Simulation-Based Inference with Balanced Neural Ratio Estimators: A Standardized Pipeline with Agent-Assisted Workflows

Anonymous Author(s)
Affiliation
Address
email

Abstract

Simulation-based inference (SBI) enables Bayesian parameter estimation in settings where likelihoods are intractable but stochastic simulators are available, yet the effort required to adapt workflows to new problems is non trivial. We present a JAX-based pipeline that combines Balanced Neural Ratio Estimation (BNRE) with Hamiltonian Monte Carlo (HMC) in a standardized, configuration-driven framework. By enforcing a consistent dataset contract and experiment structure, the pipeline enables reproducible training, inference, and diagnostic evaluation across tasks. We further introduce an agent-assisted workflow that automates experiment execution while preserving human control over interpretation. We demonstrate the approach on a cosmological application involving the Lyα forest autocorrelation function, showing that it produces well-calibrated posterior estimates even in the presence of known model degeneracies.

1 Introduction

Simulation-Based Inference (SBI) enables parameter estimation in models where the likelihood is intractable but forward simulations are available [1]. While modern SBI methods provide flexible posterior approximations, ensuring calibration and adapting workflows to new problems remain practical challenges [2]. Even when suitable inference methods exist, applying SBI to a new scientific problem often requires substantial effort, including dataset preparation, experiment design, posterior sampling, and diagnostic evaluation.

In this work, we present a standardized pipeline for calibrated SBI based on Balanced Neural Ratio Estimation (BNRE) combined with Hamiltonian Monte Carlo (HMC). Rather than introducing a new inference method, our goal is to provide a reproducible and extensible framework that unifies dataset structure, model training, inference, and diagnostic evaluation. A key design choice is the use of a fixed dataset contract and configuration-driven workflow, enabling consistent application across problems and optional integration with coding agents. In this setting, the agent acts as a constrained execution tool capable of running full experiments given a valid dataset, while all interpretation remains the responsibility of the researcher.

The main contributions of this work are: (i) a standardized workflow connecting dataset validation, ratio-estimator training, HMC posterior inference, and diagnostic evaluation; (ii) an integrated diagnostic layer based on Simulation-Based Calibration (SBC) and the Test of Accuracy with Random Points (TARP); and (iii) an agent-compatible workflow that enables rapid adaptation of the pipeline to new inference problems while preserving reproducibility.

Submitted to the 2026 Conference on Physics and AI. Do not distribute.

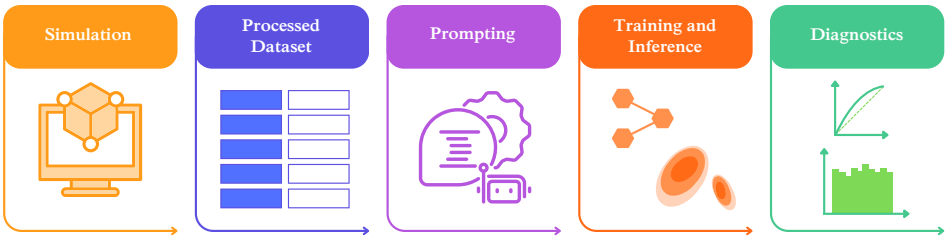

Figure 1: Standardized workflow used throughout the pipeline. Given a simulator-generated dataset, the agent validates the dataset contract, adapts a canonical experiment template, executes BNRE training and HMC inference, and generates calibration diagnostics. Human interpretation occurs after diagnostic evaluation.

We demonstrate the pipeline on a cosmological inference problem based on the Lyα forest autocorrelation function at $z = 5.5$, showing calibrated posterior estimates despite intrinsic degeneracies. The resulting code can be found here: `BNRE + HMC` Repository (currently anonymized).

2 Method

2.1 Balanced Neural Ratio Estimation

Neural Ratio Estimation (NRE) [3, 4, 5] estimates the likelihood-to-evidence ratio by training a classifier to distinguish between joint samples $(\boldsymbol{\theta}, \mathbf{x}) \sim p(\boldsymbol{\theta}, \mathbf{x})$ and independent samples $(\boldsymbol{\theta}, \mathbf{x}') \sim p(\boldsymbol{\theta})p(\mathbf{x})$:

$$r_\phi(\boldsymbol{\theta}, \mathbf{x}) \approx \frac{p(\mathbf{x} \mid \boldsymbol{\theta})}{p(\mathbf{x})}. \tag{1}$$

This ratio defines the posterior up to normalization:

$$p(\boldsymbol{\theta} \mid \mathbf{x}_{\text{obs}}) \propto r_\phi(\boldsymbol{\theta}, \mathbf{x}_{\text{obs}})\, p(\boldsymbol{\theta}). \tag{2}$$

Balanced Neural Ratio Estimation (BNRE) [2] introduces a regularization term controlled by λ that improves calibration by enforcing consistency between joint and marginal predictions.

