Simulation-Efficient Cosmological Inference with Multi-Fidelity SBI

Leander Thiele [1 2] **Adrian E. Bayer** [3 4] **Naoya Takeishi** [5]

Abstract

The simulation cost for cosmological simulation-based inference can be decreased by combining simulation sets of varying fidelity. We propose an approach to such multi-fidelity inference based on feature matching and knowledge distillation. Our method results in improved posterior quality, particularly for small simulation budgets and difficult inference problems.

1. Introduction

Cosmology is seeing an increase in attention for simulation-based inference (SBI) methods. Many proofs-of-concept have appeared, and even applications to real data are now possible (e.g., Hahn et al., 2024; Lemos et al., 2024; Gatti et al., 2024; Thiele et al., 2024; Novaes et al., 2025). Such applications do not, at the moment, enjoy the same degree of trust as do traditional likelihood-based analyses. To increase the trustworthiness of SBI in cosmology methodological developments are necessary, in particular in maximizing the posterior's quality with a realistic simulation budget.

SBI enables us to constrain model parameters in situations where the likelihood of an observed data vector is not tractable. In many such cases, we are able to simulate samples from the likelihood. Given a training set of such simulated samples, SBI (Cranmer et al., 2020) is an umbrella for algorithms that can learn approximations to the likelihood or functionally equivalent objects: e.g., posterior estimation (NPE, Greenberg et al., 2019; Lueckmann et al., 2017; Papamakarios & Murray, 2016), likelihood estimation (NLE, Papamakarios et al., 2019; Lueckmann et al., 2019), ratio estimation (NRE, Hermans et al., 2020), and quantile estimation (NQE, Jia, 2024b).

There is no free lunch in SBI. Posterior approximations can be dramatically wrong (Hermans et al., 2022). This is particularly pertinent in situations with a very limited simulation budget (Lueckmann et al., 2021; Homer et al., 2024; Bairagi et al., 2025). Cosmological simulations are extremely expensive if they are to capture all relevant physics accurately. For current and upcoming data sets, we will only be able to run very few simulations with a fidelity matching the data precision. This problem has been studied in a few recent works. By combining accurate analytic descriptions on large scales with cheaper small-scale simulations, the computational expense can be decreased (Modi & Philcox, 2023; Zhang et al., 2025). Sequential methods (e.g., Cole et al., 2022) concentrate the simulation budget in the most informative regions of parameter space.

In this work, we focus on the approach of multi-fidelity inference. In this setup, we assume a cascade of simulation sets with varying levels of fidelity, culminating in a small high-fidelity set matching the observation's precision. Prior work in this direction, specializing on the case of two fidelity levels, has proposed a calibration step in NQE (Jia, 2024a), and transfer learning via weight initialization (Krouglova

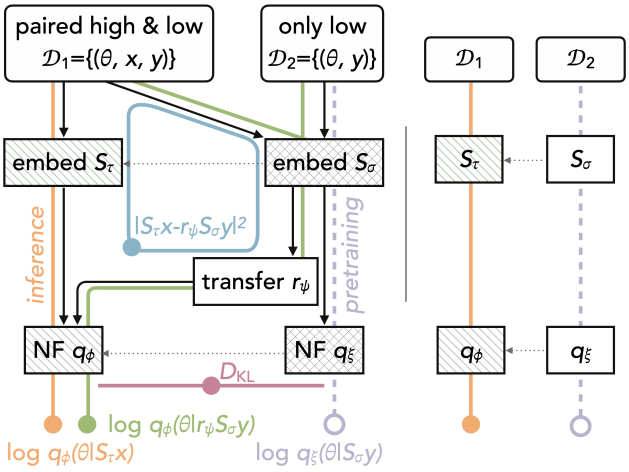

Figure 1. Graphical summary of the proposed multi-fidelity architecture and training loss. Filled circles indicate the components of our proposed training loss. The notation is given in Section 3, particularly Eq. (5). Hatched components are weight-initialized (dotted arrows) from pretraining on $\mathcal{D}_2$, the cross-hatched ones being frozen. The final model only evaluates the path indicated in orange. The right side shows the weight-initialization scheme we compare to.

