Reliable Parameter Inference for the Epoch of Reionization using Balanced Neural Ratio Estimation

Diego González-Hernández
Department of Physics
University of California, Santa Barbara
Santa Barbara, CA 93106, USA
Lawrence Berkeley National Laboratory
Berkeley, CA 94720, USA
dgonzalezhernandez@ucsb.edu

Molly Wolfson
Center for Cosmology and AstroParticle Physics
Department of Physics
Department of Astronomy
The Ohio State University
Columbus, OH 43210, USA
wolfson.63@osu.edu

Joseph F. Hennawi
Department of Physics
University of California, Santa Barbara
Santa Barbara, CA 93106, USA
Leiden Observatory, Leiden University
Niels Bohrweg 2, 2333 CA Leiden, Netherlands
joe@physics.ucsb.edu

Abstract

We present an application of the Balanced Neural Ratio Estimation (BNRE) algorithm to improve the statistical validity of parameter estimates used to characterize the Epoch of Reionization, where the common assumption of a multivariate Gaussian likelihood leads to overconfident and biased posterior distributions. Using a two-parameter model of the Lyα forest autocorrelation function, we show that BNRE yields posterior distributions that are significantly better calibrated than those obtained under the Gaussian likelihood assumption, as verified through the Test of Accuracy with Random Points (TARP) and Simulation-Based Calibration (SBC) diagnostics. These results demonstrate the potential of Simulation-Based Inference (SBI) methods, and in particular BNRE, to provide statistically robust parameter constraints within existing astrophysical modeling frameworks.

1 Introduction

The Epoch of Reionization (EoR) corresponds to the time in cosmic history when the neutral hydrogen in the intergalactic medium (IGM) was ionized by the first luminous sources [see e.g. 1]. One of the primary probes of the late stages of the EoR is the Lyman-α (Lyα) forest, a series of redshifted absorption features observed in the spectra of distant quasars caused by the presence of neutral hydrogen along the line of sight [2, 3]. Typically, two-point summary statistics of the Lyα forest (such as the one-dimensional power spectrum and the autocorrelation function) are used to understand the properties of the IGM during the EoR [e.g. 4, 5, 6, 7]. In standard analyses, parameter estimation relies on the assumption of a multivariate Gaussian likelihood for these statistics, an approximation that has been shown to yield overconfident or biased posteriors [e.g. 8, 9]. Such miscalibrations could lead to incorrect scientific conclusions, making it important to seek solutions that properly address this issue. Fortunately, the recent development of machine learning powered Simulation-Based Inference (SBI) methods has provided an alternative by directly leveraging forward simulations, allowing us to let go of explicit likelihood assumptions [10]. In this work, we present our preliminary results

obtained by applying Balanced Neural Ratio Estimation (BNRE) to a two-parameter model of the EoR, demonstrating its potential to improve the statistical validity in parameter inference problems within cosmology.[1]

2   Modeling the Epoch of Reionization

We adopt the model introduced in [8], which combines simulations of the IGM with semi-numerical calculations of the Ultraviolet Background (UVB). For our study, we focus on the redshift $z = 5.5$, where this EoR model can be used to generate the mean autocorrelation function of the Lyα forest $\boldsymbol{\xi}_m$ as a function of two parameters: the mean free path of ionizing photons λ_{mfp}, and the mean transmitted flux $\langle F \rangle$. We refer the reader to [8] for specific details, but in summary, the model employs Nyx cosmological hydrodynamical simulations [11, 12] to model the underlying density, temperature, and velocity fields, while the method from [13] is used to create spatially varying UVB realizations parameterized by λ_{mfp}. Forward modeling is used to incorporate instrumental resolution, noise, and sightline lengths consistent with realistic observational data (e.g., the XQR-30 data from [14]). For each choice of $\boldsymbol{\theta} = \{\lambda_{\mathrm{mfp}}, \langle F \rangle\}$, the model produces ensembles of mock observations of the autocorrelation function of the Ly-α forest $\boldsymbol{\xi}_i$, the mean autocorrelation function $\boldsymbol{\xi}_m$ and the corresponding covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$.

A crucial aspect of this modeling approach is that each parameter combination is represented not by a single deterministic output, but by an arbitrarily large number of mock realizations required to properly sample the subsets of randomly selected sightlines and the observational noise. These mocks are already necessary for constructing model-dependent covariance matrices in traditional analyses, but they also naturally define a stochastic simulator. This makes the model ideally suited for Simulation-Based Inference (SBI), where the availability of many forward-modeled realizations per parameter point enables the training of neural estimators without additional modifications.

