can be repeated for many observations without having to regenerate data nor retrain the normalizing flow.

Table 1. Prior distribution over the model parameters.

Parameter	Prior	Parameter	Prior
T_1	$\mathcal{U}(0, T_2)$	$\log \tilde{X}_{\text{Fe}}$ [(b)]	$\mathcal{U}(-2.3, 1)$
T_2	$\mathcal{U}(0, T_3)$	$\log \tilde{X}_{\text{MgSiO}_3}$ [(b)]	$\mathcal{U}(-2.3, 1)$
T_3	$\mathcal{U}(0, T_{\text{connect}})$ [(a)]	f_{sed}	$\mathcal{U}(0, 10)$
$\log \delta$	$P_{\text{phot}} \sim \mathcal{U}(10^{-3}, 10^2)$ [(c)]	$\log K_{zz}$	$\mathcal{U}(5, 13)$
α	$\mathcal{U}(1, 2)$	σ_g	$\mathcal{U}(1.05, 3)$
T_0	$\mathcal{U}(300, 2300)$ K	R_P	$\mathcal{U}(0.9, 2)$
C/O	$\mathcal{U}(0.1, 1.6)$	$\log g$	$\mathcal{U}(2, 5.5)$
Fe/H	$\mathcal{U}(-1.5, 1.5)$	$\log P_{\text{quench}}$	$\mathcal{U}(-6, 3)$

Notes. [(a)] T_{connect} is the uppermost temperature of the photospheric layer that `petitRADTRANS` calculates by setting the optical depth $\tau = 0.1$. [(b)] Here $\tilde{X}_i = X_0^i / X_{\text{eq}}^i$, where X_{eq} is defined as the mass fraction predicted for the cloud species when assuming equilibrium condensation at the cloud base location. [(c)] P_{phot} is defined as the pressure where the optical depth $\tau = 1$. The parameter δ is calculated accordingly.

Before training, random noise realizations were added on-the-fly to the spectra to obtain observations $x = f(\theta) + \epsilon$. Following Eq. (3), the flow and embedding networks were trained jointly to minimize the expected negative posterior log-density over the training set. The optimization was carried out through a variant of stochastic gradient descent, namely AdamW (Loshchilov & Hutter 2017). We used an initial learning rate of 10^{-3} that was halved every time the average loss over the validation set did not improve for the last 32 epochs, until it reached 10^{-6} to improve training without overfitting (Zhang et al. 2021). We also used a high weight decay of 10^{-2}. We trained for a total of 1024 epochs during which 1024 random batches of 2048 pairs $(\theta, f(\theta))$ were taken from the training set.

The architectural hyper-parameters were adjusted on validation data. We also explored a neural spline flow Durkan et al. (2020) implementation of the posterior estimator, but in the end implemented a neural autoregressive flow since it gave a lower validation loss. We performed extensive hyper-parameter tuning on the flow and embedding network parameters. For the flow, we explored different numbers of transforms and hidden layer dimensions in the range of $[3, 5]$ and $[256, 512]$, respectively. For the embedding network, we tried different number of layers in the ResMLP in the range of $[10, 15]$. We also explored different activation functions like ReLU and ELU for both networks. We explored different values for the initial learning rate and the minimum learning rate in the ranges of $[10^{-5}, 10^{-3}]$ and $[10^{-6}, 10^{-5}]$, respectively. We analyzed the impact of different schedulers like `ReduceLROnPlateau` and `CosineAnnealingLR`, available in PyTorch, with patience rates between $[8, 32]$. We tried batch sizes between $[2^8, 2^{11}]$ and the number of epochs between $[2^8, 2^{10}]$. We tuned each hyper-parameter by randomly searching over a grid within their range mentioned above, and studied their impact over ~80 runs in parallel. We selected those that led to lower validation loss and/or more stable training. Amongst all the parameters that we tuned, the parameter weight decay between $[0, 10^{-2}]$ had the most significant impact on the training. We think this is because of the high variance of the input dataset, where some spectra are six orders of magnitude brighter than the rest. This leads to the skewing of the weights to very high values, which is compensated by weight decay to improve training performance. For more details, we refer to the source code of the experiments[1].

Table 2. Parameter values θ_{obs} of the benchmark spectrum x_{obs}.

Parameter	Value	Parameter	Value
T_1	330.6 K	$\log \tilde{X}_{\text{Fe}}$	−0.86
T_2	484.7 K	$\log \tilde{X}_{\text{MgSiO}_3}$	−0.65
T_3	687.6 K	f_{sed}	3
$\log \delta$	−7.51	$\log K_{zz}$	8.5
α	1.39	σ_g	2
T_0	1063.6 K	R_P	1
C/O	0.55	$\log g$	3.75
Fe/H	0	$\log P_{\text{quench}}$	−5

3.2. Benchmark retrieval

As a demonstration of atmospheric retrieval with NPE, we present inference results for a synthetic exoplanet spectrum x_{obs} generated with the parameter values θ_{obs} given in Table 2, similarly to the benchmark retrieval of Mollière et al. (2020). The synthetic spectrum spans a wavelength range from 0.95 to 2.45 µm with a continuous wavelength spacing of $\lambda/\Delta\lambda = 400$. As in Mollière et al. (2020), we assumed a signal-to-noise ratio of 10 per spectral bin, leading to a standard deviation $\sigma = 0.1257 \times 10^{-17}\,\text{W}\,\text{m}^{-2}\,\mu\text{m}^{-1}$ for the Gaussian noise. The synthetic spectrum used for our retrieval tests is shown in Fig. 1.