[1]Kavli IPMU (WPI), UTIAS, The University of Tokyo, Japan [2]Center for Data-Driven Discovery, Kavli IPMU (WPI), Japan [3]Center for Computational Astrophysics, Flatiron Institute, New York, USA [4]Department of Astrophysical Sciences, Princeton University, USA [5]RCAST, The University of Tokyo, Japan. Correspondence to: Leander Thiele <leander.thiele@ipmu.jp>.

ML4Astro 2025, Vancouver, CA.

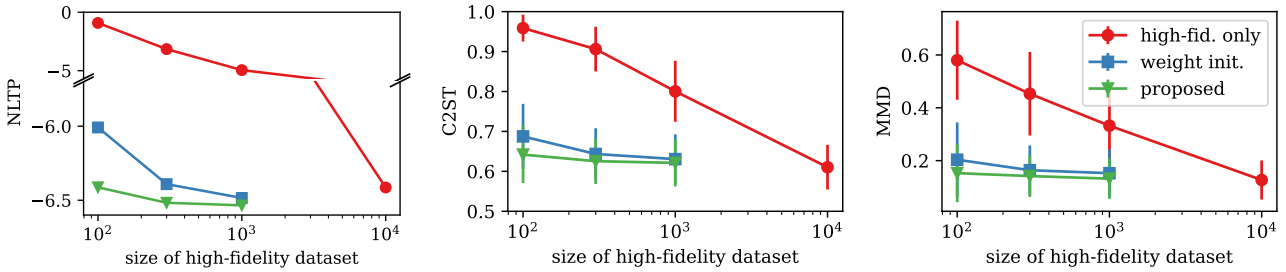

Figure 2. Performance evaluation for the cosmological inference problem. We show, from left to right, test loss, two-sample classifier accuracy, and maximum mean discrepancy. Lower is better for all.

et al., 2025; Saoulis et al., 2025). For an arbitrary number of fidelity levels, the loss function can be stabilized via control variates (Hikida et al., 2025). These existing approaches highlight the potential of multi-fidelity SBI to reduce the required simulation budget.

Our contribution employs a tailored training setup which enables training on multiple fidelity simulations simultaneously. The training constructs stochastic mappings between the embedded data vectors at different fidelity levels and a latent space corresponding to the highest fidelity. Figure 1 illustrates our architecture and training process, a derivation is presented in Section 3. Our approach is a superset of weight initialization. We demonstrate that it outperforms weight initialization in examples where the two can be compared. Furthermore, our approach can accommodate any number of fidelity levels and is applicable even in situations when the observations or embeddings at different fidelities differ in shape. Empirically, it is also found to converge faster than weight initialization.

In this work, we focus on the case of NPE. The extension to NRE appears natural but will be deferred to future work. We consider a cosmology example with a traditional summary statistic, though the extension to field-level inference, as considered in Saoulis et al. (2025), is straightforward.

2. Results

We evaluate performance by considering the example of the matter power spectrum at redshift $z = 0$ in 5-parameter ΛCDM. The high-fidelity data is taken from the BSQ set of the Quijote simulation suite (Villaescusa-Navarro et al., 2020), which uses a tree-PM algorithm for accurate small scale clustering. The low-fidelity data was produced using FastPM (Feng et al., 2016), a PM code which is cheaper by a factor of $\sim$10–100 per simulation, using the same prior. The FastPM simulations paired with Quijote ($\mathcal{D}_1$) are matched in parameters and seeds.[1] We truncate at $k_{\text{max}} = 0.5\,h\text{Mpc}^{-1}$

[1]Matching seeds is, in principle, not required as the transfer network r_ψ can absorb the additional stochasticity. However, performance of the proposed method benefits from matching the seeds.