3   Parameter Estimation

Our goal is to infer the astrophysical parameters $\boldsymbol{\theta} = \{\lambda_{\mathrm{mfp}}, \langle F \rangle\}$ given an observed Lyα forest autocorrelation function $\boldsymbol{\xi}_{\mathrm{obs}}$. Using a Bayesian framework, the posterior distribution is given by:

$$p(\boldsymbol{\theta} \mid \boldsymbol{\xi}_{\mathrm{obs}}) \propto \mathcal{L}(\boldsymbol{\xi}_{\mathrm{obs}} \mid \boldsymbol{\theta}) \, p(\boldsymbol{\theta}), \tag{1}$$

where $\mathcal{L}(\boldsymbol{\xi}_{\mathrm{obs}} \mid \boldsymbol{\theta})$ is the likelihood function and $p(\boldsymbol{\theta})$ is the prior over the model parameters. The choice of likelihood is therefore central to parameter inference, as it directly impacts the statistical validity of the resulting posterior distributions.

We evaluate the statistical validity of the posterior distributions using two methods. First, we use *coverage tests*, which assess whether the credible intervals of the inferred posterior distributions contain the true parameters with the expected frequency. Given our set of mock observations, coverage can be tested directly since we have access to their corresponding true $\boldsymbol{\theta}$. We employ the Test of Accuracy with Random Points [TARP, 15], which provides a diagnostic for posterior miscalibration by repeatedly comparing inferred credible regions against ground-truth parameter values. Second, we apply *Simulation-Based Calibration* [SBC, 16, 17], which evaluates whether the ranks of true parameters within their corresponding posterior samples are uniformly distributed. While TARP quantifies global coverage performance, SBC provides a check of local miscalibration across the parameter space. Using both methods as inference tests provides a robust check on the statistical validity of the posteriors.

In what follows, we first describe the standard assumption of a multivariate Gaussian likelihood function adopted in previous works (Section 3.1), before describing an SBI method based on Balanced Neural Ratio Estimation (Section 3.2).

3.1   Assuming a Gaussian Likelihood

A common approach in Lyα forest analyzes is to assume that the summary statistic of choice ($\boldsymbol{\xi}$ in this case) follows a multivariate Gaussian distribution with fixed model parameters. Under this

---

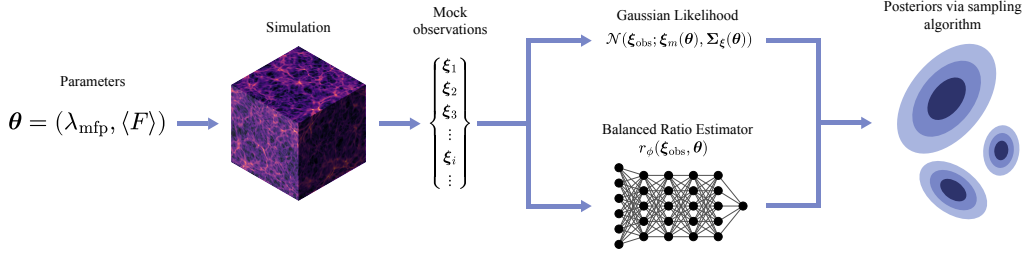[1]All code and analysis scripts are available at `https://github.com/diego-gonher/laf_sbi_bnre`.

Figure 1: Given a set of parameters $\boldsymbol{\theta}$, the model generates a set of mock Lyα forest autocorrelation functions. These mocks can be used to: (i) compute the mean autocorrelation function $\boldsymbol{\xi}_m$ and the corresponding covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ required to evaluate an assumed multivariate Gaussian likelihood, or (ii) train a neural ratio estimator (BNRE in this study). Both approaches yield posterior distributions via MCMC sampling (`emcee` in the Gaussian case, and HMC/NUTS in the BNRE case).

assumption, the likelihood function can be written as:

$$\mathcal{L}(\boldsymbol{\xi}_{\mathrm{obs}} \mid \boldsymbol{\theta}) \propto \exp\left[-\tfrac{1}{2} \left(\boldsymbol{\xi}_{\mathrm{obs}} - \boldsymbol{\xi}_m(\boldsymbol{\theta}) \right)^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\xi}}(\boldsymbol{\theta})^{-1} \left(\boldsymbol{\xi}_{\mathrm{obs}} - \boldsymbol{\xi}_m(\boldsymbol{\theta}) \right) \right], \tag{2}$$

Where $\boldsymbol{\xi}_m(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}(\boldsymbol{\theta})$ denote the mean and covariance of the autocorrelation function evaluated at $\boldsymbol{\theta}$. In practice, these quantities are typically precomputed on a grid in parameter space, and inference is carried out using Markov Chain Monte Carlo (MCMC). Following [8], we use a grid of 557 distinct parameter combinations spanning $\boldsymbol{\theta} = \{\lambda_{\mathrm{mfp}}, \langle F \rangle\}$, with each grid point having its own $\boldsymbol{\xi}_m$ and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$. We then use the `emcee` package [18] with a nearest-grid-point interpolation scheme to evaluate $\boldsymbol{\xi}_m$ and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ between grid points. As mentioned above, this likelihood assumption often leads to overconfident or biased posteriors [see Fig. 9 and Appendix C in 8].