Retrieval results are summarized in Fig. 2. The corner plot shows 1d and 2d marginal posterior distributions obtained for the benchmark spectrum x_{obs}. The marginal posterior distributions were approximated by sampling sufficiently many times the joint posterior distribution from the normalizing flow, which takes only a few seconds to obtain a smooth corner plot. We observe that the bulk of the marginal posterior distributions (in blue) is centered around the parameter values θ_{obs} (in black) used to generate the spectrum x_{obs}. The figure also illustrates the spread in the posterior P–T profiles with respect to the synthetic observation spectrum. More specifically, we computed posterior $P - T$ profiles for $\theta \sim p_\phi(\theta|x_{\text{obs}})$, and show their 68.3 %, 95.5 % and 99.7 % credible regions. We see that the $P - T$ profile for θ_{obs} (in black) is constrained mostly within the first credible region of the

[1] `https://github.com/MalAstronomy/sbi-ear`

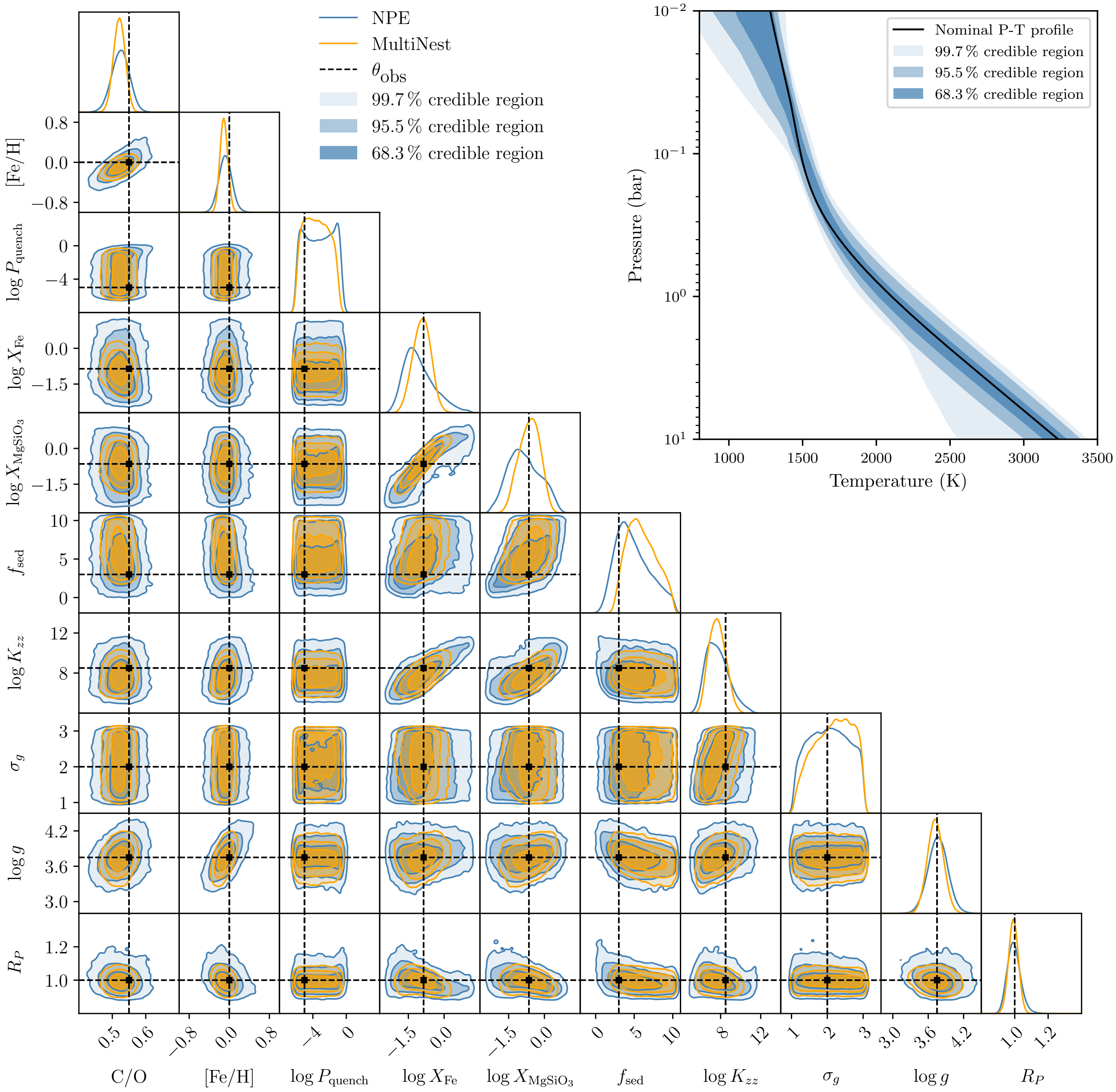

Fig. 2. Benchmark retrieval using neural posterior estimation. The corner plot shows 1d and 2d marginal posterior distributions obtained for the benchmark spectrum $x_{\rm obs}$ for NPE (in blue) and for nested sampling (in orange). We observe that the nominal parameter values $\theta_{\rm obs}$ (in black) are well identified. The top right figure illustrates the posterior distribution of the P–T profiles.

posterior. These results lead us to believe that the NPE posterior approximation produces coherent posterior distributions.

3.3. Validation

