

# Fast and Credible Likelihood-Free Cosmology with Truncated Marginal Neural Ratio Estimation

**Alex Cole** <sup>a,\*</sup> **Benjamin K. Miller**<sup>a</sup> **Samuel J. Witte**<sup>a</sup> **Maxwell X. Cai**<sup>b,c</sup> **Meiert W. Grootes**<sup>d</sup> **Francesco Nattino**<sup>d</sup> **Christoph Weniger**<sup>a,\*</sup>

<sup>a</sup>Gravitation Astroparticle Physics Amsterdam (GRAPPA), Institute for Theoretical Physics Amsterdam and Delta Institute for Theoretical Physics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

<sup>b</sup>SURF Cooperative, Science Park 140, 1098 XG Amsterdam

<sup>c</sup>Leiden Observatory, Leiden University, Niels Bohrweg 2, 2300 RA Leiden, The Netherlands

<sup>d</sup>Netherlands eScience Center, Science Park 140, 1098 XG Amsterdam, The Netherlands

E-mail: [a.e.cole2@uva.nl](mailto:a.e.cole2@uva.nl), [c.weniger@uva.nl](mailto:c.weniger@uva.nl)

**Abstract.** Sampling-based inference techniques are central to modern cosmological data analysis; these methods, however, scale poorly with dimensionality and typically require approximate or intractable likelihoods. In this paper we describe how Truncated Marginal Neural Ratio Estimation (TMNRE) (a new approach in so-called simulation-based inference) naturally evades these issues, improving the (*i*) efficiency, (*ii*) scalability, and (*iii*) trustworthiness of the inference. Using measurements of the Cosmic Microwave Background (CMB), we show that TMNRE can achieve converged posteriors using orders of magnitude fewer simulator calls than conventional Markov Chain Monte Carlo (MCMC) methods. Remarkably, in these examples the required number of samples is effectively independent of the number of nuisance parameters. In addition, a property called *local amortization* allows the performance of rigorous statistical consistency checks that are not accessible to sampling-based methods. TMNRE promises to become a powerful tool for cosmological data analysis, particularly in the context of extended cosmologies, where the timescale required for conventional sampling-based inference methods to converge can greatly exceed that of simple cosmological models such as  $\Lambda$ CDM. To perform these computations, we use an implementation of TMNRE via the open-source code `swyft`.<sup>†</sup>

---

\*Corresponding author.

<sup>†</sup>`swyft` is available at <https://github.com/undark-lab/swyft>. Demonstration on cosmological simulators used in this paper is available at <https://github.com/a-e-cole/swyft-CMB>.

---

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Simulation-based inference methodology</b>	<b>3</b>
2.1	Truncated Marginal Neural Ratio Estimation	4
2.2	“Cutting to the chase” by estimating marginal posteriors	9
2.3	“Getting it right” through coverage tests for simulation-based inference	11
2.4	Zero-cost inference through simulation re-use	13
2.5	Comparison with other methods	14
<b>3</b>	<b>Cosmological applications</b>	<b>16</b>
3.1	A CMB forecasting simulator	17
3.1.1	Inference setup and results	19
3.2	Comparing experimental configurations	20
3.3	Realistic Planck simulator	22
3.4	Neutrino masses	24
3.5	Testing statistical consistency	26
3.6	Wide priors	26
<b>4</b>	<b>Conclusions and outlook</b>	<b>28</b>
<b>A</b>	<b>Explicit sky mask</b>	<b>30</b>

---

## 1 Introduction

An impending flood of high-quality data from cosmological surveys promises to revolutionize our understanding of the universe (see *e.g.* [1] for a review of upcoming surveys). The collection of this data alone, however, is not sufficient to perform fundamental science; in order to extract cosmological information one must also “solve the inverse problem,” *i.e.* given a set of observations, one must deduce the set of cosmological models that may have produced them.

Inferring model parameters from complex observational data is one of the grand challenges of modern cosmology. Part of the difficulty arises from the fact that classical methods rely on evaluating the likelihood of data given model parameters. Often only an approximation of this likelihood, representing a fraction of the data’s information content (*e.g.* rather than the entire spatial distribution of galaxies, only the power spectrum of this distribution at large scales), is theoretically understood. In addition, these approaches require sampling the *full joint posterior*, so that time required to achieve convergence scales poorly with the dimensionality of the parameter space. This latter point is particularly troublesome for problems with large numbers of nuisance parameters, and arises independently of whether the likelihood is known.

Novel approaches in the field of simulation-based inference (SBI)<sup>1</sup> are starting to overcome these obstacles (see [5] for a recent review). SBI exploits the fact that data generated by a stochastic simulator can be regarded as sampling an exact (implicit) likelihood, even

---

<sup>1</sup>SBI is closely connected to Approximate Bayesian Computation [2, 3], which refers to a class of computational methods used to estimate posterior distributions, and has been developed since the 1980s [4].

when this function is not explicitly known. The field has recently seen accelerated progress thanks to deep learning methods [5–9], and has been applied to a wide array of problems in cosmology and astrophysics, from cosmological density fields to gravitational waves [10–15]. In order to improve the scalability with the dimensionality of the parameter space, it is important to note that in many cases the full joint posterior is overly informative. Typically, one is only interested in a subset of 1- and 2-dimensional marginal posteriors; for example, in cosmology, one may be interested in the cosmological parameters but not the large associated list of nuisance parameters parameterizing *e.g.* foregrounds and instrumental systematics. By directly targeting low-dimensional posteriors we can alleviate sampling problems associated with high dimensionality, and thus an *arbitrarily* large number of nuisance parameters can be included without additional computational overhead.<sup>2</sup>

In this paper, we describe and apply a new algorithm in simulation-based inference called *Truncated Marginal Neural Ratio Estimation* (TMNRE) [16, 17]. TMNRE uses neural networks to directly learn *marginal* likelihood-to-evidence ratios in a sequence of rounds. After each round, evaluations of the network are used to truncate the prior, so that the parameter region of interest is targeted. In other words, simulations are evaluated in regions of parameter space where the posterior for a given observation is expected to have support. As presented in [16, 17], TMNRE has a number of major advantages over classical sampling-based inference techniques (as well as some other SBI algorithms); these advantages can be broadly categorized as

1. **Efficiency:** Learning 1- and 2-dimensional marginal posteriors is far simpler than trying to sample the full joint posterior. As such, TMNRE naturally outperforms conventional sampling methods like MCMC, typically requiring orders of magnitude fewer simulator calls. In addition, one can store and reuse simulations should one wish to change the priors, the observations, or the inference network. For example, we show in Sec. 3.2 that after having trained a network to infer the posteriors of cosmological parameters using the CMB, one can include BAOs (baryonic acoustic oscillations) to learn corresponding posteriors at *no additional simulation cost*.
2. **Scalability:** As previously mentioned, sampling-based inference approaches scale poorly with the dimensionality of the parameter space. In most problems of interest, the dimensionality can be greatly augmented due to a large list of nuisance parameters, the posteriors of which are typically not of interest. TMNRE offers the possibility of learning the posteriors only for the parameter subset of interest, and thus is easily scalable to high dimensional problems. We demonstrate the excellent scaling behavior of TMNRE on a “realistic” Planck simulator in Sec. 3.3.
3. **Testability:** Novel approaches to inference, in particular those rooted in machine learning, are often met with a level of skepticism. While in some cases caution is justified, the reliability of many SBI algorithms (including TMNRE) can be explicitly tested, which is *not* easily possible for likelihood-based methods. In the context of TMNRE, thanks to so-called *local amortization* (see Sec. 2.3), the trained network has learned not only the marginal posteriors for our observed data, but also the marginal posteriors for *any* data drawn from the (truncated) prior. This enables us to efficiently inspect the statistical

---

<sup>2</sup>Note that if all parameters contribute equal variance to the data, the implicit data distribution becomes exceedingly noisy. When we refer to scaling to arbitrary number of parameters, the data variance is implicitly held fixed. This limit remains a challenge for sampling-based methods, but is tractable in our proposal.

rigor of our results, and assess fundamental statistical properties like expected coverage [18]. These sorts of tests are not feasible for results generated via sampling-based methods, since for each new observation a new chain is needed.

In this paper we use TMNRE to accelerate cosmological inference using simulators for the CMB power spectra. As first glance this might appear a somewhat peculiar application – in this case the likelihood is known, and obtaining convergent MCMC results using a  $\Lambda$ CDM cosmology is not computationally difficult. We have chosen to nevertheless apply TMNRE in this context for three important reasons: (*i*) the CMB serves as the gold standard for cosmological inference, and novel statistical tools need to be proven as robust in this context before being applied elsewhere, (*ii*) SBI trivially includes explicit likelihoods, strong advantages even in these cases, and (*iii*) because obtaining convergent MCMC results in *extended* cosmological scenarios *can* be computationally prohibitive. This latter point is particularly important, as it tends to cause cosmologists to introduce unnecessary simplifications in their cosmological simulators (*e.g.* by assuming massless neutrinos, simplified phase space distributions, or approximate collision terms in the Boltzmann equations). The current precision of the CMB is already sufficient in many cases to be sensitive to such approximation schemes, and thus TMNRE offers a practical tool for cosmologists to robustly assess the validity of their favorite models without compromise.

The structure of this paper is as follows. In Sec. 2 we briefly review SBI focusing specifically on TMNRE and comparing to the alternative deep learning-based approach of DELFI [19, 20]. In Sec. 3 we apply this framework to different cosmological examples, using the open-source implementation of TMNRE via `swyft`.<sup>3</sup> Each example demonstrates a unique attractive feature of our approach (or occasionally SBI in general). In Sec. 3.1 we explain at length how CMB likelihoods may be reformulated as stochastic simulators (defining a “CMB forecasting simulator”), showing that TMNRE requires orders of magnitude fewer simulations than MCMC to infer  $\Lambda$ CDM parameter constraints. In Sec. 3.2 we show how reusing simulations allows us to perform inferences for new combinations of experiments with *zero* additional calls to the simulator. In Sec. 3.3 we infer cosmological constraints for a simulator corresponding to the Planck HiLLiPoP likelihood, demonstrating that the inclusion of 13 nuisance parameters does not increase the simulator cost of TMNRE, while the cost of MCMC increases by several orders of magnitude. In Sec. 3.4 we show that the performance of TMNRE is not impeded when posteriors are significantly non-Gaussian, demonstrating on neutrino mass upper bounds. In Sec. 3.5 we show how *local amortization* allows us to inspect the statistical consistency of our results. Finally, in Sec. 3.6 we demonstrate that the performance gain demonstrated by TMNRE is robust even when adopting excessively wide priors. We conclude in Sec. 4.

## 2 Simulation-based inference methodology

The primary concern of this paper is to determine a probability distribution over model parameters  $\boldsymbol{\theta}$  given an observation  $\mathbf{x}$ . This probability distribution, called the posterior, is given by Bayes’ rule as

$$p(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \boldsymbol{\theta})}{p(\mathbf{x})} p(\boldsymbol{\theta}). \quad (2.1)$$

The likelihood of the data  $\mathbf{x}$  given parameters  $\boldsymbol{\theta}$  is denoted  $p(\mathbf{x} \mid \boldsymbol{\theta})$ ,  $p(\boldsymbol{\theta})$  are our prior beliefs about the parameters, and  $p(\mathbf{x})$  represents the evidence of the data.

---

<sup>3</sup><https://github.com/undark-lab/swyft>

The posterior is extremely challenging to compute for most interesting physical systems, making approximation techniques commonplace. Traditional approaches are distinguished by whether point-wise evaluation of the likelihood function is needed. When the likelihood can be computed, methods such as Markov-chain Monte Carlo (MCMC) [21–23] and nested sampling [24–28] produce samples from the posterior distribution. Nested sampling can even estimate the evidence, which can be a valuable ingredient in model comparison. Alternatively, when the likelihood function is intractable one can use Approximate Bayesian computation (ABC) [4, 29–31], which produces samples from an approximate posterior. However, on a practical level ABC requires introducing summary statistics, a choice which can strongly influence the approximation quality and the computational cost.

Beyond these traditional approaches, recent years have seen significant progress in techniques that avoid explicitly calculating the likelihood function – these methods fall under the umbrella of what is known as Simulation-Based Inference (SBI) [5]. This taxonomy includes ABC, along with *e.g.* various deep learning algorithms that train neural networks to approximate the posterior. The fundamental input in SBI is a stochastic simulator that maps from model parameters  $\boldsymbol{\theta}$  to data  $\mathbf{x}$ ; notice that this mapping is equivalent to sampling from the distribution  $\mathbf{x} \sim p(\mathbf{x} | \boldsymbol{\theta})$  (*i.e.* the likelihood). When likelihood evaluation is not tractable, we call the distribution an *implicit likelihood* function. Using this stochastic simulator, SBI methods generate a set of  $N$  sample-parameter pairs  $\{(\mathbf{x}^{(1)}, \boldsymbol{\theta}^{(1)}), (\mathbf{x}^{(2)}, \boldsymbol{\theta}^{(2)}), \dots, (\mathbf{x}^{(N)}, \boldsymbol{\theta}^{(N)})\}$ , which can then be used as training data for a neural network that estimates either the posterior, the likelihood, or the likelihood-to-evidence ratio. We follow [32] and call these methods Neural Posterior Estimation (NPE) [8, 33–36], Neural Likelihood Estimation (NLE) [7], and Neural Ratio Estimation (NRE) [6, 35], respectively. Both NPE and NLE estimate a normalized probability density, and are therefore restricted to architectures like mixture density networks [37] or normalizing flows [38, 39]. On the other hand, as we describe in Sec. 2.1, NRE requires only a parameterizable classifier, and is therefore highly flexible in its network architecture. NPE, NLE, and NRE all produce an *amortized estimate* of the posterior, which means that once the relevant network is trained, it can sample from the posterior  $p(\boldsymbol{\theta} | \mathbf{x})$  for any data  $\mathbf{x}$  without retraining. Used appropriately, SBI enables a significant reduction in computational cost when compared with sampling-based methods like MCMC. These properties have led to the application of SBI in various astrophysical contexts [9, 14, 15, 19].