3.2 Balanced Neural Ratio Estimation

Neural Ratio Estimation algorithms are a subset of SBI methods that bypass the need for an explicit likelihood function by training neural networks to approximate the ratio between the joint distribution $p(\boldsymbol{\xi}, \boldsymbol{\theta})$ and the product of marginals $p(\boldsymbol{\xi})p(\boldsymbol{\theta})$ [10]. Balanced Neural Ratio Estimation (BNRE) [19] extends this framework by introducing a tunable hyperparameter γ in the loss function, which balances likelihood-ratio estimation accuracy against posterior calibration. By adjusting γ, BNRE can be made to nearly satisfy coverage tests.

We use the BNRE implementation in the `sbi` Python package [20, 21]. The training dataset consists of 557×500 mock realizations $\boldsymbol{\xi}_i$, where 557 corresponds to the same $\boldsymbol{\theta}$ combinations on the grid described above (see Section 3.1), and for each $\boldsymbol{\theta}$ we have 500 distinct mocks (a random subset of the mocks used to estimate each $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$, which means no extra computations were required for the creation of this dataset). We use a 70–30 split for the training and validation sets and train the ratio estimator. The architecture for our estimator is the default ResNet-based classifier provided in the `sbi` package, which constructs a residual network with two blocks of 50 hidden units and ReLU activations, operating on the concatenated $(\boldsymbol{\xi}_i, \boldsymbol{\theta})$ inputs. For the training, we set $\gamma = 100$ (although we tested the effect of using $\gamma = \{10, 1000\}$, see Section 5 and Appendix B). Once trained, the ratio estimator defines the posterior via:

$$p(\boldsymbol{\theta} \mid \boldsymbol{\xi}_{\mathrm{obs}}) \propto r_{\phi}(\boldsymbol{\xi}_{\mathrm{obs}}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{3}$$

Where r_{ϕ} denotes the learned likelihood ratio. Because r_{ϕ} is differentiable with respect to $\boldsymbol{\theta}$, we can efficiently sample from the posterior using Hamiltonian Monte Carlo [HMC, 22, 23]. Specifically, we use the No-U-Turn Sampler [NUTS, 24] implementation from the `pyro` library [25], which exploits gradient information to explore parameter space more effectively. Consequently, the combination of BNRE with HMC is able to achieve both statistically valid and computationally efficient parameter estimations.

To summarize, both inference approaches rely on the same underlying dataset of mock autocorrelation functions, but use them in different ways: either to compute mean statistics and covariances for a
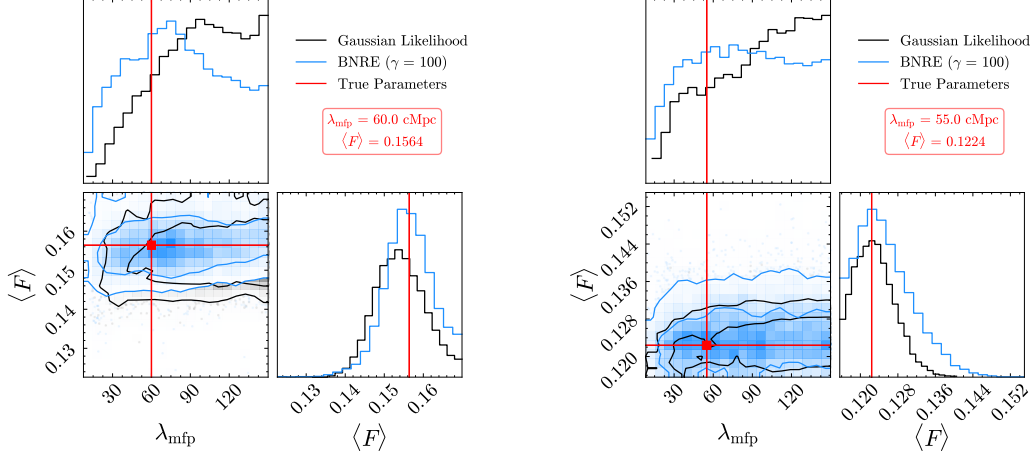
Figure 2: Corner plots of the posterior distributions obtained with both inference methods for two separate mock observations. The red text under the legend indicates the corresponding true parameter values $(\lambda_{\mathrm{mfp}}, \langle F \rangle)_{\mathrm{true}}$. Contours denote the 68% and 95% credible regions.

Gaussian likelihood, or to train a neural ratio estimator. A schematic overview of both inference methods is provided in Figure 1.

4   Results

To evaluate the performance of the two inference methods described above, we perform parameter inference on randomly selected mock observations (such that $\boldsymbol{\xi}_{\mathrm{obs}} = \boldsymbol{\xi}_i$). For the MCMC sampling, we use four chains with 2000 samples each, discarding the first 1000 as "burn-in". This allows us to make a direct comparison of both inference methods, and test their statistical validity. As illustrative examples, Figure 2 shows the posterior distributions obtained by either assuming a Gaussian likelihood or using our trained likelihood ratio estimator on two different mock observations, with additional examples shown in Figure 4 in Appendix A.