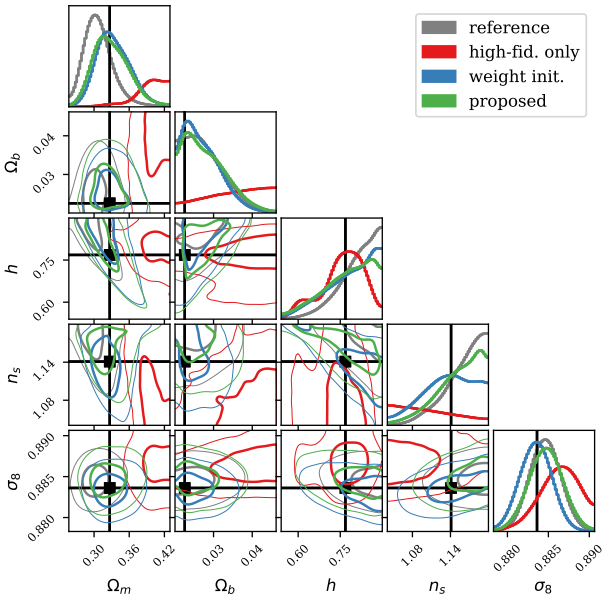

Figure 3. Posteriors on an example realization from the test set. We compare models trained on 100 high-fidelity simulations to the reference posterior (grey) trained on 28k high-fidelity simulations.

to dimensionality 79.

We take MLPs for the embedding and transfer networks, and a spline flow (Durkan et al., 2019) from `sbi` (Tejero-Cantero et al., 2020; Durkan et al., 2020) for the density estimator. Hyperparameters are optimized with `optuna` (Akiba et al., 2019).

For all multi-fidelity runs we use 10k low-fidelity samples with 100, 300, or 1k high-fidelity samples. We evaluate performance using a held-out test set of 2k high-fidelity simulations, and compare to reference posteriors computed by training standard NPE on 28k high-fidelity samples.

Figure 2 shows three different evaluation metrics as a function of the high-fidelity set size. NLTP is the test loss, C2ST is the two-sample classifier accuracy (Lopez-Paz &

Oquab, 2018), and MMD is the kernel-based maximum mean discrepancy (Gretton et al., 2012). Our approach consistently outperforms weight initialization, with most benefit for small high-fidelity sets. As an example, in Figure 3 we show a comparison of posteriors. Our method results in posteriors quite close to those obtained with weight initialization, while the estimation only with small high-fidelity data is far from the reference.

3. Methods

3.1. Preliminaries

NPE is a problem to estimate the conditional distribution $p(\theta \mid x)$ given realizations of simulator's parameters θ and outcomes x as training data. In the standard NPE, we learn a density estimator $q_\phi(\theta \mid x)$ parametrized by ϕ, typically a conditional normalizing flow, by minimizing

$$\mathbb{E}_{p(x)} \mathrm{KL}\left(p(\theta \mid x) \,||\, q_\phi(\theta \mid x)\right) = -\mathbb{E}_{p(\theta,x)} \log q_\phi(\theta \mid x) - \mathbb{E}_{p(x)} H(p(\theta \mid x)).$$

Let $\mathcal{D} = \{(\theta_1, x_1), \dots, (\theta_n, x_n)\}$ be data drawn from $p(x \mid \theta)p(\theta)$. As the entropy $H(p)$ is constant wrt ϕ, we solve

$$\underset{\phi}{\text{minimize}} \quad \frac{1}{n} \sum_{(\theta,x)\in\mathcal{D}} -\log q_\phi(\theta \mid x). \tag{1}$$

3.2. Proposed Method

Let $p(x \mid \theta)$ and $p(y \mid \theta)$ be the likelihoods of two simulators to model the same target with different fidelity: $x \in \mathcal{X}$ is with relatively high fidelity, and $y \in \mathcal{Y}$ is with lower fidelity. Evaluating the values of $p(x \mid \theta)$ and $p(y \mid \theta)$ is intractable, and we only have samples drawn from them. The high-fidelity simulator is much heavier to run than the low-fidelity simulator, so we typically have $\mathcal{D}_1 = \{(\theta_i, x_i, y_i) \mid i = 1, \dots, n_1\}$ and $\mathcal{D}_2 = \{(\theta_i, y_i) \mid i = 1, \dots, n_2\}$, where $n_1 < n_2$ by a large margin.