In Fig. 2, we compare the NPE posteriors with those obtained using MultiNest (Feroz & Hobson 2008; Feroz et al. 2009, 2019; Buchner et al. 2014) for the same noisy synthetic observation, in orange. While the results obtained with NPE appear to be coherent with respect to the nominal parameter values $\theta_{\rm obs}$ and the posterior $P-T$ profiles, we see that the approximate marginal posterior distributions computed with MultiNest, using a sampling efficiency of 0.8 (recommended for parameter estimation) with 4000 live points, are slightly less dispersed than for NPE. On performing several retrievals with different noise realizations (not shown here), it is seen that, each time, the peaks of the individual parameter posterior distributions shift in a similar way in both retrieval algorithms. This can be seen here in the parameters C/O and Fe/H, similarly shifted slightly to the left. This suggests that these shifts are directly related to the particular noise realization, and that MultiNest and NPE behave in a similar way in presence of noise.

The difference in the posterior widths for the two algorithms motivates a thorough investigation of the computational faithfulness of the NPE posterior approximations using inference diagnostics, including posterior predictive checks and

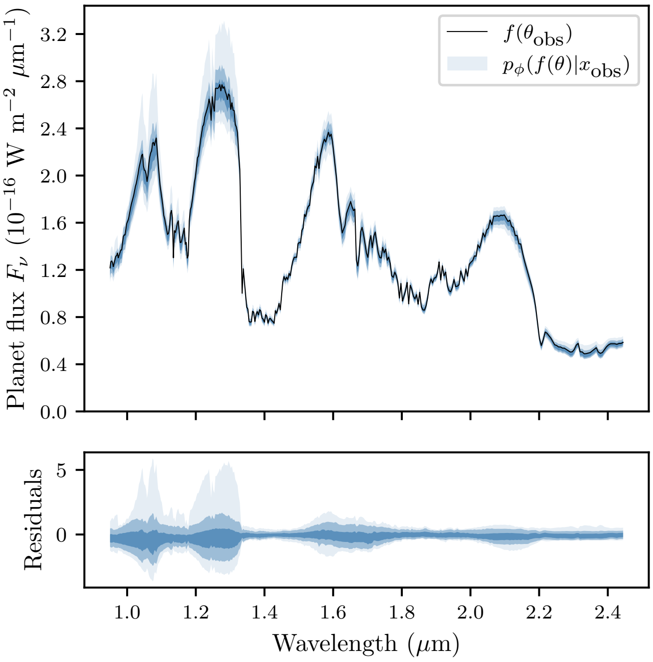

Fig. 3. *Top:* posterior predictive distribution $p(f(\theta)|x_{\text{obs}})$ of noiseless spectra (without the instrumental noise disturbance ϵ) for the 99.7%, 95% and 68.7% quartiles (hues of blue), overlaid on top of the noiseless observed spectrum $f(\theta_{\text{obs}})$, (black line). *Bottom:* residuals of the posterior predictive samples, normalized by the standard deviation of the noise distribution for each spectral channel.