To quantitatively assess posterior calibration, we apply both parameter inference methods to a common set of 500 independent mock observations that were not used in training the BNRE's ratio estimator. The resulting posteriors are used to perform the inference tests presented below.

4.1   Coverage Test

As mentioned in Section 3, we use TARP as our main coverage test. To estimate the statistical uncertainty in the coverage curves, we follow the bootstrapping procedure explained in [26]. For both inference methods, we resample the set of mock observations and their associated posterior samples, recompute the TARP curve for each bootstrap realization, and use the resulting distribution to construct confidence intervals.

The right panel of Figure 3 shows the results of our coverage test. As can be seen, assuming a Gaussian likelihood leads to clearly overconfident or biased posterior distributions (shown in black). For the EoR model that we are using, this result is in agreement with the coverage test done in [8]. In contrast, the coverage probability obtained with BNRE (shown in light blue) is close to the ideal curve, demonstrating that the posterior distributions obtained by this method are significantly less overconfident or biased. The corresponding shaded bands show the 16th–84th percentile ranges for both methods across 100 bootstrap replicates.

4.2   Simulation Based Calibration

As described in Section 3, we also perform SBC to further verify posterior validity. The left panel of Figure 3 shows the SBC rank histograms computed using all posteriors obtained by both methods, with the shaded region indicating the expected variation under perfect uniformity [16, 17]. Assuming a
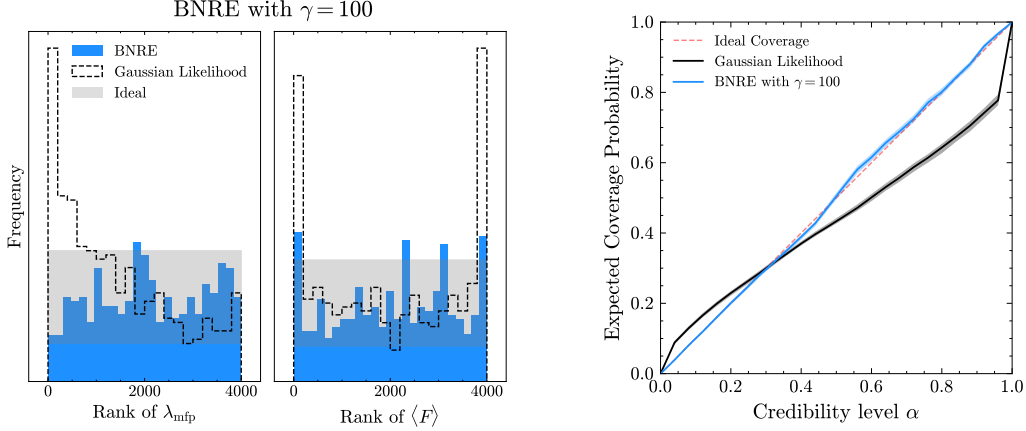
Figure 3: *Left:* SBC rank histograms for λ_{mfp} and $\langle F \rangle$ obtained by both parameter inference methods. Uniform posteriors indicate correct calibration, with the shaded region showing the expected range under sampling variability. *Right:* Coverage probabilities obtained by using TARP on a set of posterior distributions obtained with both methods. The shaded regions represent the respective 16th–84th percentile ranges obtained via bootstrap sampling.

Gaussian likelihood produces a skewed rank distribution for λ_{mfp} and a slightly U-shaped distribution for $\langle F \rangle$, indicating that this approach leads to biased posteriors for λ_{mfp} and slight overconfidence for $\langle F \rangle$. These results are consistent with the shape of the corresponding coverage curve. In contrast, the BNRE rank distributions for both parameters lie largely within the ideal uniformity region.

5 Conclusions

Our results provide encouraging evidence that the BNRE algorithm can improve the statistical validity of parameter inference in cosmological models of the Epoch of Reionization. Although our study focuses on a two-parameter model, it is representative of common Lyα forest studies and serves as a non-trivial test for evaluating different inference methods. As mentioned in Section 3.2, we also train the same BNRE architecture with $\gamma = \{10, 1000\}$ to examine the sensitivity of our results to this hyperparameter. Despite minor variations in the rank distributions of $\langle F \rangle$ (see Figures 5 and 6 in Appendix B), all trained ratio estimators yield posterior distributions that are substantially better calibrated than those obtained under the Gaussian likelihood assumption.

To obtain a simple performance comparison between BNRE and a different SBI algorithm, we also train a Neural Posterior Estimation (NPE) and evaluate its performance (see Appendix C for details). The resulting posteriors are clearly overconfident. We emphasize that this result is not representative of NPE's overall potential. We expect that with adequate hyperparameter tuning and the use of a balanced objective [as seen in 27], NPE could likely achieve a comparable performance. Nonetheless, training the NPE density estimator is substantially more computationally demanding than training the BNRE classifier, which constitutes a practical advantage for BNRE[2]

In future work, we will explore the application of BNRE and other SBI methods to more complex, higher-dimensional reionization models. Importantly, standard modeling techniques across different subfields of astrophysics require little to no modifications to adopt SBI methods, as the same mock observations used to test inference pipelines can be repurposed for training different SBI algorithms. Our results therefore highlight how SBI methods such as BNRE can serve as a practical and reliable tool for enabling statistically robust scientific inference.