In paper we apply a new technique in SBI called *Truncated Marginal Neural Ratio Estimation* (TMNRE) [16, 17]. The rest of this section is organized as follows. In Sec. 2.1, we outline TMNRE, describing how it extends and improves upon the standard NRE approach. In Sec. 2.2 we describe how TMNRE’s automatic marginalization leads to a large gain in simulation efficiency. In Sec. 2.3 we define coverage tests that can quickly assess the statistical consistency of results generated by TMNRE. In Sec. 2.4 we explain how TMNRE allows the re-use of simulations across inferences. We review related applications of SBI to cosmology problems (comparing in particular to PYDELFI [9, 19]) in Sec. 2.5.

## 2.1 Truncated Marginal Neural Ratio Estimation

Here we explain the TMNRE algorithm. We begin by introducing NRE and theory behind likelihood-to-evidence ratio estimation. We then describe how to extend NRE by directly targeting marginal posteriors and estimating a truncated prior over several rounds.

**Neural Ratio Estimation** The target of NRE is the likelihood-to-evidence ratio  $r(\mathbf{x}, \boldsymbol{\theta}) \equiv \frac{p(\mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{x})} = \frac{p(\boldsymbol{\theta}|\mathbf{x})}{p(\boldsymbol{\theta})} = \frac{p(\mathbf{x}, \boldsymbol{\theta})}{p(\mathbf{x})p(\boldsymbol{\theta})}$ . Note that this is equal to the ratio of the probability densities

of jointly drawn sample-parameter pairs,  $\mathbf{x}, \boldsymbol{\theta} \sim p(\mathbf{x}, \boldsymbol{\theta})$ , to marginally drawn pairs,  $\mathbf{x}, \boldsymbol{\theta} \sim p(\mathbf{x})p(\boldsymbol{\theta})$ . We introduce a binary random variable  $y$  corresponding to whether a pair was drawn jointly or marginally, and let each outcome of  $y$  be equally probable (in other words, we enforce that the population sizes are equal). To avoid notation clash, we use  $\tilde{p}(\mathbf{x}, \boldsymbol{\theta}, y)$  to denote the joint distribution including  $y$  and associated quantities. Given  $y$ , we have

$$\tilde{p}(\mathbf{x}, \boldsymbol{\theta} | y) = \begin{cases} p(\mathbf{x}, \boldsymbol{\theta}) & \text{if } y = 1 \\ p(\mathbf{x})p(\boldsymbol{\theta}) & \text{if } y = 0. \end{cases} \quad (2.2)$$

Then  $r(\mathbf{x}, \boldsymbol{\theta})$  may be expressed using a classifier  $\tilde{p}(y = 1 | \mathbf{x}, \boldsymbol{\theta})$ , which distinguishes sample-parameter pairs drawn jointly from those drawn marginally:

$$\begin{aligned} \frac{p(\mathbf{x}, \boldsymbol{\theta})}{p(\mathbf{x})p(\boldsymbol{\theta})} &= \frac{\tilde{p}(\mathbf{x}, \boldsymbol{\theta} | y = 1)}{\tilde{p}(\mathbf{x}, \boldsymbol{\theta} | y = 0)} = \frac{\tilde{p}(\mathbf{x}, \boldsymbol{\theta}, y = 1)}{\tilde{p}(\mathbf{x}, \boldsymbol{\theta}, y = 0)} \\ &= \frac{\tilde{p}(y = 1 | \mathbf{x}, \boldsymbol{\theta})}{\tilde{p}(y = 0 | \mathbf{x}, \boldsymbol{\theta})} \\ &= \frac{\tilde{p}(y = 1 | \mathbf{x}, \boldsymbol{\theta})}{1 - \tilde{p}(y = 1 | \mathbf{x}, \boldsymbol{\theta})}. \end{aligned} \quad (2.3)$$

This is the so-called Likelihood Ratio Trick [40], which allows one to estimate  $\frac{p(\mathbf{x} | \boldsymbol{\theta})}{p(\boldsymbol{\theta})}$  by estimating  $\tilde{p}(y = 1 | \mathbf{x}, \boldsymbol{\theta})$ . Rewriting this expression, we have

$$\tilde{p}(y = 1 | \mathbf{x}, \boldsymbol{\theta}) = \sigma(\ln r(\mathbf{x}, \boldsymbol{\theta})), \quad (2.4)$$

where we have introduced the logistic sigmoid function  $\sigma(a) \equiv \frac{1}{1 + \exp(-a)}$ . We train a neural network parameterized by weights  $\phi$  to estimate  $\tilde{p}(y = 1 | \mathbf{x}, \boldsymbol{\theta})$ , in other words

$$\hat{\tilde{p}}_\phi(y = 1 | \mathbf{x}, \boldsymbol{\theta}) = \sigma(\ln \hat{r}_\phi(\mathbf{x}, \boldsymbol{\theta})). \quad (2.5)$$

For numerical stability we do not directly train  $\hat{r}_\phi$  but rather its logarithm, extracting  $\hat{r}_\phi$  at the end when performing Bayesian inference.

The estimator  $\hat{\tilde{p}}_\phi(y | \mathbf{x}, \boldsymbol{\theta})$  is trained by maximizing its value on the observed data. Maximizing this joint probability (over multiple data points) is equivalent to maximizing the sum of log probabilities. In other words, we seek to minimize

$$\mathbb{E}_{\tilde{p}(y, \mathbf{x}, \boldsymbol{\theta})} \left[ -\ln \hat{\tilde{p}}_\phi(y | \mathbf{x}, \boldsymbol{\theta}) \right] = -\mathbb{E}_{p(\mathbf{x}, \boldsymbol{\theta})} \left[ \ln \hat{\tilde{p}}_\phi(y = 1 | \mathbf{x}, \boldsymbol{\theta}) \right] - \mathbb{E}_{p(\mathbf{x})p(\boldsymbol{\theta})} \left[ \ln \left( 1 - \hat{\tilde{p}}_\phi(y = 1 | \mathbf{x}, \boldsymbol{\theta}) \right) \right], \quad (2.6)$$

which, for our problem, means finding the parameters  $\phi$  that satisfy  $\arg \min_{\phi} \ell(\hat{r}_\phi)$ , where

$$\ell(\hat{r}_\phi) = - \int d\mathbf{x} d\boldsymbol{\theta} \{ p(\mathbf{x}, \boldsymbol{\theta}) \ln \sigma(\ln \hat{r}_\phi(\mathbf{x} | \boldsymbol{\theta})) + p(\mathbf{x})p(\boldsymbol{\theta}) \ln (1 - \sigma(\ln \hat{r}_\phi(\mathbf{x} | \boldsymbol{\theta}))) \}. \quad (2.7)$$

This final expression is also known as the binary cross-entropy [37, 41]. In practice, one draws samples from the joint distribution and trains the classifier using a sample-based approximation to the above loss function. This optimization is standard practice in deep learning and accomplished by so-called *stochastic gradient descent*. Once the algorithm converges, the parameterized ratio estimator  $\hat{r}_\phi$  allows for fast point-wise evaluation of the approximate likelihood-to-evidence ratio, enabling rejection sampling and MCMC sampling from the approximate posterior.

**Data compression** In practice, we find it beneficial to split the network structure of the ratio estimator  $\hat{r}_\phi(\mathbf{x}, \boldsymbol{\theta})$  into two distinct components, as

$$\hat{r}_\phi(\mathbf{x}, \boldsymbol{\theta}) = D_\phi(\mathbf{s} = C_\phi(\mathbf{x}), \boldsymbol{\theta}) . \quad (2.8)$$

The first component is a data compression network,  $C_\phi(\mathbf{x})$ , which learns to compress the potentially high-dimensional data into a low-dimensional feature or summary statistics  $\mathbf{s}$ . The structure of this network is heavily dictated by the specific properties of the data. For example, for image data, convolutional neural networks are often appropriate. The second component,  $D_\phi(\mathbf{s}, \boldsymbol{\theta})$  performs the actual ratio estimation, learning to discriminate between marginally and jointly drawn feature vectors  $\mathbf{s}$  and model parameters  $\boldsymbol{\theta}$ . This binary classification network is typically a simple dense network with a few layers. In this work, we use a multi-layer perceptron (MLP) with three hidden layers, each of width 256, see Fig. 1 for an illustration.

Both networks are trained simultaneously, minimizing the loss function Eq. (2.7) as described above. *Importantly, this means that both inference and data compression happens concurrently, in contrast to other algorithms in the literature* (see discussion in Sec. 2.5).

It is of interest to examine how optimization selects features  $\mathbf{s}$  if the compression network has limited bandwidth. Assuming that for a given compression network  $C_\phi(\mathbf{x})$  the ratio estimator  $r_\phi$  is fully converged, in the sense that it equals to very good approximation the posterior-to-prior ratio  $r_\phi(\mathbf{x}, \boldsymbol{\theta}) \simeq p(\boldsymbol{\theta}|\mathbf{s} = C_\phi(\mathbf{x}))/p(\boldsymbol{\theta})$ . Note that the posterior is here conditioned on the summary statistics  $\mathbf{s}$ , since it is this summary statistics that appears in the classifier in Eq. (2.8). The loss function Eq. (2.7) can then be written as<sup>4</sup>

$$\ell[r_\phi] = 2 \ln 2 - 2 \mathbb{E}_{p(\mathbf{x})} [D_{JS}(p(\boldsymbol{\theta}|\mathbf{s} = C_\phi(\mathbf{x})) \| p(\boldsymbol{\theta}))] . \quad (2.10)$$

Up to a constant, this is two times the negative data-averaged Jensen-Shannon divergence (JSD) between the posterior (given compressed data  $\mathbf{s}$ ) and prior. The JSD is given by

$$D_{JS}(P||Q) = \frac{1}{2} D_{KL}(P \| M) + \frac{1}{2} D_{KL}(Q \| M) , \quad (2.11)$$

where  $M = (P + Q)/2$ , where  $D_{KL}$  denotes the Kullback-Leibler divergence

$$D_{KL}(P||Q) = \int dx P(x) \log \left( \frac{P(x)}{Q(x)} \right) . \quad (2.12)$$

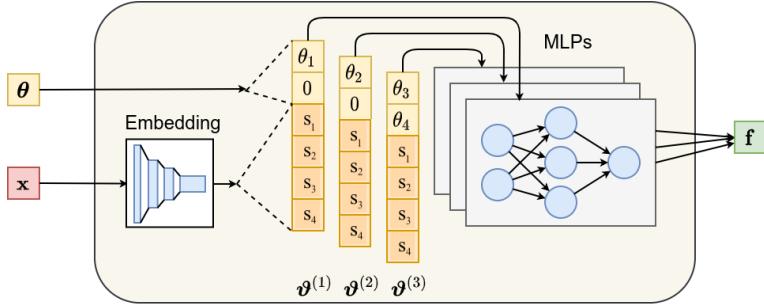
In other words, data summaries  $\mathbf{s}$  are learned such that they maximize the difference between posteriors and priors in terms of the JSD. Note that by definition the JSD is bound  $0 \leq D_{JS} \leq \ln 2$ .

---

<sup>4</sup>This follows directly from the definition of the JSD. In fact,

$$\begin{aligned} & -2 \mathbb{E}_{p(\mathbf{x})} [D_{JS}(p(\boldsymbol{\theta}|\mathbf{s}(\mathbf{x})) \| p(\boldsymbol{\theta}))] \\ &= -\mathbb{E}_{p(\mathbf{x})} \left[ D_{KL} \left( p(\boldsymbol{\theta}|\mathbf{s}(\mathbf{x})) \| \frac{1}{2} (p(\boldsymbol{\theta}|\mathbf{s}(\mathbf{x})) + p(\boldsymbol{\theta})) \right) + D_{KL} \left( p(\boldsymbol{\theta}) \| \frac{1}{2} (p(\boldsymbol{\theta}|\mathbf{s}(\mathbf{x})) + p(\boldsymbol{\theta})) \right) \right] \\ &= - \int d\mathbf{x} d\boldsymbol{\theta} \underbrace{p(\mathbf{x})p(\boldsymbol{\theta}|\mathbf{s}(\mathbf{x}))}_{\rightarrow p(\mathbf{x}, \boldsymbol{\theta})} \ln \frac{p(\boldsymbol{\theta}|\mathbf{s}(\mathbf{x}))}{\frac{1}{2} (p(\boldsymbol{\theta}|\mathbf{s}(\mathbf{x})) + p(\boldsymbol{\theta}))} + p(\mathbf{x})p(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta})}{\frac{1}{2} (p(\boldsymbol{\theta}|\mathbf{s}(\mathbf{x})) + p(\boldsymbol{\theta}))} \\ &= \ell[r_\phi] - 2 \ln 2 . \end{aligned} \quad (2.9)$$

The substitution in the third line is valid because the log depends on data  $\mathbf{x}$  only through the summary  $\mathbf{s}(\mathbf{x}) \equiv C_\phi(\mathbf{x})$ .



**Figure 1:** Illustration of the network architecture, including the “embedding” network  $C_\phi(\mathbf{x})$  and the discrimination network (“MLP”). Input data  $\mathbf{x}$  and parameters  $\boldsymbol{\theta}$  are internally mapped to the marginal parameter combinations of interest, for which individual ratio estimators with a shared compression network are trained. The outputs are estimated ratios for the marginal posteriors of interest.

**Marginalization** Scientific insight is often based on a low-dimensional marginalization of the posterior with nuisance parameters removed. NRE estimates the full joint posterior  $p(\boldsymbol{\theta} | \mathbf{x})$  over all parameters, but the marginal posterior  $p(\boldsymbol{\vartheta} | \mathbf{x})$  is sufficient for our purposes. Here we use  $\boldsymbol{\vartheta}$  to denote the parameters of interest, which are in general a low-dimensional subset of the full set of parameters  $\boldsymbol{\vartheta} \subset \boldsymbol{\theta} = (\boldsymbol{\vartheta}, \boldsymbol{\eta})$ . We propose estimating  $r(\boldsymbol{\vartheta} | \mathbf{x}) \equiv \frac{p(\mathbf{x} | \boldsymbol{\vartheta})}{p(\mathbf{x})}$  directly by Marginal Neural Ratio Estimation (MNRE) [6, 17]<sup>5</sup>, a straightforward modification of NRE. The simulation procedure for MNRE is exactly the same as it is for NRE. In MNRE we train  $\hat{r}_\phi(\mathbf{x} | \boldsymbol{\vartheta})$ , instead of the full likelihood-to-evidence ratio, by optimizing (2.7) using a classifier restricted to only the parameters of interest  $\boldsymbol{\vartheta}$ . Evaluating the marginal posterior  $p(\boldsymbol{\vartheta} | \mathbf{x})$  requires access to the marginalized prior  $p(\boldsymbol{\vartheta})$ , but this is easy to sample and evaluate.