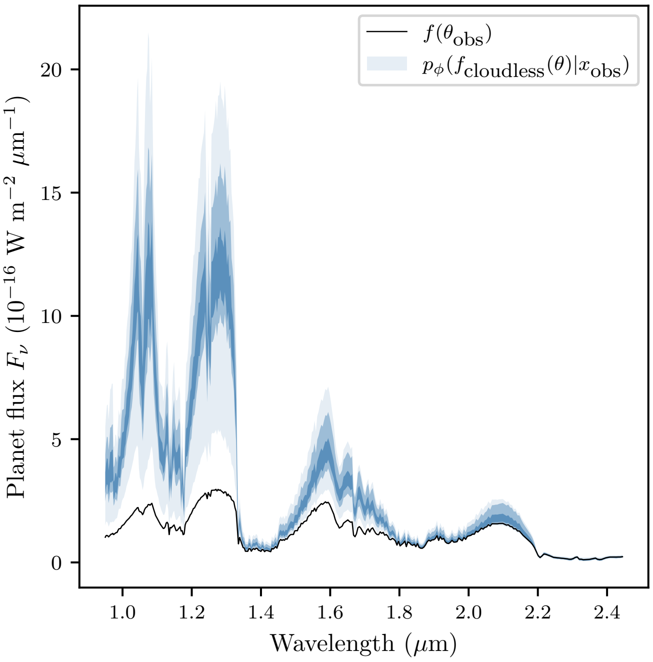

Fig. 4. Cloudless realizations of the posterior predictive distribution $p(f_{\text{cloudless}}(\theta)|x_{\text{obs}})$ overlaid on top of $f(\theta_{\text{obs}})$, where $f_{\text{cloudless}}$ artificially sets the cloud scaling factors $\log X_{\text{Fe}}$ and $\log X_{\text{MgSiO}_3}$ to a very small value of -10.

coverage. We took advantage of the quasi-instantaneous inference of NPE to perform these checks. We first performed a quantitative examination of the posterior predictive distribution $p_\phi(f(\theta)|x_{\text{obs}})$ for spectra without instrumental noise disturbance, which we obtained by sampling parameters from the posterior, $\theta \sim p_\phi(\theta|x_{\text{obs}})$, and then computed their spectra $f(\theta)$ with `petitRADTRANS`. Figure 3 shows the posterior predictive distribution $p(f(\theta)|x_{\text{obs}})$ for various quartiles against the noiseless version of the observed spectrum $f(\theta_{\text{obs}})$. We observe that (i) the posterior predictive distribution is well constrained, with the 68% quartile distribution mostly within the 1σ noise limit as expected, and (ii) that $f(\theta_{\text{obs}})$ is relatively well centered inside the 68% quartile along all bins. Had the posterior distribution $p_\phi(\theta|x_{\text{obs}})$ been too wide, we would have observed a much larger spread. Had the bulk of the posterior density been at the wrong place, we would not have observed $f(\theta_{\text{obs}})$ to be well inside the distribution. These reassuring diagnostics are a first indication of the good quality of the inference results obtained with NPE. In particular, they demonstrate that the cloud parameter distributions derived by NPE produce spectra consistent with the observed spectrum. In Fig. 4, we further demonstrate that the parameter values sampled from the somewhat wider NPE cloud posteriors are actually all leading to cloudy solutions, in good agreement with the synthetic input observation. In this figure, we sampled parameters from the (cloudy) approximate posterior, but then artificially turned off the clouds, by setting the log of the cloud mass fraction scaling factors X_{Fe} and X_{MgSiO_3} to -10, to assess their impact on the spectral shape. We see that these cloudless spectra look significantly different from the cloudy ones shown in Fig. 3. This implies that the posterior predictive distribution samples in Fig. 3 are indeed affected by clouds.

Following Hermans et al. (2020), we further evaluate the global computational faithfulness of the NPE posterior approximations in terms of expected coverage. We define the expected coverage probability of the $1-\alpha$ highest posterior density regions derived from the posterior $p_\phi(\theta|x)$ as

$$\mathbb{E}_{p(\theta,x)}\left[\mathbb{1}\left(\theta \in \Theta_{p_\phi(\theta|x)}(1-\alpha)\right)\right], \tag{5}$$

where $\mathbb{1}(\cdot)$ is the indicator function, and where the function $\Theta_{p_\phi(\theta|x)}(1-\alpha)$ yields the $1-\alpha$ highest posterior density region of $p_\phi(\theta|x)$. This diagnostic probes the consistency of the posterior estimator $p_\phi(\theta|x)$ and can be used to assess whether the approximate posterior distributions are overdispersed or underdispersed on average. It is estimated by repeatedly sampling (θ, x) from the prior and the simulation models, and then running NPE retrievals on each x. If the posteriors are well calibrated, then the parameter values θ that were used to generate the spectra x should be contained in the $1-\alpha$ highest posterior density regions of the approximate posteriors $p_\phi(\theta|x)$ exactly $(1-\alpha)$% of the time. If the coverage probability is smaller than the credibility level $1-\alpha$, then this indicates that the $1-\alpha$ highest posterior density regions are smaller than they should be, which is the sign of overconfident and usually unreliable posterior approximations. On the other hand, if the coverage probability is larger than the credibility level $1-\alpha$, then this indicates that the highest posterior density regions are wider than they should be. In this case, the posterior approximations are said to be conservative. We argue that posterior approximations should rather be conservative to guarantee reliable and meaningful inferences, even when the approximations are not faithful. Indeed, wrongly excluding plausible parameter values of exoplanet spectra could lead to wrong conclusions about the actual nature of the exoplanet, while failing to exclude actually implausible parameter values would only result in a loss of statistical power. Figure 5 summarizes the expected coverage of $p_\phi(\theta|x)$ for credibility levels from 0 to 1.