---

[2]In this study, training required approximately 7.9 CPU hours for NPE and 1.2 CPU hours for BNRE.

Acknowledgments

References

[1] Nickolay Y. Gnedin and Piero Madau. Modeling cosmic reionization. *Living Reviews in Computational Astrophysics*, 8(1):3, December 2022.

[2] James E. Gunn and Bruce A. Peterson. On the Density of Neutral Hydrogen in Intergalactic Space. *ApJ*, 142:1633–1636, November 1965.

[3] Roger Lynds. The Absorption-Line Spectrum of 4c 05.34. *ApJL*, 164:L73, March 1971.

[4] Elisa Boera, George D. Becker, James S. Bolton, and Fahad Nasir. Revealing Reionization with the Thermal History of the Intergalactic Medium: New Constraints from the Lyα Flux Power Spectrum. *ApJ*, 872(1):101, February 2019.

[5] Prakash Gaikwad, Raghunathan Srianand, Martin G. Haehnelt, and Tirthankar Roy Choudhury. A consistent and robust measurement of the thermal state of the IGM at $2 \leq z \leq 4$ from a large sample of Ly α forest spectra: evidence for late and rapid He II reionization. *MNRAS*, 506(3):4389–4412, September 2021.

[6] Michael Walther, Eric Armengaud, Corentin Ravoux, Nathalie Palanque-Delabrouille, Christophe Yèche, and Zarija Lukić. Simulating intergalactic gas for DESI-like small scale Lymanα forest observations. *JCAP*, 2021(4):059, April 2021.

[7] Molly Wolfson, Joseph F. Hennawi, Frederick B. Davies, Zarija Lukić, and Jose Oñorbe. Forecasting constraints on the high-z IGM thermal state from the Lyman-α forest flux auto-correlation function. *arXiv e-prints*, page arXiv:2309.05647, September 2023.

[8] Molly Wolfson, Joseph F. Hennawi, Frederick B. Davies, and Jose Oñorbe. Forecasting constraints on the mean free path of ionizing photons at $z \geq 5.4$ from the Lyman-α forest flux autocorrelation function. *MNRAS*, 521(3):4056–4073, May 2023.

[9] Zhenyu Jin, Molly Wolfson, Joseph F. Hennawi, and Diego González-Hernández. Neural network emulator to constrain the high-z IGM thermal state from Lyman-α forest flux autocorrelation function. *MNRAS*, 536(3):2277–2293, January 2025.

[10] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Science*, 117(48):30055–30062, December 2020.

[11] Ann S. Almgren, John B. Bell, Mike J. Lijewski, Zarija Lukić, and Ethan Van Andel. Nyx: A Massively Parallel AMR Code for Computational Cosmology. *ApJ*, 765(1):39, March 2013.

[12] Zarija Lukić, Casey W. Stark, Peter Nugent, Martin White, Avery A. Meiksin, and Ann Almgren. The Lyman α forest in optically thin hydrodynamical simulations. *MNRAS*, 446(4):3697–3724, February 2015.

[13] Frederick B. Davies and Steven R. Furlanetto. Large fluctuations in the hydrogen-ionizing background and mean free path following the epoch of reionization. *MNRAS*, 460(2):1328–1339, August 2016.

[14] Valentina D'Odorico, E. Bañados, G. D. Becker, M. Bischetti, S. E. I. Bosman, G. Cupani, R. Davies, E. P. Farina, A. Ferrara, C. Feruglio, C. Mazzucchelli, E. Ryan-Weber, J. T. Schindler, A. Sodini, B. P. Venemans, F. Walter, H. Chen, S. Lai, Y. Zhu, F. Bian, S. Campo, S. Carniani, S. Cristiani, F. Davies, R. Decarli, A. Drake, A. C. Eilers, X. Fan, P. Gaikwad, S. Gallerani,

B. Greig, M. G. Haehnelt, J. Hennawi, L. Keating, G. Kulkarni, A. Mesinger, R. A. Meyer, M. Neeleman, M. Onoue, A. Pallottini, Y. Qin, S. Rojas-Ruiz, S. Satyavolu, A. Sebastian, R. Tripodi, F. Wang, M. Wolfson, J. Yang, and M. V. Zanchettin. XQR-30: The ultimate XSHOOTER quasar sample at the reionization epoch. *MNRAS*, 523(1):1399–1420, July 2023.