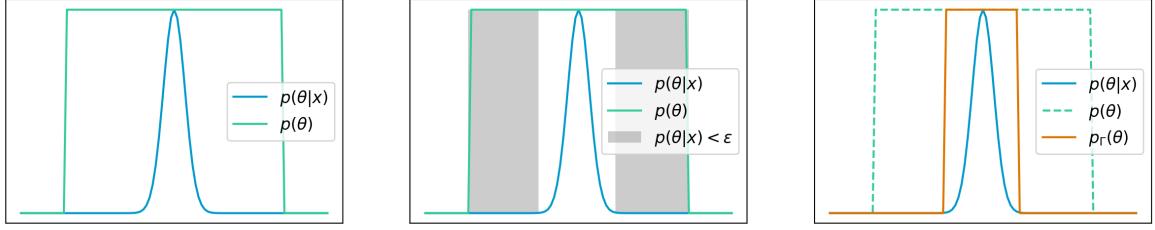
From a practical perspective, a given inference involves training several classifiers (corresponding to marginal ratio estimators) in parallel. For complex data, it is generally useful to compress the data before it is input into the inference network. In this paper, we use a *shared* compression network as input to *all* of the classification networks. The compression and inference are performed simultaneously, as described above. The structure is illustrated in Fig. 1.

**Truncation** Given observed data  $\mathbf{x}_o$ , the corresponding posterior may have (significantly) narrower approximate support (in the sense that  $p(\boldsymbol{\theta} | \mathbf{x}_o) > \tilde{\epsilon}$  for small  $\tilde{\epsilon}$ , which we denote  $\text{supp}_{\tilde{\epsilon}} p(\boldsymbol{\theta} | \mathbf{x}_o)$ ) than the prior. Therefore, although in NRE we draw samples from the full joint distribution  $p(\mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta})$ , we can learn the same ratio estimator in a local region of parameter space by drawing instead from a truncated joint distribution  $p(\mathbf{x} | \boldsymbol{\theta})p_{\Gamma}(\boldsymbol{\theta})$ , where the support of  $p_{\Gamma}(\boldsymbol{\theta})$  is smaller than the support of  $p(\boldsymbol{\theta})$ . Symbolically, given  $\mathbf{x}_o$  we denote the plausible subset of parameters lying above the posterior isocontour  $\tilde{\epsilon}$  as

$$\Gamma \equiv \{\boldsymbol{\theta} \in \text{supp } p(\boldsymbol{\theta}) \mid p(\boldsymbol{\theta} | \mathbf{x}_o) > \tilde{\epsilon}\}, \quad (2.13)$$

---

<sup>5</sup>Directly targeting marginal posteriors in the context of density estimation appears in [9, 36]. Another technique which aims to learn all marginals simultaneously, using likelihood-to-evidence ratios, appears in [42].



**Figure 2:** An illustration of truncation in one dimension. Between rounds, the prior is truncated to the region where the posterior estimator exceeds a predetermined threshold.

where  $\text{supp}$  returns the support of its argument. We then avoid simulation of parameters outside  $\Gamma$  by replacing our prior  $p(\boldsymbol{\theta})$  with a truncated prior

$$p_\Gamma(\boldsymbol{\theta}) \equiv V^{-1} \mathbb{1}_\Gamma(\boldsymbol{\theta}) p(\boldsymbol{\theta}), \quad (2.14)$$

where  $\mathbb{1}_\Gamma(\boldsymbol{\theta})$  is an indicator function that is unity on  $\Gamma \subset \text{supp } p(\boldsymbol{\theta})$  and zero otherwise, and  $V^{-1}$  is a normalizing constant (which can be interpreted as the fractional volume of the truncated prior).

Using a truncated prior  $p_\Gamma(\boldsymbol{\theta})$  focuses the simulation of parameters that feasibly lie within the highest density regions of the posterior. Truncated Marginal Neural Ratio Estimation (TMNRE) [16, 17] implements this truncation in the context of MNRE. In particular, we perform a round of MNRE, giving a set of estimated marginal posteriors  $\hat{p}(\boldsymbol{\vartheta} | \mathbf{x}_o)$ . These estimators are then used in an expression analogous to eqn. (2.13) to approximate  $\Gamma$ . An illustration of this process is shown in Fig. 2.

To be more precise, we must discuss  $\Gamma$ . How we choose to approximate  $\Gamma$  is informed by how we suspect the approximate support of the posterior factorizes. Note that since we will use MNRE to determine our approximation, we must necessarily deal with a product of low-dimensional projections of  $\Gamma$ , with each projection determined by a marginal posterior for  $\vartheta$ . In the simplest case, we might have an approximation that factorizes over each parameter individually, in other words  $\Gamma_\pi = \text{supp}_\epsilon p(\vartheta_1 | \mathbf{x}_o) \times \dots \times \text{supp}_\epsilon p(\vartheta_D | \mathbf{x}_o)$  for a  $D$ -dimensional parameter space and  $\vartheta_i = \theta_i$ . Here we have introduced  $\pi$  to denote a collection of projections we use to describe the approximated relevant region  $\Gamma_\pi$ , *i.e.*  $\pi = \{p(\vartheta_1 | \mathbf{x}_o), \dots, p(\vartheta_D | \mathbf{x}_o)\}$ . Note that  $\Gamma_\pi \supset \Gamma$ . Alternatively, it may be that significant correlations are expected in the 2-dimensional marginal posterior for *e.g.*  $\theta_1$  and  $\theta_2$ . In that case a more efficient approximation is given by  $\Gamma_\pi = \text{supp}_\epsilon p(\vartheta_1 = (\theta_1, \theta_2) | \mathbf{x}_o) \times \dots \times \text{supp}_\epsilon p(\vartheta_{D-1} | \mathbf{x})$ . For notational convenience we assume the former in the rest of this section.

We estimate  $\mathbb{1}_{\Gamma_\pi}$  by a sequence of nested indicator functions  $\mathbb{1}_{\Gamma_\pi^{(m)}}$  whose regions have the property  $\text{supp } p(\theta_1) \times \dots \times \text{supp } p(\theta_D) \equiv \Gamma_\pi^{(1)} \supset \dots \supset \Gamma_\pi^{(M)} \supset \Gamma_\pi$ . The sequence iteratively approximates the indicator function  $\mathbb{1}_{\Gamma_\pi}$  in  $M$  rounds and is generated via the following steps:

- Initialize  $\Gamma_\pi^{(1)} = \text{supp } p(\theta_1) \times \dots \times \text{supp } p(\theta_D)$ , *i.e.* we start with the unconstrained prior.
- In each round  $1 \leq m \leq M$ , we train  $D$ , one dimensional ratio estimators  $\hat{r}_{d, \Gamma_\pi^{(m)}}(\mathbf{x} | \theta_d)$  using data from within the constrained region,  $\boldsymbol{\theta} \in \Gamma_\pi^{(m)}$ . The estimated marginal posterior is  $\hat{p}_{\Gamma_\pi^{(m)}}(\theta_d | \mathbf{x}) = \hat{r}_{d, \Gamma_\pi^{(m)}}(\mathbf{x} | \theta_d) p_{\Gamma_\pi^{(m)}}(\theta_d)$ . To this end, we perform MNRE, setting  $\boldsymbol{\vartheta} = \theta_d$ ,  $d \in \{1, 2, \dots, D\}$  using the constrained prior  $p_{\Gamma_\pi^{(m)}}(\boldsymbol{\theta})$  with  $N^{(m)}$  training samples per round.

- For each round  $m < M$ , we estimate the indicator function for the next round using the approximated posteriors, the analogy of eqn. (2.13) being

$$\Gamma_{\pi}^{(m+1)} = \left\{ \boldsymbol{\theta} \in \Gamma_{\pi}^{(m)} \mid \forall d : \frac{\hat{p}_{\Gamma_{\pi}^{(m)}}(\theta_d \mid \mathbf{x}_o)}{\max_{\theta_d} \hat{p}_{\Gamma_{\pi}^{(m)}}(\theta_d \mid \mathbf{x}_o)} > \epsilon \right\}. \quad (2.15)$$

- The last round is determined either when  $m = M$  or when a stopping criterion is reached. The stopping criterion is defined by the ratio of consecutive truncated prior masses. It is satisfied when the sequence of truncated priors have the property

$$\int \mathbb{1}_{\Gamma_{\pi}^{(m)}}(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} / \int \mathbb{1}_{\Gamma_{\pi}^{(m-1)}}(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} > \beta.$$

We often set  $\beta = 0.8$ .

- Using the data from this final constrained region, we can approximate any marginal posterior of interest  $p(\boldsymbol{\vartheta} \mid \mathbf{x})$ . We emphasize that the data already generated during the truncation phase can be reused to learn arbitrary marginals of interest. When higher accuracy likelihood-to-evidence ratio estimates are needed, the user can simulate from the truncated region.

For considerations of the geometry of truncation in more than 1 dimension, see [43]. In most of the examples in Sec. 3 we perform MNRE (*i.e.* a single round) rather than TMNRE. We demonstrate TMNRE (*i.e.* multiple rounds) in Sec. 3.6. An illustration of the truncation in that example is shown in Fig. 13.

With the details of TMNRE in hand, we now address how our algorithm fares in comparison to alternative sampling-based inference methods, specifically highlighting the benefits offered by TMNRE in terms of efficiency, scalability, and testability.

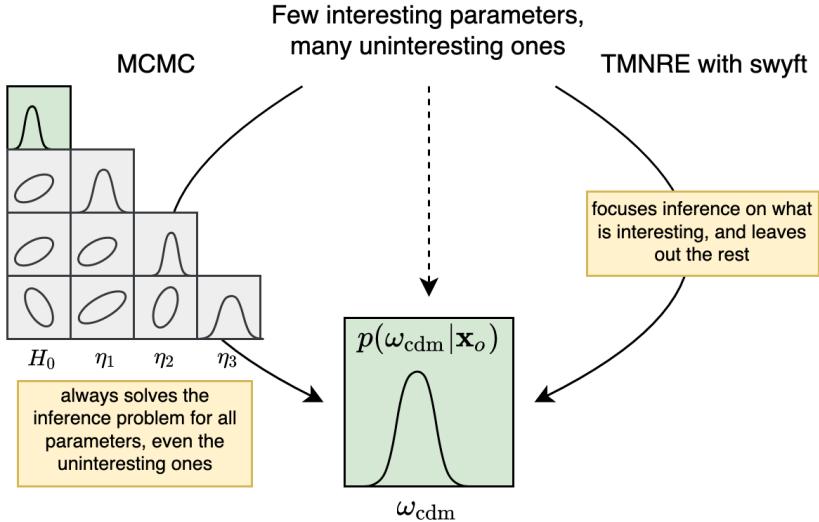
## 2.2 “Cutting to the chase” by estimating marginal posteriors

As we have mentioned, scientific conclusions about the measured value of a parameter of interest, say the cold dark matter energy density  $\omega_{cdm}$ , are based on *marginal* posteriors  $p(\omega_{cdm} \mid \mathbf{x}_o)$ . As illustrated in Fig. 3, SBI methods such as TMNRE allow us to directly estimate marginal posteriors of interest, whereas likelihood-based methods must always first estimate the joint posterior for all model parameters. More specifically, the marginal posterior is related to the joint posterior,  $p(\omega_{cdm}, \eta_1, \dots, \eta_N \mid \mathbf{x}_o)$ , via integration,

$$p(\omega_{cdm} \mid \mathbf{x}) = \int d\boldsymbol{\eta} p(\omega_{cdm}, \eta_1, \dots, \eta_N \mid \mathbf{x}). \quad (2.16)$$

Likelihood-based inference techniques do not directly generate samples from the one-dimensional or two-dimensional posteriors of interest, but instead always generate first samples from the full high-dimensional joint posterior. Those can then be marginalized (*i.e.* projected onto  $\omega_{cdm}$  direction) to generate samples from the marginal posterior  $p(\omega_{cdm} \mid \mathbf{x})$ .

Estimating low-dimensional marginals of interest with simulation-based techniques can therefore be significantly more computationally efficient than estimating the joint posterior with likelihood-based methods, especially when the number of parameters and/or modes is high. In fact, the numbers of samples required for convergence of likelihood-based methods increases very quickly (typically cubically [27] or even exponentially [26], in the case of nested



**Figure 3:** TMNRE can handle complex models and directly estimate any marginal posterior of interest, which typically requires far fewer simulation runs than estimating the full joint posterior.

sampling) as the number of model parameters grows.<sup>6</sup> In contrast, the required number of simulations for simulation-based techniques can be nearly independent of the number of parameters involved if only low-dimensional marginals are of interest.

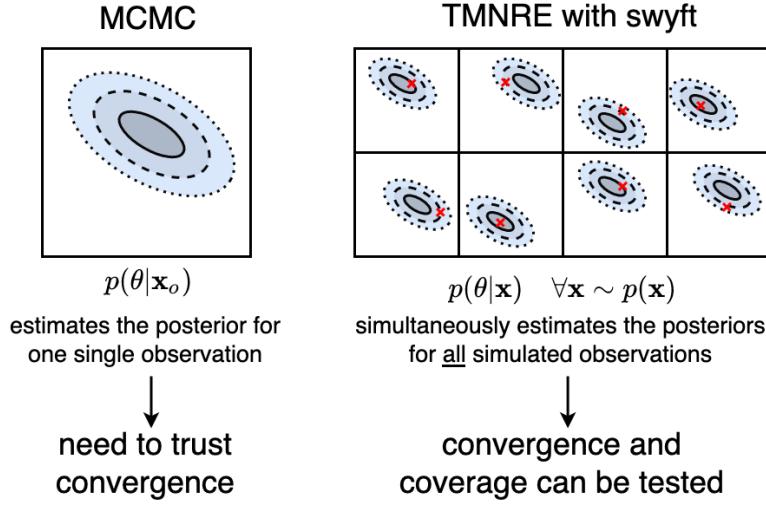
As a result, likelihood-free inference methods, like TMNRE, enable the estimation of any marginal posteriors of interest with orders of magnitude fewer simulations runs than likelihood-based methods. We demonstrate this in the context of CMB data throughout Sec. 3.