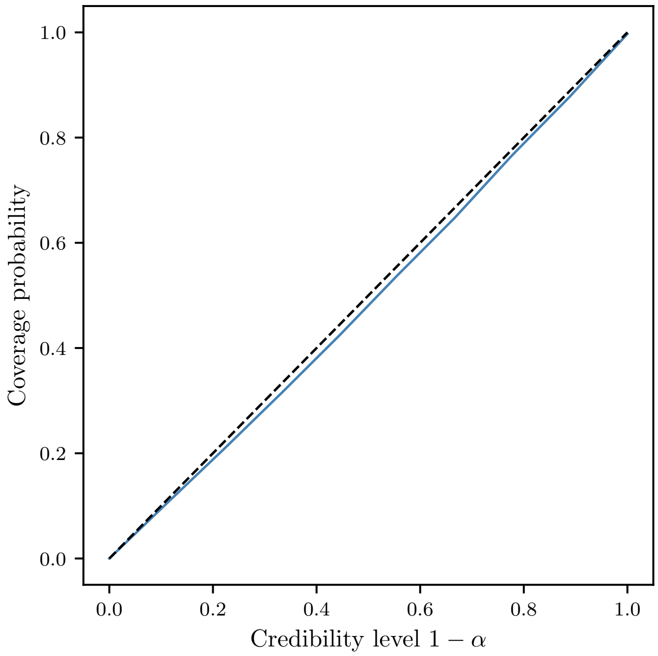

Fig. 5. Coverage plot assessing the computational faithfulness of $p_\phi(\theta|x)$ in terms of expected coverage. The coverage probability is close to the credibility level $1 - \alpha$, which indicates that the posterior approximations produced by NPE are neither significantly overdispersed (the coverage curve would otherwise be above the diagonal) nor significantly underdispersed (the coverage curve would be below the diagonal).

The coverage curve closely fits the diagonal, which indicates that the posterior distributions produced by NPE are well calibrated – even though we note a trend for the posteriors to be very slightly underdispersed.

Unfortunately, running the same coverage diagnostic for a sampling algorithm such as MCMC or nested sampling is not possible within a reasonable computation time, since it requires the repeated construction of posterior distributions over many distinct random realizations x in order to approximate the expectation in Eq. (5). For this reason, we cannot conclude whether `MultiNest` is computationally faithful in terms of expected coverage. However, given that the approximate posterior distribution produced by `MultiNest` in Fig. 2 is slightly narrower than the NPE posterior distribution, it suggests that `MultiNest` is slightly more underdispersed than NPE. This conclusion is in line with the analysis of `MultiNest` posterior distributions in Ardévol Martínez et al. (2022), where 4000 retrievals were performed on simulated observations using CNNs and MultiNest, which on comparison suggest that MultiNest tends to underestimate the uncertainties of the parameter it retrieves.

3.4. Computational cost

One of the main advantages of neural posterior estimation is its amortization of the inference procedure. Once trained, inference does not require on-the-fly simulations and can be repeated several times with different observations at very low computational cost. We demonstrate the true potential of NPE by performing 1000 retrievals and comparing how long it would take for `MultiNest` to produce comparable results. The 1000 observations were produced by randomly sampling parameters values θ from the prior distribution and rendering them through the forward simulation model to produce $x \sim p(x|\theta)$. We then retrieved their corresponding approximate posterior distributions $p_\phi(\theta|x)$. A single retrieval consists of sampling 30 740 posterior parameter vectors (as many as `MultiNest` yields) and rendering the corner plot, took respectively 6 and 10 s in average. In total, 1000 retrievals would take approximately 4.5 h. With the upfront generation of the dataset (17 h on 1000 CPUs) and the training of the neural network (13 h on a standard NVIDIA GTX 1080 Ti GPU), we reach a total computing time of 34.5 h. By contrast, each retrieval with `MultiNest` takes around 134 h on a cluster of 440 CPUs (totaling about 60 000 CPU hours) so that retrieving atmospheric parameters on 1000 spectra would require an extrapolated time of 134 000 h (15 yr). In summary, NPE is around 4000 times faster for a thousand retrievals, and almost 30 000 times faster if we do not take the upfront generation and training into account.

It is important to note that the computational speedup comes with the overhead cost of building the training set (one per atmospheric model) and training NPE on it. In our case, simulating a single parameter-spectrum pair $(\theta, f(\theta))$ took around 5 s, which results in a total of 17 000 CPU hours for the generation of the 12 millions pairs used in this study. The actual wall-clock time, however, can be largely reduced by simulating the pairs in parallel on a large computing cluster, contrary to the on-the-fly and sequential simulations required in MCMC or nested sampling methods. In our case, the training set was generated in less than 17 h using a cluster of 1000 CPUs. Generating as many samples may not be necessary in all cases, since sufficiently good performance is likely to be possible from smaller training sets. Instead, the main challenge with amortized inference will be to identify a simulation model that is general enough to be applicable and valid in many situations, so that the whole training process does not need to be repeated for each individual case. This may be possible for studies focusing on specific classes of planets, such as hot Jupiters observed in transit, or self-luminous giant planets observed with direct imaging. We also note that, when performing retrievals on a single object, a given training data set can potentially be reused several times when exploring various levels of wavelength binning or different noise models in the retrieval. In this case, only the cost of the NPE training needs to be paid several times.