Our goal is to estimate the conditional distribution of $\theta \mid x$. However, the standard NPE loss in Equation (1) for x can be computed only on the θ-x pairs in small $\mathcal{D}_1$. We thus want to utilize the information from the θ-y pairs, available in larger $\mathcal{D}_2$ as well, for learning the θ-x relation. To this end, we introduce additional loss functions as presented in Sections 3.2.1 and 3.2.2. For the sake of simplicity, we here only discuss the case of two levels of fidelity, but extending the method to more fidelity levels is straightforward.

3.2.1. Feature matching

We aim to minimize the y-based posterior KL:

$$\mathbb{E}_{p(y)} \mathrm{KL}\left(p(\theta \mid y) \,||\, q_\phi(\theta \mid y)\right) = -\mathbb{E}_{p(\theta,y)} \log q_\phi(\theta \mid y) - \mathbb{E}_{p(y)} H(p(\theta \mid y)).$$

Recall that the network of the density estimator, q_ϕ, is designed to receive x (and not y) as the condition, so we cannot compute $-\log q_\phi(\theta \mid y)$ directly. It can be upper bounded by Jensen's inequality as

$$\begin{aligned} -\log q_\phi(\theta \mid y) &= -\log \int r_\psi(x \mid y) \frac{q_\phi(\theta \mid x)p(x \mid y)}{r_\psi(x \mid y)} dx \\ &\le -\int r_\psi(x \mid y) \log \frac{q_\phi(\theta \mid x)p(x \mid y)}{r_\psi(x \mid y)} dx \\ &= -\mathbb{E}_{r_\psi(x|y)} \log q_\phi(\theta \mid x) + \mathrm{KL}\left(r_\psi(x \mid y) \,||\, p(x \mid y)\right), \end{aligned}$$

where $r_\psi(x \mid y)$ is a conditional distribution of x given y parametrized with ψ. We thus should minimize

$$-\mathbb{E}_{p(\theta,y)}\mathbb{E}_{r_\psi(x|y)} \log q_\phi(\theta \mid x) \quad \text{and} \tag{2}$$

$$\mathbb{E}_{p(y)} \mathrm{KL}\left(r_\psi(x \mid y) \,||\, p(x \mid y)\right). \tag{3}$$

The newly introduced model, $r_\psi(x \mid y)$, probabilistically transforms y to x.[2] Equation (2) reads as the NPE loss computed on the transformed condition $x \sim r_\psi(x \mid y)$. Equation (3) enforces $r_\psi(x \mid y)$ close to the true transformation, but computing the KL is challenging because we cannot evaluate $p(x \mid y)$ and only have access to the samples from $p(x, y)$. As a surrogate of (3), we suggest minimizing the KL in the opposite direction:

$$\mathbb{E}_{p(y)} \mathrm{KL}\left(p(x \mid y) \,||\, r_\psi(x \mid y)\right) = -\mathbb{E}_{p(x,y)} \log r_\psi(x \mid y) + \text{const.} \tag{3'}$$

3.2.2. Response distillation

Let $\tilde{q}_\xi(\theta \mid y)$ be the density estimator of $p(\theta \mid y)$ trained on the large dataset $\mathcal{D}_2$ with the standard NPE loss, i.e.,

$$\xi = \arg\min_{\xi'} \frac{1}{n_2} \sum_{(\theta,y)\in\mathcal{D}_2} -\log \tilde{q}_{\xi'}(\theta \mid y).$$

We aim to inform our estimator, $q_\phi(\theta \mid x)$, from $\tilde{q}_\xi(\theta \mid y)$ based on knowledge distillation (Hinton et al., 2015). We minimize the (expectation of) KL between the two:

$$\begin{aligned} &\mathbb{E}_{p(y)} \mathrm{KL}(\tilde{q}_\xi(\theta \mid y) \,||\, q_\phi(\theta \mid y)) \\ &= -\mathbb{E}_{p(y)}\mathbb{E}_{\tilde{q}_\xi(\theta|y)} \log q_\phi(\theta \mid y) - \mathbb{E}_{p(y)} H(\tilde{q}_\xi(\theta \mid y)). \end{aligned}$$