[15] Pablo Lemos, Adam Coogan, Yashar Hezaveh, and Laurence Perreault-Levasseur. Sampling-Based Accuracy Testing of Posterior Estimators for General Inference. *40th International Conference on Machine Learning*, 202:19256–19273, January 2023.

[16] Samantha R Cook, Andrew Gelman, and Donald B Rubin. Validation of software for bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692, 2006.

[17] Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. *arXiv e-prints*, page arXiv:1804.06788, April 2018.

[18] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. emcee: The MCMC Hammer. *PASP*, 125(925):306, March 2013.

[19] Arnaud Delaunoy, Joeri Hermans, François Rozet, Antoine Wehenkel, and Gilles Louppe. Towards Reliable Simulation-Based Inference with Balanced Neural Ratio Estimation. *arXiv e-prints*, page arXiv:2208.13624, August 2022.

[20] Alvaro Tejero-Cantero, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro J. Gonçalves, David S. Greenberg, and Jakob H. Macke. sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, 2020.

[21] Jan Boelts, Michael Deistler, Manuel Gloeckler, Álvaro Tejero-Cantero, Jan-Matthis Lueckmann, Guy Moss, Peter Steinbach, Thomas Moreau, Fabio Muratore, Julia Linhart, Conor Durkan, Julius Vetter, Benjamin Miller, Maternus Herold, Abolfazl Ziaeemehr, Matthijs Pals, Theo Gruner, Sebastian Bischoff, Nastya Krouglova, Richard Gao, Janne Lappalainen, Bálint Mucsányi, Felix Pei, Auguste Schulz, Zinovia Stefanidi, Pedro Rodrigues, Cornelius Schröder, Faried Zaid, Jonas Beck, Jaivardhan Kapoor, David Greenberg, Pedro Gonçalves, and Jakob Macke. sbi reloaded: a toolkit for simulation-based inference workflows. *The Journal of Open Source Software*, 10(108):7754, April 2025.

[22] Radford Neal. MCMC Using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. 2011.

[23] Simon Duane, A.D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics Letters B*, 195(2):216–222, 1987.

[24] Matthew D. Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *arXiv e-prints*, page arXiv:1111.4246, November 2011.

[25] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019.

[26] A. Ruzza, G. Lodato, G. P. Rosotti, and P. J. Armitage. DBNets2.0: Simulation-based inference for planet-induced dust substructures in protoplanetary discs. *A&A*, 700:A190, August 2025.

[27] Arnaud Delaunoy, Benjamin Kurt Miller, Patrick Forré, Christoph Weniger, and Gilles Louppe. Balancing Simulation-based Inference for Conservative Posteriors. *arXiv e-prints*, page arXiv:2304.10978, April 2023.

[28] David S. Greenberg, Marcel Nonnenmacher, and Jakob H. Macke. Automatic Posterior Transformation for Likelihood-Free Inference. *arXiv e-prints*, page arXiv:1905.07488, May 2019.

[29] Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, Volodimir Begy, and Gilles Louppe. A crisis in simulation-based inference? beware, your posterior approximations can be unfaithful. *Transactions on Machine Learning Research*, 2022.

A   Additional Corner Plots

Figure 4 presents four additional (randomly selected) examples comparing the posterior distributions obtained using the Gaussian likelihood and BNRE with $\gamma = 100.0$. These examples show the typical differences in posterior shape and coverage across independent mock observations.
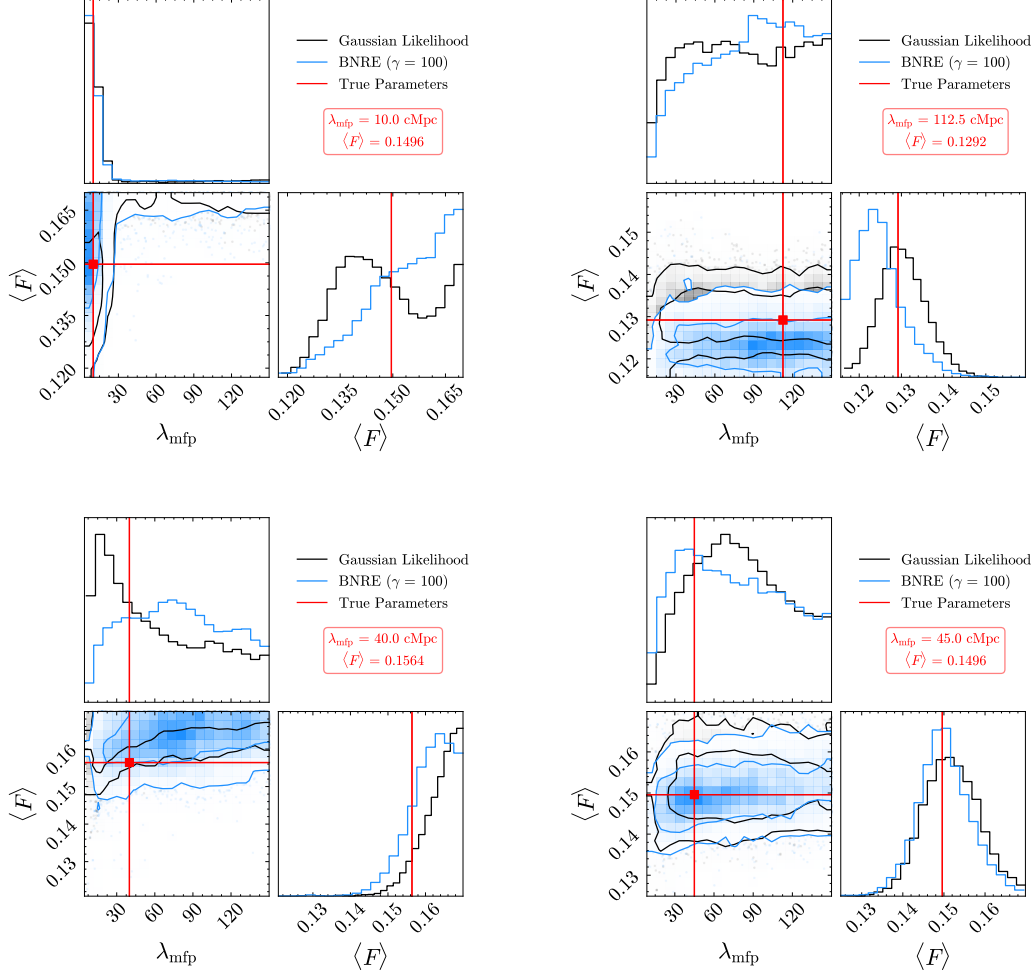