4. Related work

The closest work to our study is the recent work of Yip et al. (2022), who investigated variational inference and normalizing flows for atmospheric exoplanet retrieval. In contrast to NPE, their approach was nonamortized, and variational inference was targeted at a single spectrum. For this reason, the expected KL divergence trick we used in Eq. (3) to bypass the sampling of the unknown posterior $p(\theta|x)$ is no longer applicable. Instead, the parameters of the normalizing flow were trained by maximizing a variational lower bound on the evidence $p(x)$, provided that the likelihood function associated with the simulation model is both tractable and differentiable. In NPE, none of these requirements are necessary – the inference algorithm can be applied to any kind of simulation model, tractable or not, differentiable or not. Nevertheless, the amortization in NPE comes at the price of the upfront simulation of a large training set, whereas direct variational inference as in Yip et al. (2022) only requires a limited number of on-the-fly simulations. These on-the-fly simulations, however, make inference significantly slower than the quasi-instantaneous inference produced by an already-trained normalizing flow. In particular, this prevents posterior diagnostics (as in Sect. 3.3), as they become computationally very expensive.

Beyond exoplanet retrieval, NPE has been used increasingly for inference problems found in astronomy. Close to exoplanet atmospheric retrieval, Baso et al. (2022) and Ramos et al. (2022) used NPE to determine the thermodynamic and magnetic properties of solar and stellar atmospheres as well as their high-dimensional temperature maps. They similarly advocate for amortized and rapid parameter estimation if complex models are used to analyze the large amounts of data that the next generation of telescopes will produce. In gravitational wave science, Dax et al. (2021) used NPE for fast and accurate inference of the properties of binary black holes from gravitational waves. The inference time was reduced from 1 day using MCMC to 20s using NPE, making a strong case for inference in real-time. Similarly, Zhang et al. (2020) and Hahn & Melchior (2022) used NPE to perform inference on binary microlensing events. Complex high-dimensional physical models result in time-consuming forward simulations and complex likelihood surfaces that MCMC methods find challenging to sample from. NPE offers a way to infer from an upcoming catalog of binary events more accurately and in real-time. In astroparticle physics, Mishra-Sharma & Cranmer (2022) used NPE to improve the characterization of the sources that contribute to the Fermi γ-ray Galactic Center Excess (GCE), by directly sampling from high-dimensional γ-ray maps instead of defining a simplified and tractable likelihood function that loses some information. Likewise, Bister et al. (2022) studied the inference of cosmic-ray source properties from cosmic-ray observations on Earth. They concluded that inference with NPE provides accurate, fast, and verifiable results for a large phase space of the source parameters. Finally, as a last example, Kodi Ramanah et al. (2020) used NPE to characterize the dynamical mass of galaxy clusters directly from their 2d phase-space distributions.

For the same reasons of efficiency, scalability, and testability, simulation-based inference algorithms beyond NPE are being increasingly used across astronomy and other fields of science. Prominent algorithms include neural ratio estimation (Hermans et al. 2020; Durkan et al. 2020), which builds a surrogate of the likelihood-to-evidence ratio, and neural likelihood estimation (Papamakarios & Murray 2016; Alsing et al. 2018; Papamakarios et al. 2019), which learns a fast and tractable surrogate of the likelihood.

5. Conclusion

In this paper, we implemented a simulation-based inference algorithm called NPE to perform Bayesian retrievals of exoplanet atmospheres. Unlike the commonly used nested sampling and MCMC methods, which perform sequential sampling to construct a joint posterior of all model parameters using an explicit and tractable likelihood function, NPE relies on normalizing flows to estimate the posterior in an amortized way, without requiring an explicit or tractable likelihood. This offers several benefits over standard algorithms.

First, NPE is time efficient due to amortization. The inference network needs to be trained only once, and the same network can be used to perform quasi-instantaneous retrievals over several observations without starting from scratch. We demonstrated this by performing 1000 retrievals with synthetic observations constructed by sampling randomly from the prior. This procedure took 34.5 h in total, leading to a computational speed up of 4000 over `MultiNest`. The initial overhead cost of simulations was around 17 h, which can be easily compensated as the number of observations increases. In the case where several simulation models f need to be tested for the retrieval on a given observation, NPE still provides a speed up of over a factor four (134/30).

Second, NPE is scalable. Since the inference network is trained on the parameters of interest only, performance does not deteriorate as quickly as sampling-based algorithms that must navigate the full joint posterior over both the parameters of interest and the nuisance parameters. This is especially important for future simulation models that are likely to include a large number of nuisance parameters.