Since ξ is fixed already, the last term is constant. As $-\log q_\phi(\theta \mid y)$ can be upper bounded as discussed earlier, the quantity to minimize is

$$\begin{aligned} &-\mathbb{E}_{p(y)}\mathbb{E}_{\tilde{q}_\xi(\theta|y)} \log q_\phi(\theta \mid y) \\ &\le -\mathbb{E}_{p(y)}\mathbb{E}_{\tilde{q}_\xi(\theta|y)}\mathbb{E}_{r_\psi(x|y)} \log q_\phi(\theta \mid x) \\ &\quad + \underbrace{\mathbb{E}_{p(y)} \mathrm{KL}(r_\psi(x \mid y) \,||\, p(x \mid y))}_{\text{same as (3)}}. \end{aligned}$$

[2] We will discuss in Section 3.2.3 more practical cases where embedding nets are applied to x and y.

Note that the KL in the last term already appeared in Equation (3). Finally we are interested in minimizing

$$-\mathbb{E}_{p(y)}\mathbb{E}_{\tilde{q}_\xi(\theta|y)}\mathbb{E}_{r_\psi(x|y)} \log q_\phi(\theta \mid x), \tag{4}$$

which is equivalent to minimizing the KL divergence between $\tilde{q}_\xi(\theta \mid y)$ and $q_\phi(\theta \mid x)$.

3.2.3. Implementation

We often use so-called embedding networks to extract summary statistics from observations, which were omitted in the discussion so far for notational simplicity. Let $S_\tau : x \mapsto S_\tau(x)$ and $S_\sigma : y \mapsto S_\sigma(y)$ be the embedding networks for x and y, parametrized with τ and σ, respectively. The embedding networks are taken as multi-layer perceptrons reducing the dimensionality in order to simplify the learning problem for the normalizing flow. With these embedding nets taken into consideration, the proposed method proceeds as follows.

Step 1. Train an NPE model on $\mathcal{D}_2$ ordinarily by

$$\underset{\xi,\sigma}{\text{minimize}} \quad \frac{1}{n_2} \sum_{(\theta,y)\in\mathcal{D}_2} -\log \tilde{q}_\xi(\theta \mid S_\sigma(y)).$$

Step 2. Train an NPE model $q_\phi(\theta \mid S_\tau(x))$ with the proposed loss. First, if possible, initialize ϕ with ξ. We then solve the following problem:

$$\underset{\phi,\tau,\psi}{\text{minimize}} \quad L_1 + \alpha L_2 + \beta L_3, \tag{5}$$

where each loss term is given as[3]

$$L_1 = \frac{1}{n_1} \sum_{(\theta,x)\in\mathcal{D}_1} -\log q_\phi(\theta \mid S_\tau(x)),$$

$$L_2 = \frac{1}{n_{1,2}} \sum_{(\theta,y)\in\mathcal{D}_{1,2}} \mathbb{E}_{u\sim r_\psi(\cdot|S_\sigma(y))} -\log q_\phi(\theta \mid u) + \frac{1}{n_1} \sum_{(x,y)\in\mathcal{D}_1} -\log r_\psi(S_\tau(x) \mid S_\sigma(y)),$$

$$L_3 = \frac{1}{n_{1,2}} \sum_{y\in\mathcal{D}_{1,2}} \mathbb{E}_{\theta\sim\tilde{q}_\xi(\theta|S_\sigma(y)),\ u\sim r_\psi(\cdot|S_\sigma(y))} -\log q_\phi(\theta \mid u).$$

Here, $\mathcal{D}_{1,2} = \mathcal{D}_1 \cup \mathcal{D}_2$ and $n_{1,2} = n_1 + n_2$. The expectations can be approximated with samples. We empirically found that instead of rigorously computing $-\log r_\psi(S_\tau(x) \mid S_\sigma(y))$, merely minimizing the squared error $\mathbb{E}_v\|S_\tau(x) - v\|^2$ where $v \sim r_\psi(\cdot \mid S_\sigma(y))$ resulted in more stable learning. We note that further analysis of this heuristic may lead to an interesting discussion. We take the relative weights of the components in Eq. (5) as constant throughout Step 2.