2.2 Hamiltonian Monte Carlo

We use the learned ratio as a differentiable surrogate likelihood within Hamiltonian Monte Carlo (HMC) [6, 7]. Gradients of the log-posterior enable efficient exploration of parameter space, allowing us to obtain posterior samples using the No-U-Turn Sampler (NUTS).

2.3 Calibration Diagnostics

We evaluate posterior validity using two diagnostics. TARP measures global coverage via empirical credibility curves [8], while Simulation-Based Calibration (SBC) evaluates marginal calibration using rank statistics [9, 10].

3 Pipeline and Workflow

The pipeline is implemented in JAX [11], using Flax [12] and NumPyro [13]. A standardized HDF5 dataset contract defines training, validation, and test splits of $(\boldsymbol{\theta}, \mathbf{x})$ pairs, ensuring consistent interfaces across tasks. Experiments are configured using Hydra [14], enabling reproducible specification of preprocessing, model architecture, and inference settings, and each run produces structured outputs including posterior samples and diagnostic summaries.

3.1 Agent-Assisted Workflow

The standardized structure enables integration with coding agents via a constrained workflow. Given a valid dataset, the agent validates the dataset, adapts a canonical experiment template, and executes training and inference. For example:

```
Follow AGENTS.md strictly.  Run experiment using the
dataset located in '<dataset_file>.h5', experiment name
'<experiment_name>', there is no missing data, use MinMax
scaling to [-1, 1] for both theta and x (fit on train only).
```

The purpose of the agent-assisted workflow is not to improve inference performance, but to accelerate adaptation of the pipeline to new datasets. The agent acts as an operational tool that validates datasets, adapts experiment templates, executes training and inference, and collects diagnostic outputs. Scientific interpretation remains the responsibility of the researcher, and all outputs are independently evaluated using SBC and TARP diagnostics.

We emphasize that agent use is intentionally operational rather than interpretive. The acceptance criterion is successful execution with valid outputs, including posterior samples, diagnostic summaries, and calibration plots, followed by human review. In this way, the agent functions as a constrained execution tool that accelerates experimentation while preserving reproducibility and researcher oversight.

4 Experiment

4.1 Lyα Forest Autocorrelation Function

We apply the pipeline to the Lyα forest autocorrelation function $\boldsymbol{\xi}_{\mathrm{Ly}\alpha}$, a summary statistic sensitive to the ionization structure of the intergalactic medium during the Epoch of Reionization [15]. We use simulations based on Nyx hydrodynamics [16, 17] combined with UV background modeling [18]. The model depends on two parameters: the mean free path λ_{mfp} and the mean transmitted flux $\langle F \rangle$. Each parameter pair generates multiple stochastic realizations due to observational noise and sightline sampling, naturally defining a simulator suitable for SBI. To avoid data leakage, dataset splits are performed at the parameter level.

5 Results

We perform inference on 500 mock observations using HMC with 4 chains and 2000 samples per chain. Figure 2 shows posterior distributions for three representative mock observations. In all three cases, the true parameters (blue lines) fall within the inferred credible regions, indicating successful parameter recovery. As expected, a degeneracy appears at large λ_{mfp}, where the autocorrelation function becomes progressively less sensitive to increases in the mean free path. This behavior is consistent with previous studies of this model [15] and demonstrates that the pipeline is able to recover known physical limitations of the forward model rather than obscuring them.

To assess calibration across the full test set, we perform Simulation-Based Calibration (SBC) and the Test of Accuracy with Random Points (TARP) using the resulting 500 posterior distributions.

Figure 3 summarizes these diagnostics. The SBC rank histograms are approximately uniform for both parameters, with KS p-values of 0.0509 and 0.0715, indicating no strong evidence against calibrated marginal posteriors. The TARP empirical coverage curve closely follows the ideal diagonal, supporting good global coverage behavior. Together, these diagnostics provide complementary evidence that the inferred posteriors are statistically consistent with the true parameters.