Lastly, NPE is testable. Since the inference of many observations takes only seconds to perform, one can easily check the validity of NPE by performing posterior predictive checks and producing coverage plots, which is almost impossible to achieve in the case of sequential algorithms. The results presented in this study confirm that NPE provides computationally faithful posteriors, without any simplifying assumption on the shape of the posterior, yet with a possible sign of being slightly underdispersed. While such tests cannot be performed with `Multinest` to provide a fair comparison, our mock retrievals suggest that NPE may be less underdispersed and more faithful than `Multinest`.

NPE's computational speed up opens the possibility of efficiently retrieving atmospheric parameters from large datasets of exoplanet spectra. The speed up provided in the retrieval of individual spectra also enables the exploration of several different simulation models over limited observations in a reasonable time. The prospect of subjecting these retrievals to evaluation metrics such as posterior predictive checks and coverage plots ensures a statistical rigour to the associated results. This establishes NPE as a robust algorithm to perform time-efficient retrievals in the future.

Acknowledgements. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreements No 819155 and 832428), and from the Wallonia-Brussels Federation (grant for Concerted Research Actions). G.L. is recipient of the ULiège – NRB Chair on Big Data and is thankful for the support of the NRB.

References

Ackerman, A. S., & Marley, M. S. 2001, ApJ, 556, 872
Agakov, D. B. F. 2004, Adv. Neural Inform. Process. Syst., 16, 201
Alsing, J., Wandelt, B., & Feeney, S. 2018, MNRAS, 477, 2874
Amundsen, D. S., Tremblin, P., Manners, J., Baraffe, I., & Mayne, N. J. 2017, A&A, 598, A97
Ardévol Martínez, F., Min, M., Kamp, I., & Palmer, P. I. 2022, A&A, 662, A108
Ballard, S., Fabrycky, D., Fressin, F., et al. 2011, ApJ, 743, 200
Baso, C. D., Ramos, A. A., & de la Cruz Rodríguez, J. 2022, A&A, 659, A165
Benneke, B., & Seager, S. 2013, ApJ, 778, 153
Bister, T., Erdmann, M., Köthe, U., & Schulte, J. 2022, Eur. Phys. J. C, 82, 1
Blecic, J. 2016, arXiv e-prints [arXiv:1604.02692]
Buchner, J., Georgakakis, A., Nandra, K., et al. 2014, A&A, 564, A125
Burningham, B., Marley, M. S., Line, M. R., et al. 2017, MNRAS, 470, 1177
Clevert, D.-A., Unterthiner, T., & Hochreiter, S. 2015, arXiv e-prints [arXiv:1511.07289]
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, J. Cosmol. Astropart. Phys., 2022, 004
Cranmer, K., Brehmer, J., & Louppe, G. 2020, PNAS, 117, 30055
Dax, M., Green, S. R., Gair, J., et al. 2021, Phys. Rev. Lett., 127, 241103
Durkan, C., Murray, I., & Papamakarios, G. 2020, in International Conference on Machine Learning, PMLR, 2771
Evans, T. M., Sing, D. K., Kataria, T., et al. 2017, Nature, 548, 58
Feautrier, P. 1964, Compt. Rendus Acad. Sci., 258, 3189

Feroz, F., & Hobson, M. P. 2008, MNRAS, 384, 449

Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601

Feroz, F., Hobson, M. P., Cameron, E., & Pettitt, A. N. 2019, Open J. Astrophys., 2, 10

Gandhi, S., & Madhusudhan, N. 2018, MNRAS, 474, 271

Greenberg, D., Nonnenmacher, M., & Macke, J. 2019, in International Conference on Machine Learning, PMLR, 2404

Hahn, C., & Melchior, P. 2022, ApJ, 938, 11

Hayes, J. J., Kerins, E., Awiphan, S., et al. 2020, MNRAS, 494, 4492

He, K., Zhang, X., Ren, S., & Sun, J. 2016, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770

Helling, C., & Woitke, P. 2006, A&A, 455, 325

Hermans, J., Begy, V., & Louppe, G. 2020, in International Conference on Machine Learning, PMLR, 4239

Hermans, J., Delaunoy, A., Rozet, F., et al. 2021, ArXiv e-prints [arXiv:2110.06581]

Himes, M. D., Harrington, J., Cobb, A. D., et al. 2022, Planet. Sci. J., 3, 91

Huang, C.-W., Krueger, D., Lacoste, A., & Courville, A. 2018, in International Conference on Machine Learning, PMLR, 2078

Kodi Ramanah, D., Wojtak, R., Ansari, Z., Gall, C., & Hjorth, J. 2020, MNRAS, 499, 1985

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

Lee, E. K. H., Blecic, J., & Helling, C. 2018, A&A, 614, A126

Line, M. R., Wolf, A. S., Zhang, X., et al. 2013, ApJ, 775, 137

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

Loshchilov, I., & Hutter, F. 2017, ArXiv e-prints [arXiv:1711.05101]

Lueckmann, J.-M., Goncalves, P. J., Bassetto, G., et al. 2017, ArXiv e-prints [arXiv:1711.01861]