Figure 4: Corner plots of the posterior distributions obtained using the Guassian likelihood and BNRE with $\gamma = 100.0$ for four randomly selected mock observations. The true parameter values are shown in red for each case respectively. The contours denote the 68% and 95% credible regions.

B   Performance of BNRE with $\gamma = 10$ and $\gamma = 1000$

Using the same training set and architecture described in Section 3.2, we train two additional ratio estimators with $\gamma = \{10, 1000\}$. Following the procedure outlined in Section 4, we perform parameter inference with both estimators to obtain two respective sets of 500 posterior distributions, and then apply TARP and SBC to directly compare their performance. The corresponding results are shown in Figures 5 and 6.

Both models produce coverage curves that are close to ideal, demonstrating that the calibration of the posterior distributions is robust to moderate changes in γ. The SBC histograms show that the rank distributions for λ_{mfp} and $\langle F \rangle$ also remain largely within the ideal region. However, the rank distribution for $\langle F \rangle$ exhibits a slight U-shape for the $\gamma = 10$ case and a slight inverted U-shape for $\gamma = 1000$. This behavior is expected, as increasing γ reduces the relative weight of the likelihood-ratio loss term, leading to less confident posteriors [see 19].
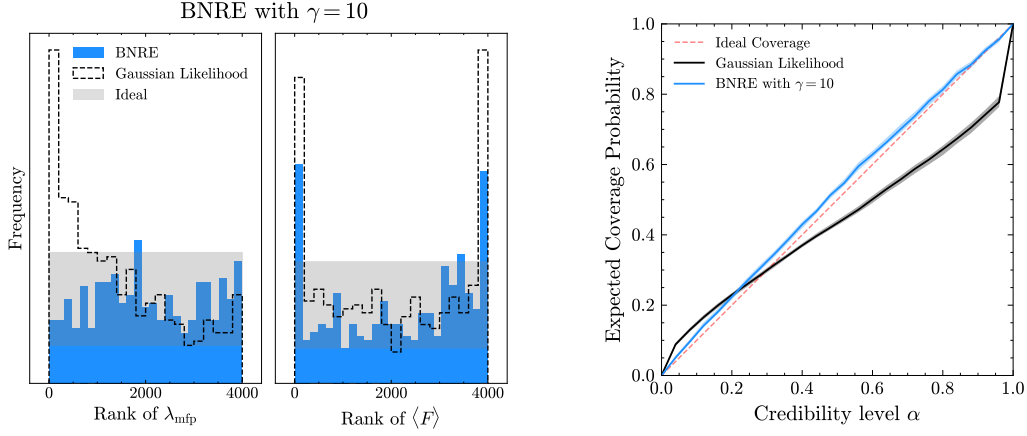
Figure 5: *Left:* SBC rank histograms for λ_{mfp} and $\langle F \rangle$ obtained using BNRE with $\gamma = 10$. Uniform posteriors indicate correct calibration, with the shaded region showing the expected range under sampling variability. *Right:* Coverage probabilities obtained by using TARP on a set of posterior distributions obtained with both methods. The shaded regions represent the respective 16th–84th percentile ranges obtained via bootstrap sampling.
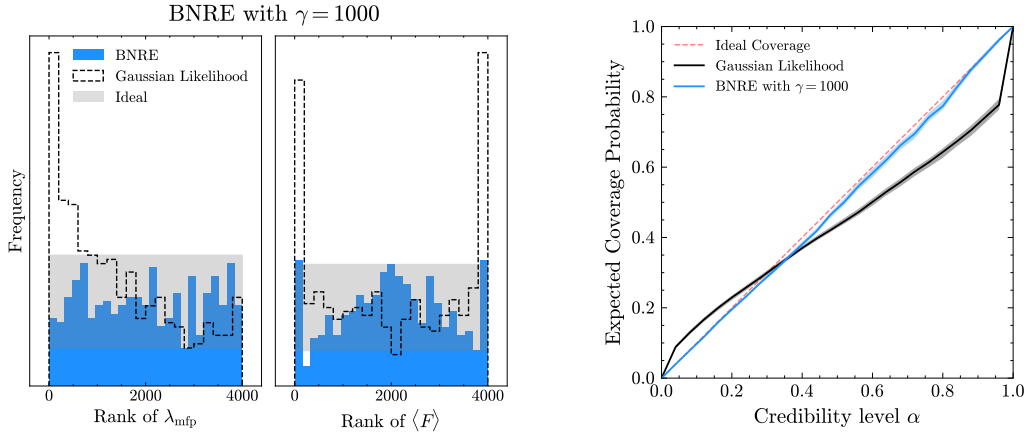