[3] In the notation of Fig. 1, line by line the loss terms are orange, green, blue, and magenta.

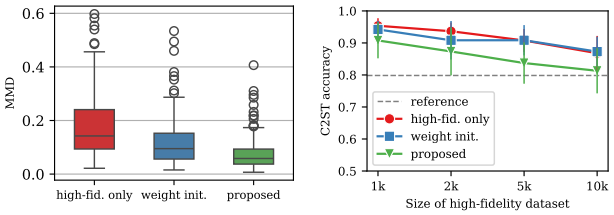

Figure 4. (*Left*) MMD between the true and learned posteriors for GAUSSIAN. (*Right*) C2ST metrics for SLCP. The reference performance is by NPE on a large high-fid. set with $n_1 = 100$k.

4. Some More Numerical Examples

For controlled evaluation, we applied the proposed method to two synthetic tasks: GAUSSIAN and SLCP.

In GAUSSIAN, θ is drawn from the standard normal distribution, and x and y are given as the affine transformation of θ plus Gaussian noise. y is supposed to have "low-fidelity" by setting $\dim y < \dim x$ and being added stronger noise than x is. We used $\dim\theta = 10$, $\dim x = 10$, and $\dim y = 5$. We prepared a high-fidelity dataset of size $n_1 = 100$ and a low-fidelity dataset of size $n_2 = 5$k.

SLCP is a standard task in SBI benchmarking (see, e.g., Lueckmann et al., 2021), featuring multimodal posteriors. Originally the task is to infer 5 parameters of a 2D Gaussian distribution given 4 samples drawn from it. We made a "low-fidelity" version by taking only 3 samples instead. So for SLCP, $\dim\theta = 5$, $\dim x = 8$, and $\dim y = 6$. We tried different sizes of high-fidelity dataset, $n_1 = 1$k, 2k, 5k, or 10k, and used the fixed size of low-fidelity dataset with $n_2 = 100$k. To provide a reference performance, we also run an ordinary NPE on a large high-fidelity dataset of size $n_1 = 100$k.

The results are shown in Figure 4. We compare the proposed method with two baselines: models trained only on the θ-x pairs in $\mathcal{D}_1$ (high-fid. only) and models pretrained on $\mathcal{D}_2$ and then finetuned on $\mathcal{D}_1$ (weight init.). For the GAUSSIAN task, the MMD between the true and the learned posterior distributions are reported in the left panel. For SLCP, the metrics of C2ST are reported in the right panel.[4] Both sets of results show the efficacy of the proposed method. It achieves better performance in both tasks than the weight initialization.

5. Conclusions

We have presented a new method for multi-fidelity SBI. Our method contains transfer learning via weight initialization as

[4] For GAUSSIAN, MMD should be reliable because the posteriors are unimodal. For SLCP, MMD may be unreliable as the posteriors are multimodal, thus we report C2ST metrics.

a special case but extends it in several ways. Additional loss terms contribute information from lower-fidelity samples during the fine-tuning phase. Our setup allows for training with an arbitrary number of fidelity levels, as well as shape mismatch in the data summaries or embeddings.

In our evaluation on a cosmology example problem we find moderate improvement from our method over weight initialization. The improvement is more pronounced in examples with complicated posterior shape (such as SLCP). Our proposed method contains several heuristics which could enable improvement in future work.

Acknowledgements

We thank S. Wagner-Carena, O. Philcox, and C. Modi for discussions and comments on this draft. We thank the anonymous reviewers for the ICML-colocated ML4Astro 2025 workshop for their constructive feedback. This research used computing resources at Kavli IPMU and the Flatiron Institute. The Kavli IPMU is supported by the WPI (World Premier International Research Center) Initiative of the MEXT (Japanese Ministry of Education, Culture, Sports, Science and Technology). Leander Thiele was supported by JSPS KAKENHI Grant 24K22878. Adrian Bayer was supported by the Simons Foundation. Naoya Takeishi was supported by JSPS KAKENHI Grant Numbers JP20K19869 and JP25H01454 and JSPS International Joint Research Program JPJSJRP20221501.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

Bairagi, A., Wandelt, B., and Villaescusa-Navarro, F. How many simulations do we need for simulation-based inference in cosmology?, 2025. arXiv:2503.13755.