Note that we are not claiming that the end goal of cosmological data analysis is to produce 6 one-dimensional posteriors corresponding to each  $\Lambda$ CDM parameter. Indeed, it is often of interest to assess higher-dimensional structure. For example, if several experiments give high-dimensional posteriors that are in tension with each other, this can be obscured by a low-dimensional projection.<sup>7</sup> Additionally, it is generally of interest to assess degeneracies between constraints on different parameters (hence the typical “triangle plot”). Finally, although we are using a language in which it is useful to “ignore” nuisance parameters (i.e. learn a low-dimensional distribution), note that in general we propose learning low-dimensional posteriors for the nuisance parameters and using them in the truncation process. Therefore at the end posteriors for nuisance parameters may be inspected to assess consistency and extract information about e.g. foregrounds, systematics, and astrophysics. Indeed, from the perspective of assessing the  $S_8$  tension by examining the  $(\sigma_8, \Omega_m)$  the other  $\Lambda$ CDM parameters are ignored, and these are clearly relevant to cosmology!

<sup>6</sup>For a pedagogical introduction to Markov Chain Monte Carlo see Ref. [44].

<sup>7</sup>If one is interested in quantifying a tension between two datasets, this is in fact possible in the framework of neural ratio estimation without examining marginal posteriors. Instead, one can consider the contrastive problem of distinguishing between  $(x_1, x_2) \sim p(x_1, x_2)$  and  $(x_1, x_2) \sim p(x_1)p(x_2)$  (where  $x_i$  denotes the data from a particular experiment), which proceeds in the full (high-dimensional) parameter space. We leave this avenue for future work.

## 2.3 “Getting it right” through coverage tests for simulation-based inference



**Figure 4:** Testability: Many SBI methods, including TMNRE do not estimate just a single posterior, but *all* of them simultaneously (“amortization”). This enables the user to test the reliability of the inference results.

Although the internal workings of the trained network are usually hard to understand without detailed inspection, the enormous evaluation speed of trained neural networks enables important consistency tests that we introduce below. These tests are generally infeasible for likelihood-based methods, where the estimate of one single posterior is already very costly (illustrated in the left panel of Fig. 4). For that reason, the expected statistical properties of Bayesian inference results (like biases, credibility) are usually not directly tested but are instead inferred based on convergence criteria [45, 46] and asymptotic properties of stationary Markov Chains. On the other hand, trained inference networks (like the ratio estimators in TMNRE), constitute essentially ultra-fast black-box data analysis pipelines. They take as input any observed data  $\mathbf{x} \sim p(\mathbf{x})$ , and very quickly generate the corresponding posteriors  $p(\boldsymbol{\theta} | \mathbf{x})$ . This is illustrated in the right panel of Fig. 4.

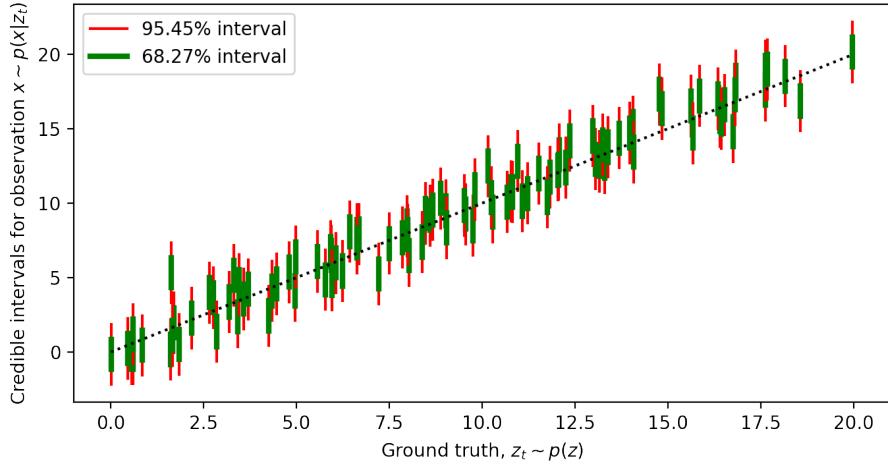
Let us denote with  $\Theta_{p(\boldsymbol{\theta}|\mathbf{x})}(1-\alpha)$  the  $1-\alpha$  highest posterior density region. For example, a 95% highest posterior density region would have an expected error rate of  $\alpha = 0.05$ . In other words, when randomly drawing observations  $\mathbf{x}, \boldsymbol{\vartheta} \sim p(\mathbf{x} | \boldsymbol{\vartheta})p(\boldsymbol{\vartheta})$ , we expect the true parameters  $\boldsymbol{\vartheta}$  to fall outside of the 95% region in 5% of the cases. Following Ref. [18], the expected coverage probability of the  $1 - \alpha$  highest posterior density region (HPDR) of some estimated posterior  $\hat{p}(\boldsymbol{\vartheta}|\mathbf{x})$  is given by

$$1 - \hat{\alpha} = \mathbb{E}_{p(\boldsymbol{\vartheta}, \mathbf{x})} [\mathbb{1} [\boldsymbol{\vartheta} \in \Theta_{\hat{p}(\boldsymbol{\vartheta}|\mathbf{x})}(1 - \alpha)]] \quad (2.17)$$

This quantity can be interpreted as both expected Bayesian credibility as well as expected Frequentist coverage probability of  $\hat{p}(\boldsymbol{\vartheta}|\mathbf{x})$  [18]. It enables us to estimate the *actual* error rate  $\hat{\alpha}$  of a  $1 - \alpha$  highest posterior density region of some estimated posterior  $\hat{p}(\boldsymbol{\vartheta}|\mathbf{x})$ .

**Coverage figure construction.** In the case where  $\hat{p}(\boldsymbol{\vartheta}|\mathbf{x}) = p(\boldsymbol{\vartheta}|\mathbf{x})$ , it follows that  $\hat{\alpha}(\alpha) = \alpha$ . We can hence use  $\hat{\alpha}$  as function of  $\alpha$  as quality measure for posteriors estimated with TMNRE. Since small values of  $\alpha$  are of particular interest, corresponding to posterior regions

with high mass, we find it convenient to reparameterize  $\alpha$  in terms of a new variable  $z$ . We define  $z$  as the  $1 - \frac{1}{2}\alpha$  quantile of the standard normal distribution, for instance  $\alpha = 0.05$  corresponds to  $z = 1.96$ . The commonly quoted “ $1\sigma$ ”, “ $2\sigma$ ” and “ $3\sigma$ ” regions, correspond to  $z = 1, 2, 3$  and hence have  $1 - \alpha = 0.6827, 0.9545, 0.9997$ .

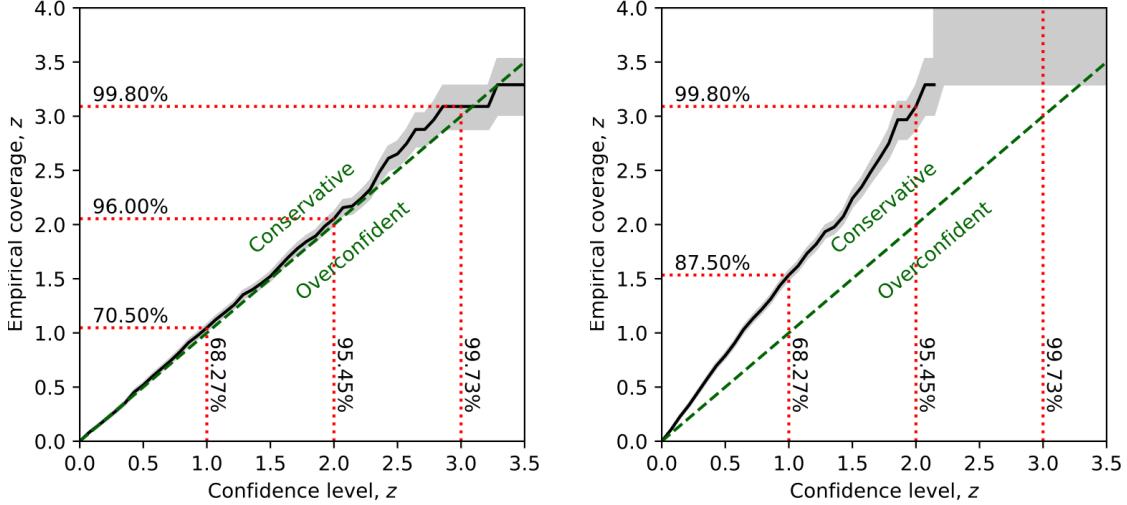


**Figure 5:** Illustration of a coverage test for a one-dimensional interval. The “ground truth” parameter values  $z_t$  for a range of test simulations are drawn from the prior,  $z_t \sim p(z)$ . The credible intervals are derived from the corresponding simultaneously generated mock observations,  $x \sim p(x|z_t)$ . One then obtains the number of times the ground truth is contained in, say, the 68.27% interval (66 times in our 100 examples), or the 95.45% interval (95 times).

We estimate empirically the actual error rate  $\hat{\alpha}$  (and the corresponding  $\hat{z}$ ) for nominal  $1 - \alpha$  regions as follows (see Fig. 5 for an illustrative example). We generate  $n$  samples from the joint model  $\mathbf{x}, \boldsymbol{\vartheta} \sim p(\mathbf{x}, \boldsymbol{\vartheta})$ , and count how often a  $1 - \alpha$  highest posterior density region predicted by the trained network does *not* contain the true parameter  $\boldsymbol{\vartheta}$ . We denote this number by  $k$ , which follows a binomial distribution with fractional probability  $\hat{\alpha}$  and total number of draws  $n$ . The mean is then simply given by  $\hat{\alpha} = k/n$ . We propose to estimate uncertainties of this quantity using the so-called Jeffreys interval, which is effectively the Bayesian credible interval obtained from a non-informative Jeffreys prior for a binomial distribution [47].<sup>8</sup>

We summarize the results of such a test in Fig. 6. The empirical coverage  $\hat{z}$  is shown as function of the confidence levels  $z$  for 1000 test simulations. We also show the error band based on Jeffreys interval, with the widths determined by the number of test simulations used. We furthermore indicate the empirical coverage in terms of  $1 - \hat{\alpha}$  (horizontal numbers) for typical values of the confidence level  $1 - \alpha$  (vertical numbers). In this particular example, the 68.27% (95.45%) highest posterior density region contains the ground truth value in 70.50% (96.00%) of the cases, and is hence slightly conservative, or equivalently the confidence intervals slightly over-cover (and are thus conservative). In the case of perfect coverage, one would expect the

<sup>8</sup>More specifically, the interval is obtained from the 68.27% central interval of a Beta distribution with parameters  $(n - k + \frac{1}{2}, k + \frac{1}{2})$ . One of the advantages of the Jeffreys interval is that it is approximately equally tailed, with intervals evenly distributed above and below the true value. This convenient property is preserved even when reparametrizing the inferred parameter in terms of  $\hat{z}$ .



**Figure 6:** Empirical coverage as function of the confidence level, as function of the significance level  $z$  (see text). In cases where  $\hat{z} > z$ , estimated credible contours are conservative and contain the true value with a frequency higher than nominally expected. The nominal probabilities  $1 - \alpha$  are shown in the figure as vertical text, the empirical estimates  $1 - \hat{\alpha}$  are shown as horizontal text. *Left panel:* Converged network with good coverage. *Right panel:* Conservative coverage, as typically shown by networks during training before convergence.

black line and the green dashed line in Fig. 6 to perfectly overlap. Should the black line lie above the green line, the confidence intervals over-cover (and are thus conservative; see the right panel of Fig. 6), while if the black line lies below the confidence intervals under-cover (*i.e.* are over-confident).

Figures like Fig. 6 can be used to test and validate the correct statistical behavior of an inference network, which is particularly useful when one does not have a ground-truth against which to compare the results (*e.g.* in the examples given below, our ground truth is taken to be a converged MCMC). For over-conservative or overconfident networks, those figures can be also used to provide corrections to quoted credible levels. In Sec. 3.5, we use empirical coverage plots in order to double-check and confirm the convergence of our posterior estimators for CMB data. Note that similar tests and plots have appear in the literature in [14, 18, 48]. Additionally, while what we propose is a necessary condition for the calibration of our approximate posteriors, it is not sufficient. Moreover, we only test the consistency of our approximation of the posterior given the simulation model and prior choice. Developing more tests in order to trust results generated by SBI is worthwhile and ongoing work.

## 2.4 Zero-cost inference through simulation re-use

Simulation runs in likelihood-based methods are commonly difficult to re-use, because the generation of proposal points for simulations is directly intertwined with the inference process for a specific observation. In contrast, the generation of training data through simulation runs and the training of inference networks are distinct steps in the context of TMNRE.<sup>9</sup> This makes

---

<sup>9</sup>Note that this is also a feature of other variational SBI methods.

the re-use of previous simulations almost trivial: the same training data, sampled from the generative model, can be used train multiple networks, it can be combined with other training data, or it can sub-sampled to change priors or reduce training data.

In the context of TMNRE, and for the purpose of this paper, we will adopt a simple sample cache model (for a more general case, see Ref. [16]). Let us assume we generated training  $\mathcal{S}_1$  data consisting of  $N$  samples from the truncated prior  $p_{\Gamma_1}(\boldsymbol{\vartheta})$ . Now we are interested in generating  $M$  samples from the truncated prior  $p_{\Gamma_2}(\boldsymbol{\vartheta})$ , where we assume that  $\Gamma_2 \subset \Gamma_1$ . We can now simply re-use all samples in  $\mathcal{S}_1$  that are inside  $\Gamma_2$  (let us assume there are  $N' \leq N$  of those), and then add to those  $M - N'$  samples from  $p_{\Gamma_2}(\boldsymbol{\vartheta})$ . In cases where the truncation region does not change, we can re-use all previous samples and hence simulations.