Figure 6: *Left:* SBC rank histograms for λ_{mfp} and $\langle F \rangle$ obtained using BNRE with $\gamma = 1000$. Uniform posteriors indicate correct calibration, with the shaded region showing the expected range under sampling variability. *Right:* Coverage probabilities obtained by using TARP on a set of posterior distributions obtained with both methods. The shaded regions represent the respective 16th–84th percentile ranges obtained via bootstrap sampling.
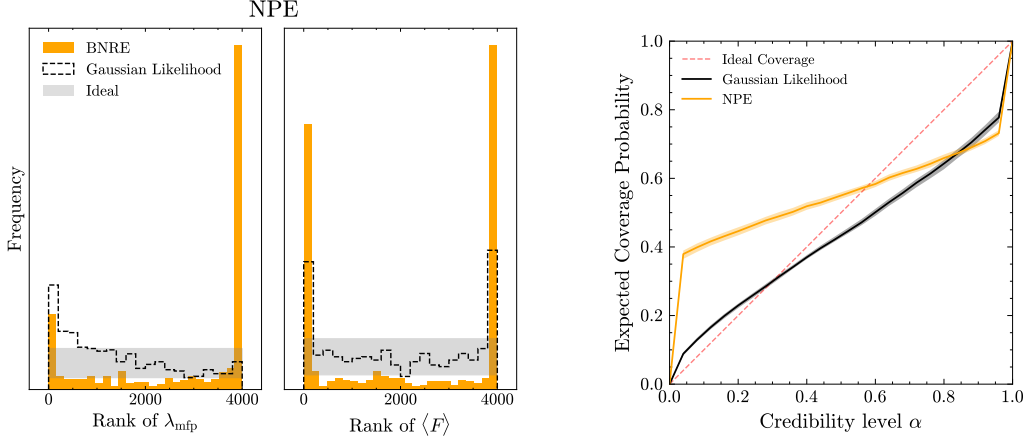
Figure 7: *Left:* SBC rank histograms for λ_{mfp} and $\langle F \rangle$ obtained NPE. Uniform posteriors indicate correct calibration, with the shaded region showing the expected range under sampling variability. *Right:* Coverage probabilities obtained by using TARP on a set of posterior distributions obtained with both methods. The shaded regions represent the respective 16th–84th percentile ranges obtained via bootstrap sampling.

C Neural Posterior Estimation

To provide a simple comparison against an alternative simulation-based inference (SBI) method, we also train a Neural Posterior Estimation (NPE) algorithm. In NPE, a density network is used to model the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{\xi})$ directly. We use the same training and validation datasets as for the BNRE networks (see Section 3.2) to train a posterior estimator, employing the NPE method from [28] available in the `sbi` package [20].

We use the default configuration for the density estimator provided in `sbi`, which consists of a Masked Autoregressive Flow (MAF) composed of five affine autoregressive transforms with random permutations between them. Each transform uses fully connected layers with 50 hidden units and $\tanh$ activations. The model is optimized using maximum likelihood with early stopping based on the validation loss, again using the default settings in `sbi`.

To evaluate its performance, we use the trained density estimator to infer posterior distributions for the same 500 mock observations as in Section 4, drawing 4000 samples per mock. Figure 7 shows the results of the TARP and SBC tests.

The rank distributions for both parameters exhibit a strong U-shape, indicating that the trained NPE model is overconfident, consistent with the shape of the corresponding coverage curve. This overconfidence a well-known issue in some standard SBI algorithms [29]. However, we strongly emphasize that this result is not representative of NPE's overall potential. We expect that with appropriate hyperparameter tuning or the use of a balanced objective [as proposed by 27], NPE could likely achieve substantially better calibration.