Cole, A., Miller, B. K., Witte, S. J., Cai, M. X., Grootes, M. W., Nattino, F., and Weniger, C. Fast and credible likelihood-free cosmology with truncated marginal neural ratio estimation. *Journal of Cosmology and Astroparticle Physics*, 2022(09):004, 2022.

Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Neural spline flows. In *Advances in Neural Information Processing Systems*, volume 32, pp. 7511–7522, 2019.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. nflows: normalizing flows in PyTorch, 2020. doi: 10.5281/zenodo.4296287.

Feng, Y., Chu, M.-Y., Seljak, U., and McDonald, P. Fastpm: a new scheme for fast simulations of dark matter and haloes. *Monthly Notices of the Royal Astronomical Society*, 463(3):2273–2286, 2016.

Gatti, M., Jeffrey, N., Whiteway, L., Williamson, J., Jain, B., Ajani, V., Anbajagane, D., Giannini, G., Zhou, C., Porredon, A., Prat, J., Yamamoto, M., Blazek, J., Kacprzak, T., Samuroff, S., Alarcon, A., Amon, A., Bechtol, K., Becker, M., Bernstein, G., Campos, A., Chang, C., Chen, R., Choi, A., Davis, C., Derose, J., Diehl, H. T., Dodelson, S., Doux, C., Eckert, K., Elvin-Poole, J., Everett, S., Ferte, A., Gruen, D., Gruendl, R., Harrison, I., Hartley, W. G., Herner, K., Huff, E. M., Jarvis, M., Kuropatkin, N., Leget, P. F., MacCrann, N., McCullough, J., Myles, J., Navarro-Alsina, A., Pandey, S., Raveri, M., Rollins, R. P., Roodman, A., Sanchez, C., Secco, L. F., Sevilla-Noarbe, I., Sheldon, E., Shin, T., Troxel, M., Tutusaus, I., Varga, T. N., Yanny, B., Yin, B., Zhang, Y., Zuntz, J., Aguena, M., Alves, O., Annis, J., Brooks, D., Carretero, J., Castander, F. J., Cawthon, R., Costanzi, M., da Costa, L. N., Pereira, M. E. S., Evrard, A. E., Flaugher, B., Fosalba, P., Frieman, J., García-Bellido, J., Gerdes, D. W., Gruen, D., Gruendl, R. A., Gschwend, J., Gutierrez, G., Hollowood, D. L., Honscheid, K., James, D. J., Kuehn, K., Lahav, O., Lee, S., Marshall, J. L., Mena-Fernández, J., Menanteau, F., Miquel, R., Ogando, R. L. C., Pereira, M. E. S., Pieres, A., Plazas Malagón, A. A., Sanchez, E., Smith, M., Suchyta, E., Swanson, M. E. C., Tarle, G., Weaverdyck, N., Weller, J., Wiseman, P., and DES Collaboration. Dark Energy Survey Year 3 results: Simulation-based cosmological inference with wavelet harmonics, scattering transforms, and moments of weak lensing mass maps. Validation on simulations. *Phys. Rev. D*, 109(6): 063534, 2024.

Greenberg, D., Nonnenmacher, M., and Macke, J. Automatic posterior transformation for likelihood-free inference. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2404–2414, 2019.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Hahn, C., Lemos, P., Parker, L., Régaldo-Saint Blancard, B., Eickenberg, M., Ho, S., Hou, J., Massara, E., Modi, C., Moradinezhad Dizgah, A., and Spergel, D. Cosmological constraints from non-Gaussian and nonlinear galaxy clustering using the SIMBIG inference framework. *Nature Astronomy*, 8:1457–1467, 2024.

Hermans, J., Begy, V., and Louppe, G. Likelihood-free MCMC with amortized approximate ratio estimators. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 4239–4248, 2020.