MacDonald, R. J., & Madhusudhan, N. 2017, MNRAS, 469, 1979

Madhusudhan, N. 2018, Atmospheric Retrieval of Exoplanets, eds. H. J. Deeg, & J. A. Belmonte (Cham: Springer International Publishing), 1

Madhusudhan, N., Harrington, J., Stevenson, K. B., et al. 2011, Nature, 469, 64

Madhusudhan, N., Crouzet, N., McCullough, P. R., Deming, D., & Hedges, C. 2014, ApJ, 791, L9

Márquez-Neila, P., Fisher, C., Sznitman, R., & Heng, K. 2018, Nat. Astron., 2, 719

Mishra-Sharma, S., & Cranmer, K. 2022, Phys. Rev. D, 105, 063017

Mollière, P., van Boekel, R., Dullemond, C., Henning, T., & Mordasini, C. 2015, ApJ, 813, 47

Mollière, P., van Boekel, R., Bouwman, J., et al. 2017, A&A, 600, A10

Mollière, P., Wardenier, J., van Boekel, R., et al. 2019, A&A, 627, A67

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

Nixon, M. C., & Madhusudhan, N. 2020, MNRAS, 496, 269

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

Papamakarios, G., & Murray, I. 2016, Adv. Neural Inform. Process. Syst., 29

Papamakarios, G., Sterratt, D., & Murray, I. 2019, in The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 837

Papamakarios, G., Nalisnick, E. T., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. 2021, J. Mach. Learn. Res., 22, 1

Ramos, A. A., Baso, C. D., & Kochukhov, O. 2022, A&A, 658, A162

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. 2018, arXiv e-prints [arXiv:1804.06788]

Todorov, K. O., Line, M. R., Pineda, J. E., et al. 2016, ApJ, 823, 14

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

Waldmann, I. P., Rocchetto, M., Tinetti, G., et al. 2015a, ApJ, 813, 13

Waldmann, I. P., Tinetti, G., Rocchetto, M., et al. 2015b, ApJ, 802, 107

Woitke, P., Helling, C., & Gunn, O. 2020, A&A, 634, A23

Yip, K. H., Changeat, Q., Al-Refaie, A., & Waldmann, I. 2022, ApJ, submitted [arXiv:2205.07037]

Zhang, K., Bloom, J. S., Gaudi, B. S., et al. 2020, arXiv e-prints [arXiv:2010.04156]

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. 2021, arXiv e-prints [arXiv:2106.11342]

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

Appendix A: Normalizing flows 101

Normalizing flows (Papamakarios et al. 2021) are invertible mappings transforming a simple probability distribution to a complex one. The change in the probability density of a random variable u due to a invertible transformation g is given by the change of variables theorem as

$$\log p(v) = \log p(u) - \log \left| \det \frac{\partial g(u)}{\partial u} \right|,$$

where $v = g(u)$, and the determinant accounts for the change in volume between the two distributions. Because the transformation is invertible, the opposite direction $u = t(v)$ where $t = g^{-1}$ is also tractable. In this direction, the density "flows" from a complex distribution to a Normal distribution, hence the name "normalizing flows" (NFs).

To increase the expressiveness of NFs, parametric transformations can be stacked up as $v = g_n \circ g_{n-1} \circ \ldots \circ g_1(u)$, which results in the probability density

$$\log p(v) = \log p(u) - \sum_{i=1}^{n} \log \left| \det \frac{\partial g_i(u_{i-1})}{\partial u_{i-1}} \right|,$$

where $u_i = g_i(u_{i-1})$ and $u_0 = u$.

In this work, we similarly model the posterior density $p_\phi(\theta|x)$ of the variable θ through a series of transformations of a Normal random variable u with probability density $p(u)$, as illustrated in Fig. A.1. In the case of neural autoregressive flows, the transformations are invertible neural networks conditioned to the observation x.

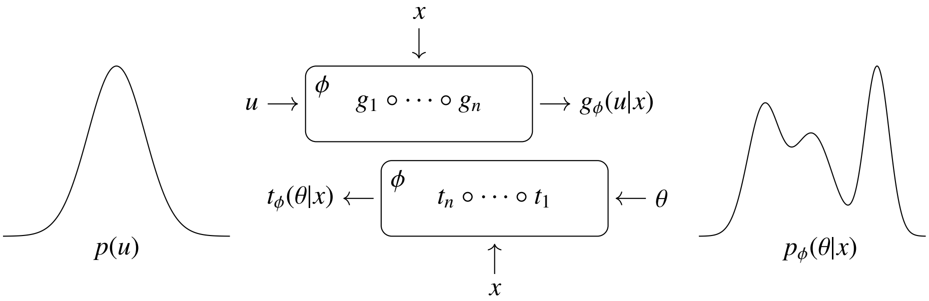

Fig. A.1: Transformation of a random variable z with probability density $p(z)$ toward a variable θ with probability density $p_\phi(\theta|x)$.