In Sec. 3.2 we re-use simulations to efficiently compare distinct experimental configurations. We demonstrate in Sec. 3.6 that this scheme allows us to quickly zoom from wide initial priors into the parameter region relevant for a given observation, while keeping the number of required simulations low.

## 2.5 Comparison with other methods

Although our focus in this work is a deep learning-based method for SBI, progress continues to be made in existing approaches as well. For example, ABC continues to be successfully applied to astrophysics, with public codes incorporating importance sampling via e.g. Population Monte Carlo [49, 50] or Sequential Monte Carlo [51]. Similarly, approaches that leverage Gaussian approximations of likelihoods have been used and combined with Gaussian processes [52, 53]. Gaussian process regression has also been used to enable Bayesian optimization [54–56]. Along the lines of Gaussian approximations, including perturbations around Gaussian likelihoods has also been pursued [57, 58].

On the deep learning side, applications to cosmology have appeared on several fronts [10–13]. Most applications have employed density estimation methods (DELF) [9, 19]. In the rest of this section, we review DELFI, specifying to its implementation via PYDELFI, and describe how it relates to TMNRE.

DELFI constitutes Neural Likelihood Estimation (NLE) or Neural Posterior Estimation (NPE). For notational simplicity, we refer to DELFI in NLE mode. NLE learns the sampling distribution  $\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta})$  of data  $\mathbf{x}$  as a function of model parameters  $\boldsymbol{\theta}$ . NLE becomes increasingly difficult when the dimension of the data is large, or the data distribution is highly complex [39]. This problem can be alleviated by compressing the data to some simple low-dimensional summary statistics before performing inference with NLE. In PYDELFI this is often done by introducing a compression network such as an Information Maximizing Neural Network, which maximizes the Fisher information of the compression variables [59]. This compression network is typically trained upfront before inference, generally at additional simulation costs. The network architecture used for the subsequent NLE inference step is restricted to density estimation networks such as mixture density networks [37] or normalizing flows [38, 39] (while simpler network structures such as multi-layer perceptron (MLP) are not an option). With compressed data summaries in hand, PYDELFI uses one of two approaches to draw new training data relevant for a given observation (sometimes coined “active learning”). In the first approach, PYDELFI uses Sequential NLE (SNLE) [7], in which the parameters for each batch of simulations are drawn from a proposal density based on the current posterior estimator, and the neural density estimators are re-trained after each simulation batch. Commonly the proposal density is the current posterior estimator itself. In the second approach, each new batch of simulations is given by a deterministic acquisition function that seeks to maximize

both relevance (*i.e.* large posterior density) and uncertainty (usually quantified by comparing the estimates of multiple networks).

PYDELFI and TMNRE are similar in several respects:

- *Amortization*: in non-active learning mode (*i.e.* NLE or MNRE), networks are manifestly amortized throughout the full prior. In active learning mode, TMNRE is manifestly amortized throughout the truncated prior, and the network trained by *e.g.* SNLE can still be evaluated throughout the prior. Both approaches learn quantities that are independent of the proposal distribution. While the proposal distribution means that SNLE is expected to be less accurate away from the observation of interest, the empirical tests afforded by amortization may still be carried out. This amortization is a unique aspect of methods that train variational approximations of Bayesian quantities that can be rapidly evaluated or sampled.
- *Sampling*: after training, both approaches use *e.g.* MCMC to sample from the (marginal) neural posterior. This evaluation is generally very fast thanks to the massive evaluation speed and parallelization afforded by neural networks, see *e.g.* [60]. This neural posterior sampling can also be combined with sampling from additional explicit likelihoods as needed.
- *Compression*: for complex data, it is advantageous for both methods to employ a compression network. The training of the compression network is slightly different in the two cases, see below.
- *Marginalization*: both approaches can directly target marginal posteriors [16, 17, 20, 36]. In the context of PYDELFI, this marginalization can be integrated into the compression step by constructing “nuisance-hardened” summary statistics [20]. PYDELFI has been shown to outperform MCMC even when targeting the full joint posterior [19]. The ability to directly target marginal posteriors is only possible for variational methods.
- *Active learning*: both approaches preferentially simulate with model parameters belonging to some region of interest. In TMNRE this is done via truncation. In PYDELFI this is done either via SNLE or optimization of a deterministic acquisition function. As mentioned previously, Bayesian optimization is also relevant to non-neural SBI [54–56].

We therefore emphasize that many of the attractive features of TMNRE we demonstrate via examples in Sec. 3 also have analogues using PYDELFI or other variational SBI approaches.

There are two main differences between TMNRE and PYDELFI:

- *Ratio vs. density estimation*: ratios (as learned by TMNRE) can be estimated with simple battle-tested classification networks, which are often easier to train than density estimation models (as used by PYDELFI) [39]. However, especially in the case where the functional shape of posteriors are sufficiently simple and Gaussian, the inductive bias of density estimation models can increase the simulation efficiency of density-estimation based approaches [17]. Moreover, density estimation learns a cheap and reusable generative model for the sampling distribution of the data, which is not true for NRE parameterized via *e.g.* an MLP. One option, which we leave to future work, is to parameterize ratio estimation using a flow-based model, in which case these benefits are shared.

The difference between density and ratio estimation affects how the compression network is trained. In PYDELFI, the compression network is trained before the inference network, while in TMNRE the compressed representation is learned simultaneously with the inference network.

- *Interplay between marginalization and active learning dynamics:* in TMNRE, active learning is accomplished through truncation, while in PYDELFI, it is accomplished via e.g. sequential methods. These approaches are especially different when targeting marginal posteriors. Note first that when a joint likelihood-to-evidence ratio or likelihood is estimated, the proposal distribution for model parameters  $\boldsymbol{\theta}$  drops out.<sup>10</sup> On the other hand, when targeting marginals so that  $\boldsymbol{\theta} = (\vartheta, \eta)$ , the proposal distribution for  $\eta$  is important. For example, the marginal likelihood is defined by  $p(\mathbf{x} | \vartheta) = \int d\eta p(\mathbf{x} | \vartheta, \eta)p(\eta)$ . The nuisance parameters  $\eta$  must be sampled according to the correct distribution. In truncation, the marginal posteriors of the nuisance parameters are learned and used to define a truncated prior *including the nuisance parameters*. This strategy is successful even when the priors for nuisance parameters are very wide, see [61] for an example. On the other hand, when PYDELFI uses sequential methods for marginals, the nuisance parameters are sampled from the full prior distribution. In practice the user aims to reduce the dependence of the summary statistics on the nuisance parameters via nuisance-hardening. In the optimally nuisance-hardened case there is no dependence of the summary statistics on the nuisance parameters, but this may be difficult to achieve in extreme cases.

More research is needed to compare in detail the performance of NRE, NLE, and NPE approaches in various settings. We expect that further progress can be made by combining insights from approaches that are regarded as distinct within the current taxonomy.

### 3 Cosmological applications

In this section we apply TMNRE to several cosmological examples. Each example is designed to showcase a particular attractive feature of the algorithm:

1. **Directly estimating marginals is much more efficient than estimating the joint and then marginalizing.** In Sec. 3.1 we describe at length how to define a simulator for CMB forecasting, and show that by learning marginal rather than joint posteriors, TMNRE converges much more quickly than MCMC.
2. **Rerunning the analysis with different experimental configurations does not require new simulations.** In Sec. 3.2 we demonstrate simulation reuse in the context of comparing CMB inferences with and without BAO measurements.
3. **Increasing the number of nuisance parameters does not increase the number of required simulations.** In Sec. 3.3 we perform inference for a realistic Planck simulator, showing that the 13 varying nuisance parameters do not impact the simulator cost for TMNRE.

---

<sup>10</sup>For NRE, this is only approximately true, since the evidence  $p(\mathbf{x}) = \int d\boldsymbol{\theta} p(\mathbf{x} | \boldsymbol{\theta})p(\boldsymbol{\theta})$  depends on  $p(\boldsymbol{\theta})$ ; however, this is a constant for fixed  $\mathbf{x}$ .

4. **Non-Gaussian posteriors are a non-issue.** In Sec. 3.4 we show that TMNRE accurately and cheaply infers an upper bound for the sum of neutrino masses, the marginal posterior of which is highly non-Gaussian.
5. **Amortization allows cheap consistency checks.** In Sec. 3.5 we demonstrate the rigorous statistical testing of TMNRE results, which can be performed even when the ground truth posteriors are unknown.
6. **Wide priors are not a problem.** In Sec. 3.6 we take a wide prior and show that TMNRE’s truncation efficiently identifies the relevant region of parameter space, so that simulations are not wasted on irrelevant parameters.

### 3.1 A CMB forecasting simulator

To use SBI, we require a simulator. In this section we define very explicitly a simulator with the same statistical content as CMB likelihoods commonly used for forecasting. We couple this simulator to TMNRE to show that:

**Directly estimating marginal posteriors is much more efficient than estimating the joint posterior and then marginalizing.**

We now describe the simulator at some length. Implementations of the corresponding likelihood are available via `MontePython`, in the `Likelihood_mock_cmb` class [62]. Our notation follows the discussion in [63]. In particular, the multipole coefficients  $a_{\ell m}$  of CMB maps receive contributions from the CMB signal  $s_{\ell m}$  and experimental noise  $n_{\ell m}$

$$a_{\ell m}^P = s_{\ell m}^P + n_{\ell m}^P \quad (3.1)$$

where the index  $P$  runs over temperature, polarization, and weak lensing degrees of freedom. For simplicity, we restrict to  $P = T, E$  in our analysis, but the extension is straightforward. In the full sky limit, the two-point correlation of  $a_{\ell m}$  is given by

$$\langle a_{\ell m}^{P*} a_{\ell' m'}^{P'} \rangle = \left( C_{\ell}^{PP'} + N_{\ell}^{PP'} \right) \delta_{\ell\ell'} \delta_{mm'} \equiv \bar{C}_{\ell}^{PP'} \delta_{\ell\ell'} \delta_{mm'} \quad (3.2)$$

where  $\langle \cdots \rangle$  denotes an ensemble average. For specified cosmological parameters, the power spectra  $C_{\ell}^{PP'}$  can be computed via a Boltzmann code such as `CLASS` [64]. The effects of spatially uniform Gaussian white noise and a Gaussian beam give rise to a noise contribution

$$N_{\ell}^{PP'} \equiv \langle n_{\ell m}^{P*} n_{\ell m'}^{P'} \rangle = \delta_{PP'} \theta_{\text{fwhm}}^2 \sigma_P^2 \exp \left( \ell(\ell+1) + \frac{\theta_{\text{fwhm}}^2}{8 \ln 2} \right) \quad (3.3)$$

where  $\theta_{\text{fwhm}}$  is the full width at half maximum of the beam and  $\sigma_P$  is the root-mean-square of the instrumental noise. In this section we will use values for  $N_{\ell}^{PP'}$  compatible with the Planck experiment (accessible within `MontePython` as `fake_planck_realistic`) [65].

Given  $\bar{C}_{\ell}(\boldsymbol{\theta})^{PP'}$  (making explicit the dependence on underlying cosmology parameters, collectively denoted  $\boldsymbol{\theta}$ ), the  $a_{\ell m}$  are distributed according to the likelihood

$$p(\mathbf{a} \mid \boldsymbol{\theta}) \propto \frac{1}{|\bar{C}(\boldsymbol{\theta})|^{1/2}} \exp \left( -\frac{1}{2} \mathbf{a}^\dagger [\bar{C}(\boldsymbol{\theta})^{-1}] \mathbf{a} \right) \quad (3.4)$$

where  $\mathbf{a} = \{a_{\ell m}^T, a_{\ell m}^E\}$  and  $\bar{C}(\boldsymbol{\theta})^{-1}$  denotes matrix inversion. In general, one may sample from this using the Cholesky decomposition of the covariance matrix. For fixed  $\ell$  and  $m$ , one has

$$\begin{pmatrix} a_{\ell m}^T \\ a_{\ell m}^E \end{pmatrix} = L \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} \quad (3.5)$$

where  $n_i$  are sampled from normal distributions with unit variance  $n_i \sim \mathcal{N}(0, 1)$  and  $L$  is a matrix satisfying  $LL^T = \begin{pmatrix} \bar{C}_\ell^{TT} & \bar{C}_\ell^{TE} \\ \bar{C}_\ell^{TE} & \bar{C}_\ell^{EE} \end{pmatrix}$ . For each  $\ell$  there are  $2\ell + 1$  modes. Therefore for a single realization of the universe,  $\bar{C}_\ell$  can only be determined to finite precision. Given a realization of  $a_{\ell m}^P$ , the maximum likelihood estimator for  $\bar{C}_\ell$ , which we denote  $\hat{C}_\ell^{PP'}$ , is given by

$$\hat{C}_\ell^{PP'} = \frac{1}{2\ell + 1} \sum_{m=-\ell}^{\ell} a_{\ell m}^{P*} a_{\ell m}^{P'} = \frac{1}{2\ell + 1} \left( a_{\ell 0}^P a_{\ell 0}^{P'} + 2 \sum_{m=1}^{\ell} a_{\ell m}^{P*} a_{\ell m}^{P'} \right) \quad (3.6)$$

Therefore the map  $\boldsymbol{\theta} \rightarrow \bar{C}(\boldsymbol{\theta}) \rightarrow \hat{C}$  is stochastic in the last step. For fixed  $\bar{C}_\ell$ , the Gaussian distribution of  $a_{\ell m}$  propagates to a Wishart distribution for  $\hat{C}_\ell$  [66]. Explicitly, given  $\bar{C}_\ell^{PP'}$ , the likelihood for  $\hat{C}(\boldsymbol{\theta})$  is given by

$$-2 \ln p(\hat{C}(\boldsymbol{\theta}) | \bar{C}) = \chi_{\text{eff}}^2 = \sum_{\ell} (2\ell + 1) \left[ \frac{D}{|\bar{C}|} + \ln \frac{|\bar{C}|}{|\hat{C}|} - 2 \right] \quad (3.7)$$

where  $D = \bar{C}_\ell^{TT} \hat{C}_\ell^{EE} + \hat{C}_\ell^{TT} \bar{C}_\ell^{EE} - 2\bar{C}_\ell^{TE} \hat{C}_\ell^{TE}$  and we have used our freedom in shifting the log-likelihood by a constant so that  $\ln p(\hat{C}(\boldsymbol{\theta}) = \bar{C} | \bar{C}) = 0$ . The Wishart distribution is significantly non-Gaussian at low  $\ell$ , which is necessary to reflect the fact that the autocorrelations  $\hat{C}_\ell^{PP}$  are positive. At large  $\ell$ , the Wishart distribution can be approximated as a Gaussian distribution with