Hermans, J., Delaunoy, A., Rozet, F., Wehenkel, A., Begy, V., and Louppe, G. A trust crisis in simulation-based inference? Beware, your posterior approximations can be unfaithful. *Transactions on Machine Learning Research*, 2022.

Hikida, Y., Bharti, A., Jeffrey, N., and Briol, F.-X. Multilevel neural simulation-based inference, 2025. URL https://arxiv.org/abs/2506.06087.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network, 2015. arXiv:1503.02531.

Homer, J., Friedrich, O., and Gruen, D. Simulation-based inference has its own Dodelson-Schneider effect (but it knows that it does), 2024. arXiv:2412.02311.

Jia, H. Cosmological analysis with calibrated neural quantile estimation and approximate simulators, 2024a. arXiv:2411.14748.

Jia, H. Simulation-based inference with quantile regression. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 21731–21752, 2024b.

Krouglova, A. N., Johnson, H. R., Confavreux, B., Deistler, M., and Gonçalves, P. J. Multifidelity simulation-based inference for computationally expensive simulators, 2025. arXiv:2502.08416.

Lemos, P., Parker, L., Hahn, C., Ho, S., Eickenberg, M., Hou, J., Massara, E., Modi, C., Dizgah, A. M., Blancard, B. R.-S., and Spergel, D. Field-level simulation-based inference of galaxy clustering with convolutional neural networks. *Phys. Rev. D*, 109:083536, 2024.

Lopez-Paz, D. and Oquab, M. Revisiting classifier two-sample tests, 2018. arXiv:1610.06545.

Lueckmann, J.-M., Goncalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., and Macke, J. H. Flexible statistical inference for mechanistic models of neural dynamics. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1289–1299, 2017.

Lueckmann, J.-M., Bassetto, G., Karaletsos, T., and Macke, J. H. Likelihood-free inference with emulator networks. In *Proceedings of the 1st Symposium on Advances in Approximate Bayesian Inference*, pp. 32–53, 2019.

Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., and Macke, J. Benchmarking simulation-based inference. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pp. 343–351, 2021.

Modi, C. and Philcox, O. H. E. Hybrid SBI or how I learned to stop worrying and learn the likelihood, 2023. arXiv:2309.10270.

Novaes, C. P., Thiele, L., Armijo, J., Cheng, S., Cowell, J. A., Marques, G. A., Ferreira, E. G. M., Shirasaki, M., Osato, K., and Liu, J. Cosmology from HSC Y1 weak lensing data with combined higher-order statistics and simulation-based inference. *Phys. Rev. D*, 111(8):083510, 2025.

Papamakarios, G. and Murray, I. Fast ϵ-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, volume 29, pp. 1036–1044, 2016.

Papamakarios, G., Sterratt, D., and Murray, I. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 837–848, 2019.

Saoulis, A. A., Piras, D., Jeffrey, N., Mancini, A. S., Ferreira, A. M. G., and Joachimi, B. Transfer learning for multifidelity simulation-based inference in cosmology, 2025. arXiv:2505.21215.

Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Gonçalves, P. J., Greenberg, D. S., and Macke, J. H. sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, 2020.

Thiele, L., Massara, E., Pisani, A., Hahn, C., Spergel, D. N., Ho, S., and Wandelt, B. Neutrino Mass Constraint from an Implicit Likelihood Analysis of BOSS Voids. *Astrophys. J.*, 969(2):89, 2024.

Villaescusa-Navarro, F., Hahn, C., Massara, E., Banerjee, A., Delgado, A. M., Ramanah, D. K., Charnock, T., Giusarma, E., Li, Y., Allys, E., Brochard, A., Uhlemann, C., Chiang, C.-T., He, S., Pisani, A., Obuljen, A., Feng, Y., Castorina, E., Contardo, G., Kreisch, C. D., Nicola, A., Alsing, J., Scoccimarro, R., Verde, L., Viel, M., Ho, S., Mallat, S., Wandelt, B., and Spergel, D. N. The quijote simulations. *The Astrophysical Journal Supplement Series*, 250(1):2, 2020.

Zhang, G., Modi, C., and Philcox, O. H. E. Modeling galaxy surveys with hybrid SBI, 2025. arXiv:2505.13591.