Importantly, these results are obtained through the same standardized workflow used throughout the pipeline. The posterior examples in Figure 2 demonstrate parameter recovery for individual observations, while the SBC and TARP diagnostics in Figure 3 provide quantitative validation across the full ensemble of inferred posteriors. Taken together, these results suggest that the BNRE+HMC pipeline produces reliable posterior estimates despite the known degeneracies of the Lyα forest model.

6 Conclusion

We presented a standardized pipeline for simulation-based inference that combines BNRE with HMC to produce calibrated posterior estimates. Our results on the Lyα forest autocorrelation function

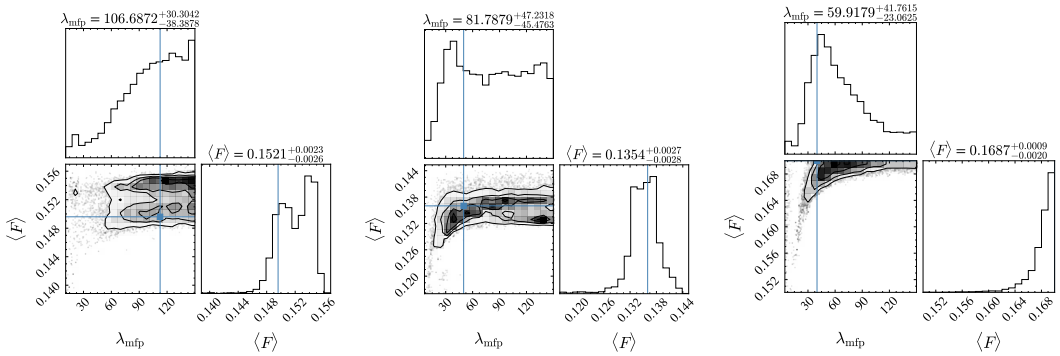

Figure 2: Posterior distributions for three mock observations. Blue lines indicate true parameters.

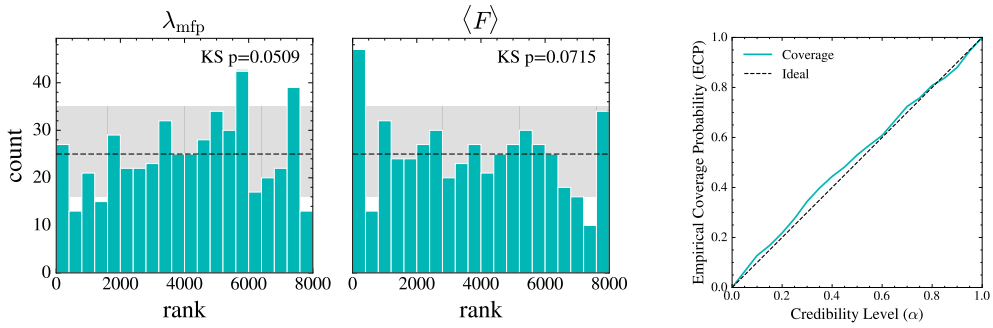

Figure 3: SBC rank histograms and TARP coverage curve.

demonstrate reliable inference in a realistic, degenerate setting, while integrated SBC and TARP diagnostics provide quantitative validation of posterior quality.

The structured workflow enables efficient and reproducible experimentation across problems, while the agent-assisted execution layer facilitates rapid adaptation to new datasets without altering the underlying inference methodology. By separating experiment execution from scientific interpretation and embedding calibration diagnostics directly into the workflow, the pipeline provides a practical framework for applying simulation-based inference in both manual and agent-assisted settings.

References

[1] 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.

[2] 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.

[3] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free MCMC with Amortized Approximate Ratio Estimators. *arXiv e-prints*, page arXiv:1903.04057, March 2019.

[4] Conor Durkan, Iain Murray, and George Papamakarios. On Contrastive Learning for Likelihood-free Inference. *arXiv e-prints*, page arXiv:2002.03712, February 2020.

[5] Benjamin Kurt Miller, Christoph Weniger, and Patrick Forré. Contrastive Neural Ratio Estimation for Simulation-based Inference. *arXiv e-prints*, page arXiv:2210.06170, October 2022.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

[12] Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. Flax: A neural network library and ecosystem for JAX, 2024.

[13] Du Phan, Neeraj Pradhan, Martin Jankowiak, Eli Bingham, Jonathan P. Chen, Fritz Obermeyer, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. NumPyro: Probabilistic programming with NumPy. Astrophysics Source Code Library, record ascl:2505.005, May 2025.

[14] Omry Yadan. Hydra - a framework for elegantly configuring complex applications. Github, 2019.

[15] 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. , 521(3):4056–4073, May 2023.

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

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

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