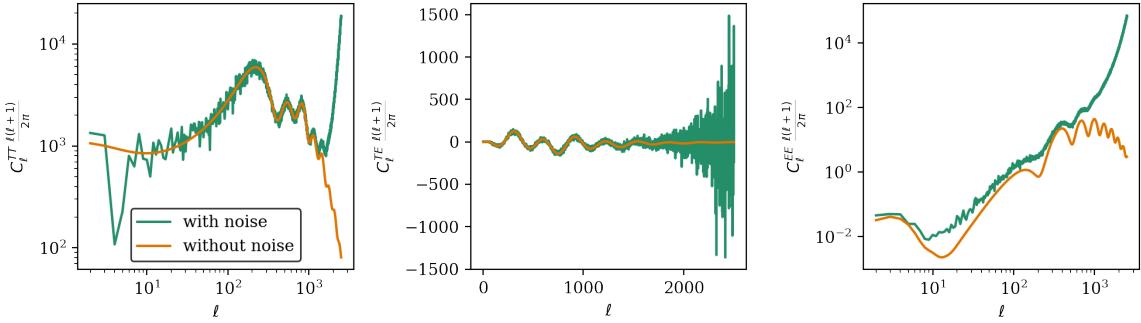
$$\text{Cov}_{C_\ell} = \frac{2}{2\ell + 1} \begin{pmatrix} (\bar{C}_\ell^{TT})^2 & \bar{C}_\ell^{TT} \bar{C}_\ell^{TE} & (\bar{C}_\ell^{TE})^2 \\ \bar{C}_\ell^{TT} \bar{C}_\ell^{TE} & \frac{1}{2} (\bar{C}_\ell^{TT} \bar{C}_\ell^{EE} + (\bar{C}_\ell^{TE})^2) & \bar{C}_\ell^{TE} \bar{C}_\ell^{EE} \\ (\bar{C}_\ell^{TE})^2 & \bar{C}_\ell^{TE} \bar{C}_\ell^{EE} & (\bar{C}_\ell^{EE})^2 \end{pmatrix} \quad (3.8)$$

Sampling from this distribution is more efficient than sampling from the full Wishart distribution, especially at large  $\ell$  where the approximation is valid. We will take the Gaussian approximation starting at  $\ell = 52$ .

Finally, we account for partial sky coverage by modifying the effective degrees of freedom at each  $\ell$ . We rescale the covariance matrix (3.8) by  $2\ell + 1 \rightarrow f_{\text{sky}}(2\ell + 1)$  at high  $\ell$ . At low  $\ell$ , we round  $f_{\text{sky}}(2\ell + 1)$  to the nearest integer and sample this many  $a_{\ell m}$ , subsequently using the maximum likelihood estimator to determine  $\hat{C}_\ell^{PP'}$ . (We show a power spectrum analysis with explicit mask in App. A.)

In other words, a simulator that is equivalent to the likelihood (3.7) is defined by the following steps:

1. Given cosmological parameters, compute  $C_\ell^{PP'}$  via a Boltzmann code.
2. Add experimental noise:  $\bar{C}_\ell^{PP'} = C_\ell^{PP'} + N_\ell^{PP'}$



**Figure 7:** Samples drawn from the CMB forecasting simulator. Here “without noise” denotes the output of `CLASS` (*i.e.*  $C_\ell^{PP'}$ ) and “with noise” includes both experimental noise and cosmic variance (*i.e.*  $\hat{C}_\ell^{PP'}$ ).

3. From  $\bar{C}_\ell^{PP'}$ , compute  $\hat{C}_\ell^{PP'}$  by sampling  $a_{\ell m}$  and the maximum likelihood estimator at low  $\ell$  (equivalently, sampling the Wishart distribution). At large  $\ell$ , use  $\hat{C}_\ell^{PP'} = \bar{C}_\ell^{PP'} + n_\ell^{PP'}$ , where  $n_\ell^{PP'}$  is sampled from the multivariate normal distribution (3.8).

Finally, in likelihood-based approaches uncertainties are incorporated in the form of the likelihood and therefore not sampled explicitly (for forecasting, sometimes even for the fiducial observation) [63]. Note that in SBI we must sample from the noise model. Samples from this simulator are shown in Fig. 7. We now proceed to perform inference with this simulator.

### 3.1.1 Inference setup and results

Using this simulator, we perform inference for a  $\Lambda$ CDM cosmology with  $\ell_{\max} = 2500$ . We use noise  $N_\ell^{PP'}$  given by `fake_planck_realistic` [65]. In the context of forecasting, we are free to fix the fiducial cosmology. We choose a fiducial cosmology<sup>11</sup>

$$\boldsymbol{\theta}^0 = (\omega_b, \omega_{cdm}, 100\theta_s, \ln(10^{10} A_s), n_s, \tau) = (0.0224, 0.12, 1.0411, 3.0753, 0.965, 0.054) \quad (3.9)$$

A Gaussian approximation to the likelihood can be computed using the Fisher information matrix

$$F_{ij} \equiv - \left. \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} \right|_{\boldsymbol{\theta}^0} \quad (3.10)$$

In this approximation, the formal error on the parameter  $\theta_i$  is given by

$$\sigma_i^F = \sqrt{(F^{-1})_{ii}} \quad (3.11)$$

For this simulator with known likelihood, the Fisher matrix can be computed:

$$F_{ij} = \sum_{\ell=2}^{\ell_{\max}} \sum_{PP',QQ'} \frac{\partial C_\ell^{PP'}}{\partial \theta_i} \left( \text{Cov}_{C_\ell}^{-1} \right) \frac{\partial C_\ell^{QQ'}}{\partial \theta_j} \quad (3.12)$$

---

<sup>11</sup>For this experiment, we use Asimov data, *i.e.*  $\hat{C}_\ell^{PP'} = \bar{C}_\ell^{PP'}$ , as is customary for forecasting. Since the likelihood for  $\bar{C}_\ell^{PP'}$  is non-Gaussian at low multipole, we rescale these modes accordingly for the fiducial data. This is mainly to show that the typical forecasting workflow is possible. In later sections the fiducial data includes non-trivial noise contributions.

where  $\text{Cov}_{C_\ell}$  is given by (3.8) (including a factor of  $f_{\text{sky}}$ ). We compute the Fisher matrix using a two-sided numerical approximation for the derivatives  $\partial_{\theta_i} C_\ell^{PP'}$ . We use the resulting  $\sigma_i^F$  to define a uniform prior on each cosmological parameter  $\theta_i$

$$\theta_i \sim \mathcal{U}([\theta_i^0 - 5\sigma_i^F, \theta_i^0 + 5\sigma_i^F]) \quad (3.13)$$

The final ingredient to our setup is a compression network. Since there are only 6 varying parameters in our model, it is reasonable to expect that the data can be optimally compressed into 6 features. When the likelihood is known, the “score function”  $\vec{t} = \nabla_{\boldsymbol{\theta}} \ln \mathcal{L}(\hat{C}(\boldsymbol{\theta})|\boldsymbol{\theta})$  compresses the data into one summary per parameter, such that the Fisher information is preserved [19, 67]. In the case of a Gaussian likelihood, score compression is equivalent to MOPED [68] (if  $\boldsymbol{\theta}$  determines the mean) or the optimal quadratic estimator [69] (if  $\boldsymbol{\theta}$  determines the covariance). If a likelihood is not known, neural compression schemes can be trained to find reduced representations of the data that maximize information content [59].

We choose to remain agnostic about the compression, apart from noting that within our prior, the power spectra  $\bar{C}_\ell^{PP'}$  are well-described by a linear approximation in the cosmological parameters.<sup>12</sup> We therefore use as a compression network a linear map from 7497 to 15 features. We use 15 features since this allows more capacity for the network without significantly increasing computational cost.

We use TMNRE to train *every* 1- and 2-dimensional marginal posterior using  $N_{\text{sim}} \leq 5000$  simulations. Training all of the relevant posteriors in parallel takes about 10 minutes on our GPU. Additionally, we note that the same 5000 simulations are used to train every posterior. These results are shown in Fig. 8, as well as posteriors derived via MCMC for  $N_{\text{sim}} \leq 10^5$  simulator calls. The MCMC chain had an acceptance rate of  $\sim 0.3$ , so that  $10^5$  simulator calls were required to produce  $\sim 3 \times 10^4$  samples. MCMC, which *samples* the *joint* (6-dimensional) posterior, requires this many samples in order to produce converged  $2\sigma$  contours, while TMNRE, which *learns* the *marginal* posteriors, produces results in excellent agreement at significantly reduced simulator cost.<sup>13</sup> Moreover, the number of simulations given to TMNRE can be decreased without significantly impacting the quality of the posteriors; we found that at  $N_{\text{sim}} = 3000$  simulations, TMNRE produces 1-dimensional marginal posteriors whose central values and  $1\sigma$  credible regions were accurate to within  $0.05\sigma$  of the ground truth.

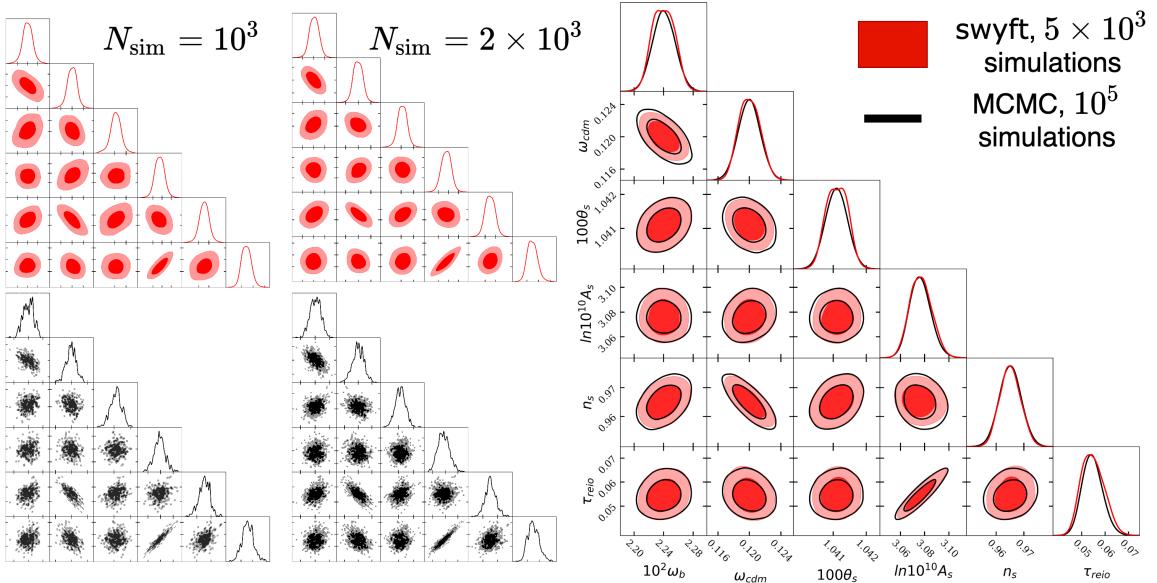
### 3.2 Comparing experimental configurations

It is often of interest to study how constraints on parameters (or from a forecasting perspective, expected sensitivities) change when different combinations of experiments are considered. Given the large number of upcoming cosmological experiments at any given time, the number of relevant experimental configurations can be daunting (e.g. 30 configurations of CMB and LSS experiments for forecasting constraints on neutrino masses in [62]). In the context of

---

<sup>12</sup>This is not strictly necessary for using a linear compression, since the subsequent classifier can learn non-linearities of the posteriors. TMNRE accurately reproduces non-Gaussian (i.e. non-linear) features in the posterior of this simulator.

<sup>13</sup>A common criterion for the convergence of MCMC results is that the Gelman-Rubin statistic  $\hat{R} - 1$  is sufficiently small [70], for example  $< 0.01$ . We employ this as a necessary condition for converged results. In addition, we require that 1- and  $2\sigma$  credible regions are visually converged when the kernel smoothing used to produce plots is fixed. Note that this sometimes means running our chains beyond  $\hat{R} - 1 < 0.01$  (in particular for high-dimensional examples). We use the same kernel smoothing to produce plots for both MCMC and TMNRE; the advantage to TMNRE is that once trained (on a potentially small number of simulations), a very large number of samples may be drawn from the variational posterior.



**Figure 8:** Left: convergence of marginal posteriors for TMNRE vs. MCMC for the CMB forecasting simulator. Since MCMC must sample the full joint posterior, it is slow to converge. On the other hand, TMNRE directly learns the marginal posteriors, and begins to converge quickly. Right: comparison of TMNRE at  $5 \times 10^3$  simulation to ground-truth MCMC with  $10^5$  simulations. TMNRE agrees with the ground-truth to excellent precision and at significantly reduced simulator cost. Note that tick mark values are the same for all plots.

MCMC, each experimental configuration must be treated with its own chain(s) in order to preserve Markov properties. The situation, however, is significantly improved for TMNRE.

### Re-using simulations allows for additional inferences at zero simulator cost.

This opens the opportunity to perform massive global scans with significantly reduced simulator cost. As demonstrated in the previous section, for a single inference TMNRE can be orders of magnitude more simulator-efficient than MCMC. By reusing simulations across inferences, TMNRE can gain another order of magnitude of simulator-efficiency relative to sampling methods.

More explicitly, for MCMC, the simulator cost for generating posteriors for  $N_{\text{config}}$  experimental configurations scales as  $N_{\text{config}} N_{\text{sim},\text{MCMC}}$ , where  $N_{\text{sim},\text{MCMC}}$  is the typical number of simulator calls for MCMC to converge for an individual experimental configuration. In the case where simulations can be reused, the computational cost for TMNRE scales as  $N_{\text{sim},\text{TMNRE}} \ll N_{\text{sim},\text{MCMC}}$ .

There are two alternative approaches for reusing simulations in the standard framework. One is the reweighting of MCMC chains via an alternative likelihood. One drawback of this approach is that deviations beyond 1- or 2- $\sigma$  can not be addressed. Moreover, reweighting requires either recomputing the data (in which case one might as well run a new chain) or having these data saved. As we have seen, MCMC requires orders of magnitude more simulator calls than TMNRE, so storing these simulations quickly becomes a burden.

A second approach is to use simulations to train emulators, enabling ultra-fast likelihood evaluations [55, 60, 71–73]. Alternatively, one may pair expensive and cheap simulations or

models to construct approximate likelihoods in a simulation-efficient way [74, 75]. While these approaches are also attractive, there are a few drawbacks. In particular, these approaches are generally possible only for restricted summary statistics such as power spectra. Even when neural networks are trained to output *e.g.* full boxes of galaxy surveys [76–78], these methods are typically validated on low-order or otherwise restricted information; the degree to which higher-order information can be trusted (and used as input for inference) is more limited. Even in the case of a power spectrum analysis, when a new cosmological model is developed, it is more efficient to do inference with TMNRE before training an entire emulator.

We demonstrate the principle of simulation reuse across different experimental configurations in Fig. 9. We first use TMNRE to compute posteriors for the CMB forecasting simulator with  $N_{\text{sim}} = 10^4$  simulations. Then we use the *same* simulation bank to train TMNRE for an alternative experimental configuration corresponding to the combination of CMB and BAO data. To generate the BAO data, when performing simulations for the first inference, we include a simulator component corresponding to BAO measurements at  $z = 0.106, 0.32, 0.57, 0.15$ , with sensitivities proportional to the 6dF Galaxy Survey [79], Sloan Digital Sky Survey (SDSS-III) Baryon Oscillation Spectroscopic Survey (BOSS) data release 11 and CMASS data [80], and SDSS Main Galaxy Sample [81] respectively.<sup>14</sup><sup>15</sup> The simulator corresponding to these measurements takes the BAO value derived by CLASS and adds Gaussian noise inferred from experimental sensitivities. It should be clear from the previous section that likelihood approaches to LSS that use more information from the power spectra (*e.g.* the full shape of the galaxy power spectrum [82, 83]) can also be phrased as simulators.

We perform inference using the same settings as in Sec. 3.1, including a linear compression network that maps from 7497 (7501 for CMB+BAO) to 15 features. As in the previous section, for *each* inference MCMC requires  $N_{\text{sim}, \text{MCMC}} = \mathcal{O}(10^5)$  simulator calls for converged  $2\sigma$  contours, so that for  $N_{\text{config}} = 2$  experimental configurations, the total cost is  $2 \times 10^5$ . On the other hand, after the first inference, TMNRE requires *zero* new simulator calls, so that the total simulator cost saturates at the single-inference value, in this case  $\mathcal{O}(10^4)$ . This logic clearly extends to the case of more experimental configurations, the number of which often exceeds an order of magnitude for realistic cosmological inferences.

### 3.3 Realistic Planck simulator

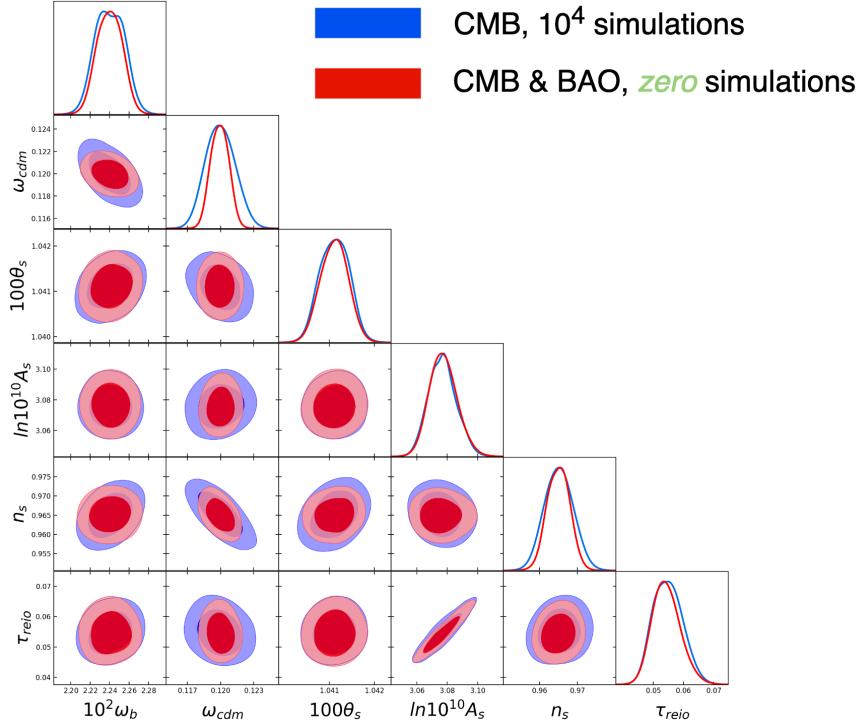
Classical inference becomes more difficult as the instrument model becomes more realistic. This is because realistic instrument models involve extra parameters related to the calibration of the instrument, foreground residuals, etc. These quantities, which are not interesting to the scientific problem at hand, are called *nuisance parameters*. Sampling cost for the joint posterior scales poorly with the addition of extra parameters. On the other hand, TMNRE is not significantly affected by nuisance parameters.

#### Directly learning marginal posteriors alleviates sampling problems associated with the curse of dimensionality.

---

<sup>14</sup>We shrink the noise uncertainties by a factor of two for the sake of visualization.

<sup>15</sup>The same logic applies to considering an upgraded CMB experiment. In this case, we may simply change the simulator’s noise model. Note that it is efficient to save the “noise-free” simulations and only add noise later, when performing the inference. Often this second step is much quicker than the initial simulation.



**Figure 9:** Demonstration of simulation reuse with TMNRE. We infer posteriors corresponding to the forecasting CMB simulator of Sec. 3.1 with  $N_{\text{sim}} = 10^4$  simulations. We subsequently use the *same* simulations to infer posteriors for the combination of CMB and BAO data.

We now turn to `HiLLiPoP`<sup>16</sup> [84], which is a multifrequency CMB likelihood for Planck data. In particular, `HiLLiPoP` treats high- $\ell$  modes within a Gaussian approximation, including cross-correlation between Planck’s 100, 143, and 217 Ghz split-frequency maps. The model covers multipoles from  $\ell = 30$  to  $\ell = 2500$ , and includes contributions from galactic dust, the cosmic infrared background, thermal and kinetic Sunyaev-Zeldovich emission, and point sources.

Since the likelihood is Gaussian, it can be naturally rephrased in simulator language, as in Sec. 3.1. Namely, `HiLLiPoP` computes the likelihood using a  $\chi^2$

$$\chi^2 = (d - \mu)^T C^{-1} (d - \mu) \quad (3.14)$$

where  $d$  is computed by binning, adding foregrounds, etc. to a spectrum computed from a Boltzmann code and  $\mu$  is the observed data. We therefore peel off  $d$  from the likelihood and use this as the output of our simulator, drawing noise from the covariance matrix  $C$  as needed.

We use the temperature autocorrelation part of the simulator, which has 10816 bins and 13 varying nuisance parameters in addition to 6 cosmological parameters. We choose this likelihood because (i) it has many nuisance parameters (ii) it has a native Python implementation.<sup>17</sup>

---

<sup>16</sup><https://github.com/planck-npipe/hillipop>

<sup>17</sup>Among other Planck likelihoods, `plik` [85] has a similar number of nuisance parameters but no native Python implementation. The nuisance-marginalized version of this likelihood is available in Python as `plik_lite` via Cobaya [86] or [87]. We have checked that inference for the simulator corresponding to this likelihood is as roughly as easy with TMNRE as the CMB forecasting simulator of Sec. 3.1.

The model takes 6 cosmological parameters and 13 varying nuisance parameters. For priors on the cosmological parameters, we use uniform distributions with  $\pm 5\sigma$  from Planck 2018 [88]. Since this is a high-multipole simulator, we impose a Gaussian prior on the optical depth,  $\tau = 0.0544 \pm 0.0073$ , corresponding to Planck low-multipole constraints.<sup>18</sup> For priors on the nuisance parameters, we use the recommendations shipped with `HILLiPoP`: Gaussian priors for all 6 varying calibration parameters (with  $\sigma = 0.0025$  for  $A_{\text{Planck}}$  and  $\sigma = 0.002$  for varying  $c_X$ ) and Gaussian priors (with  $\sigma = 0.2$ ) for all foreground amplitudes other than  $A_{\text{kSZ}}$  and  $A_{\text{SZ} \times \text{CIB}}$ , which are both uniformly drawn from the interval  $[0, 10]$ .<sup>19</sup>

In order to have control over our setup, we infer posteriors for a fiducial spectrum generated by our simulator rather than Planck data. For all parameters with Gaussian priors except  $\tau$  and  $A_{\text{radio}}$ , we displace the fiducial parameter by  $-0.5\sigma$  from the mode. For parameters with uniform priors in the interval  $[l, h]$ , we set the parameter equal to  $l + \frac{3(h-l)}{10}$ . For  $\tau$  and  $A_{\text{radio}}$  (where the difference between the prior and posterior is most dramatic in [84]), we displace the fiducial parameter from the mode of the prior by  $+2\sigma$ .

With this prior, we draw  $N_{\text{sim}} = 3000$  simulations and infer posteriors for 6 cosmological parameters using a linear compression network from 10816 to 10 features. We show posteriors for the cosmological parameters in Fig. 10. With only  $N_{\text{sim}} = 3000$  simulations, TMNRE learns accurate 1- and 2-dimensional marginal posteriors, while MCMC requires  $\mathcal{O}(5 \times 10^6)$  simulator calls for converged 1- and 2- $\sigma$  contours (cf. footnote 13). While the simulator cost for MCMC is significantly larger due to the addition of nuisance parameters, TMNRE directly learns marginal posteriors of interest and is not afflicted by the curse of dimensionality.

### 3.4 Neutrino masses

As a final application, we turn to neutrino masses, which provide a cosmologically-relevant instance of significantly non-Gaussian posteriors. We use this example to show that for TMNRE:

#### Non-Gaussian posteriors are a non-problem.

The existence of a cosmological background of relic neutrinos is a standard prediction of the  $\Lambda$ CDM cosmology. Observations of neutrino oscillations require at least two of the three neutrinos to be massive (with the sum of the three masses  $\sum m_\nu \gtrsim 0.06$  eV) [90–92] – as such, a majority of the energy density stored today in the cosmic neutrino background (CNB) is expected to be in the form of non-relativistic neutrinos. The transition from relativistic to non-relativistic neutrinos modifies the characteristic free-streaming length, the physical effect of which is to damp the growth of small scale perturbations. From an observational perspective, the presence of neutrino masses appears as a suppression of the matter power spectrum on scales  $k \gtrsim k_{nr}$ , where  $k_{nr} \sim 0.018 \Omega_m \sqrt{m/1 \text{ eV}} h \text{ Mpc}^{-1}$  is related to the comoving scale at which the transition occurs (see e.g. [93–96]).

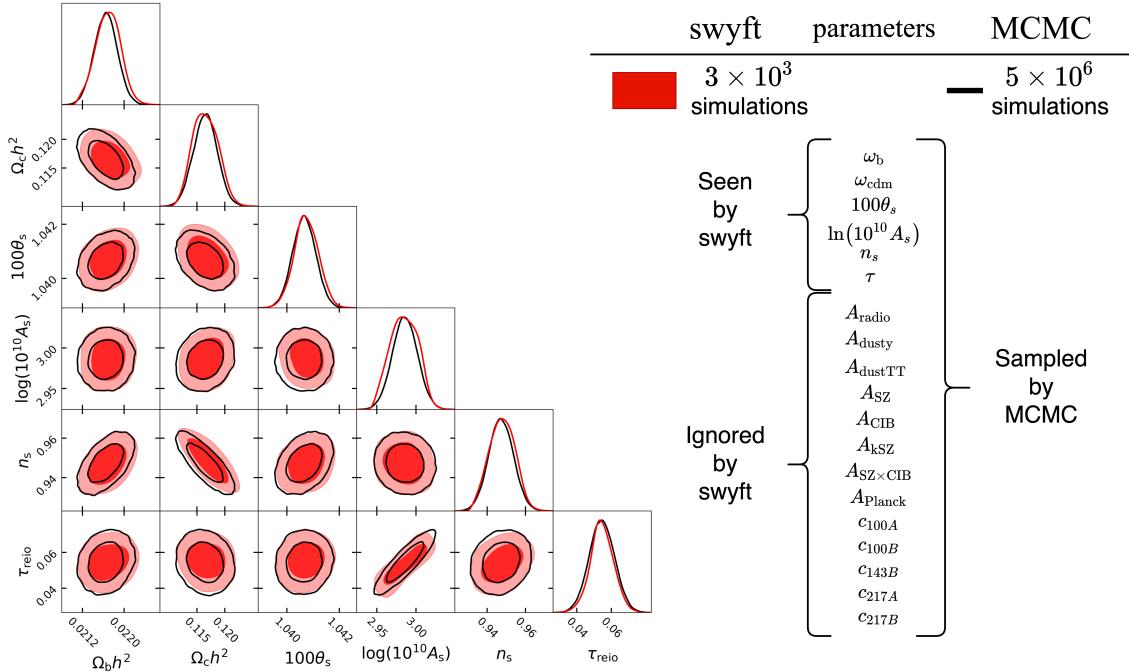
At first order, cosmology is only sensitive to the net sum of neutrino masses, placing constraints today at the level of  $\sum m_\nu \lesssim 0.12$  eV [88] (to be compared with terrestrial experiments, which currently have constrained  $\sum m_\nu \lesssim 0.9$  eV [97]). Terrestrial experiments<sup>20</sup> on the other hand, are typically most sensitive to the relative differences in neutrino masses,

---

<sup>18</sup>For progress towards simulation-based inference for low- $\ell$  modes, see [89].

<sup>19</sup>See [https://github.com/planck-npipe/hillipop/blob/master/examples/hillipop\\_example.yaml](https://github.com/planck-npipe/hillipop/blob/master/examples/hillipop_example.yaml) for the full list of priors.

<sup>20</sup>The best terrestrial sensitivity to the sum of neutrino masses comes from tritium beta decay experiments such as KATRIN [98].



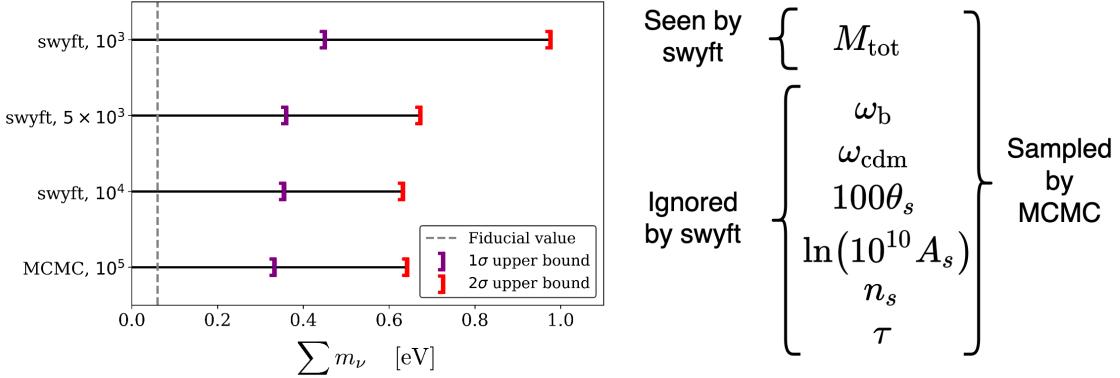
**Figure 10:** 1- and 2-dimensional posteriors inferred by TMNRE trained on  $N_{\text{sim}} = 3000$  realistic simulations, corresponding to the Planck HiLLiPoP likelihood. In addition to the 6 cosmological parameters shown, there are 13 varying nuisance parameters. While TMNRE ignores irrelevant parameters when learning posteriors, MCMC must sample each parameter. Converged  $2\sigma$ -contours for MCMC require  $\mathcal{O}(5 \times 10^6)$  simulations.

their mixings, and possible CP violating phases [90–92]; this implies that cosmological probes of the small scale structure of the universe offer a highly complementary route to testing the properties of neutrinos. Moreover, it seems likely that cosmology will become sensitive to the lower mass limit (*i.e.*  $\sum m_\nu \sim 0.06$  eV) long before their terrestrial counterpart [94] (although it is worth emphasizing that the cosmological measurements are sensitive to exotic physics, particularly those that can modify the neutrino phase space distribution). Given the importance of a cosmological neutrino mass measurement, we investigate a final application of the methods outlined here by forecasting the sensitivity of future experiments. This in particular demonstrates TMNRE’s ability to estimate non-Gaussian posteriors. (For the demonstration of this algorithm on more exotic posteriors, see [16, 17].)

Here we consider the simulator of Sec. 3.1, with the cosmology extended to include massive neutrinos. We include three neutrinos degenerate in mass, each with mass  $m_\nu = M_{\text{tot}}/3$ . Results from the inference are shown in Fig. 11.

In particular, we show the 1- and  $2\sigma$  upper bounds for  $M_{\text{tot}}$  as computed by TMNRE with various simulator budget. As our fiducial cosmology we use the same parameters as in Sec. 3.1 and  $M_{\text{tot}} = 0.06$  eV. For TMNRE we use  $N_{\text{sim}} \leq 10^4$  simulations drawn from a prior determined, as in Sec. 3.1 using a Fisher matrix formalism and a uniform distribution over  $\pm 5\sigma$  ranges. It is well-known (see e.g. [63]) that the Fisher matrix gives a conservative (*i.e.* too wide) estimate of the posteriors in this case. Regardless, TMNRE produces accurate posteriors with orders of magnitude fewer simulator calls than the ground-truth MCMC, which requires  $\mathcal{O}(10^5)$  simulations for convergence (cf. footnote 13). Already at  $N_{\text{sim}} = 5000$ , the

$2\sigma$  upper bound computed by TMNRE agrees with the long-run MCMC result to within 3%.



**Figure 11:** Upper bounds on  $M_{\text{tot}}$  as estimated by TMNRE for various simulator cost. We see that already at  $N_{\text{sim}} = 5000$ , TMNRE accurately approximates the long-run ( $10^5$  simulations) MCMC result. TMNRE directly learns the marginal posterior for  $M_{\text{tot}}$  and thus circumvents the sampling cost associated with  $\Lambda\text{CDM}$  parameters.

### 3.5 Testing statistical consistency

So far we have demonstrated the accuracy of TMNRE results by comparing against long-run MCMC chains. Since one of our points is that TMNRE avoids the simulator cost associated with MCMC, we would like a way to test the consistency of TMNRE results without referring to MCMC. This is especially important for inferring parameters when a likelihood is not available and the ground-truth is unknown. We now show how consistency checks that are inaccessible to sampling-based methods are possible with TMNRE.

**Statistical under/overconfidence can be cheaply assessed using local amortization.**

In particular, we perform the test defined in Sec. 2.3 for the marginal posteriors trained in Sec. 3.1. We show the results in Fig. 12. We see that the empirical coverage and confidence level generally match to excellent precision. Even if they did not match, this information could be used to modify reported constraints compared to network predictions. This test makes no reference to the true parameter values of the observed data or long-run MCMC results, and therefore will prove very important in the inference of observational data with unknown true parameters and implicit likelihood.

### 3.6 Wide priors

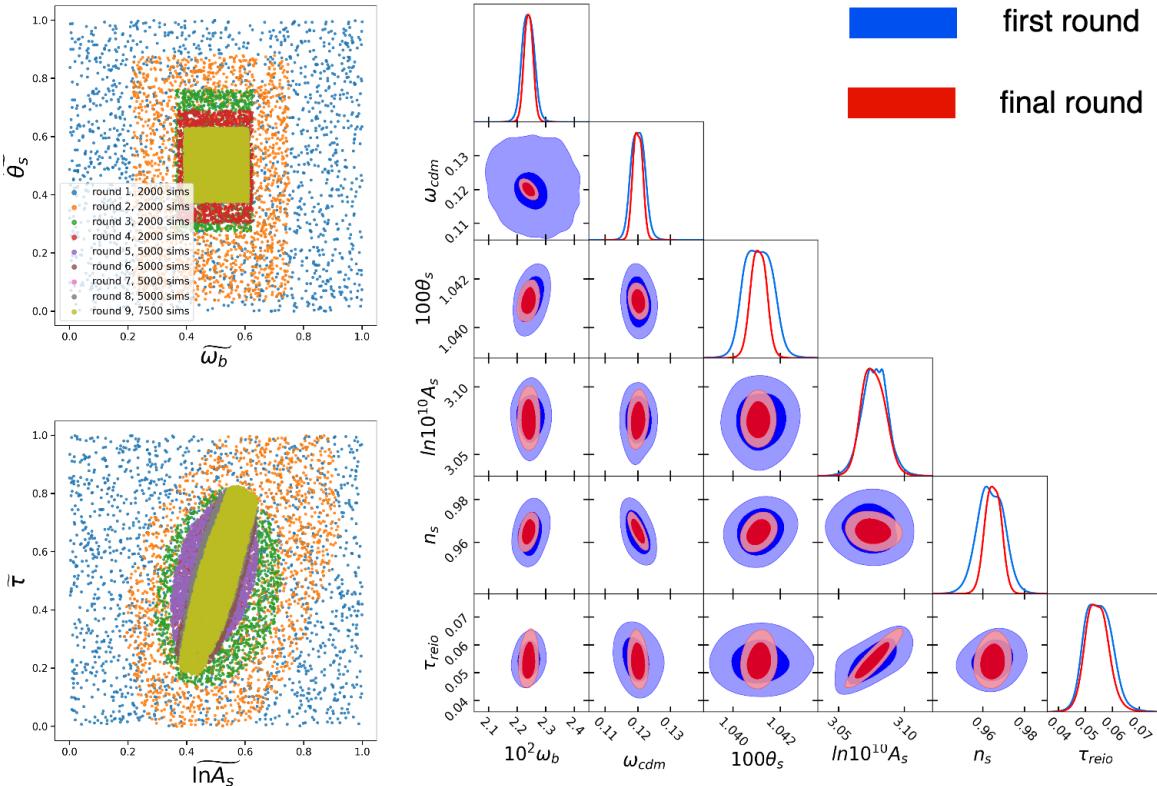
In the previous sections we have dealt with reasonably narrow priors, estimated either via a Fisher matrix formalism or related observational constraints. However, there are some situations in which the prior is much wider than the support of the posterior, for example in the context of bias parameters in large-scale structure [99]. In this section we demonstrate how wide priors can be accommodated by using multiple rounds of training and truncation with TMNRE:

**Truncation efficiently identifies the parameter region of interest.**



**Figure 12:** Local amortization allows for cheap consistency checks. Here we show the results of our coverage test applied to all 1- and 2-dimensional posteriors for the network trained in Sec. 3.1, for which the posteriors are shown in Fig. 8.

Examples outside the cosmological context can be found in [16, 17]. As a concrete test, we consider the CMB forecasting simulator from Sec. 3.1. This time, each cosmological parameter is drawn from a uniform prior with  $\theta_i \in [\theta_i^0 - 25\sigma_F^i, \theta_i^0 + 25\sigma_F^i]$  with  $\sigma_F$  the Fisher estimate, except for  $\tau$ , for which  $\pm 8\sigma_F$  is used to stay within the physical range. The volume of this prior is 5000 times larger than that in Sec. 3.1. We perform multiple rounds of training with the same settings as in Sec. 3.1. When truncating, we take into account correlations for the parameter pairs  $(\omega_{cdm}, n_s)$  and  $(\ln(10^{10} A_s), \tau)$  respectively, with no correlations considered for  $\omega_b$  and  $100\theta_s$ . Starting from 2000 simulations in the first round, the amount of training data was increased every time the volume of the truncated prior shrunk by a factor



**Figure 13:** Left: evolution of training data over 9 rounds. Parameters are shown scaled to the unit hypercube. Using expected correlations from the Fisher matrix,  $\ln 10^{10} A_s$  and  $\tau$  are truncated with correlation to each other, as are  $\omega_{cdm}$  and  $n_s$  (not shown).  $\omega_b$  and  $100\theta_s$  are truncated independently. Whenever the volume of the truncated prior remains roughly constant between rounds, the amount of training data is increased. Right: comparison of posteriors from the first and last rounds.

of less than 5. Over 9 rounds, 20,000 simulations are used, a factor of 4 increase from the inference with “reasonable” prior in Sec. 3.1. Some simulations are reused between rounds.<sup>21</sup> The evolution of the training data as well as posteriors from the first and last rounds are shown in Fig. 13. After the sixth round of training (with  $10^4$  simulations seen by the network), the volume of the truncated prior has shrunk by a factor of 5000. In other words, truncation efficiently identifies the parameter region of interest.

## 4 Conclusions and outlook

In this work we have illustrated how a novel technique in simulation-based inference (SBI) called Truncated Marginal Neural Ratio Estimation (TMNRE, as implemented in the open-source code `swyft`<sup>22</sup>) can be used to improve cosmological parameter inference. In particular,

<sup>21</sup>We ran this experiment with 10,000 simulations in the initial round, and convergence was also achieved at a total simulation cost of roughly 20,000, in this case distributed over three rounds of training.

<sup>22</sup><https://github.com/undark-lab/swyft>

we have argued that TMNRE has several advantages over conventional sampling-based approaches such as MCMC, including:

- **Efficiency.** TMNRE generically requires significantly fewer simulator calls than traditional sampling-based methods simply because it targets the 1- and 2-dimensional marginal posteriors rather than the full joint posterior. In addition, one can alter the priors, the observables, or the inference network without having to re-run the simulator from scratch; in Sec. 3.2, we illustrated that adding BAO data to the network that was generated using data from a Planck-like simulator required *zero* additional calls to the simulator.
- **Scalability.** Nuisance parameters are commonplace in cosmology, and tend to significantly impede the convergence timescale of MCMC. TMNRE is designed in part to soften this scaling with respect to the number of nuisance parameters introduced; we present here several examples where parameters “ignored” by the network do not increase the number of simulator calls required for accurate and converged posteriors. This is one of the major advantages offered by TMNRE over other SBI approaches.
- **Trustability.** One the most attractive features of TMNRE is that the reliability of the network can be directly tested via local amortization, which effectively amounts to a rigorous check of the coverage of the Bayesian credible interval. This type of test is crucial for establishing that TMNRE is trustable, offering a major advantage over MCMC and some other forms of SBI.

We have illustrated the power of this method using the CMB power spectra and BAO, showing TMNRE generates perfect Bayesian confidence intervals for the  $\Lambda$ CDM parameters with orders of magnitude less computation time than traditional algorithms. Despite the fact that TMNRE (and more generally SBI) provide their biggest improvements when the likelihood is unknown and simulations are computationally expensive, we have shown that these methods can be powerful even when in circumstances that are more favorable for conventional methods. In particular, we believe that TMNRE offers cosmologists a novel computational tool that mitigates the need for approximation schemes in extended cosmologies (such as those involving massive neutrinos, interacting species, and non-standard distribution functions).

These aspects bode well for extracting cosmological information from future surveys. While in this paper we have worked with simulators for CMB summary statistics, recent work has shown that density-based SBI performs well when presented with field-level data [11–13] (and see [100] for a comparison with field-level likelihood methods). We note that neural ratio estimation, which scales favorably compared to density estimation, has not yet been applied to field-level analysis. There is no theoretical obstruction to using SBI for full high-resolution simulations; such an approach will likely be crucial to using future data to understand our universe.

To make contact with observational data, it is therefore of high importance for future cosmology experiments to release simulators (which incorporate *e.g.* systematic effects) which are capable of reproducing their mock data. Importantly, this does not necessitate a major philosophical shift, since most experiments already release their likelihoods. Moreover, experimental collaborations have already done the hard work of calibrating and validating their simulators. Moving to realistic simulators will not only allow for significant progress in the field of SBI, but it will also improve the extent to which we can extract cosmological information and build cosmological models.

## Acknowledgments and code availability

We thank Justin Alsing and Ben Wandelt for very useful discussions about the comparison between PYDELFI and TMNRE, and Justin Alsing, Matteo Biagetti, Sascha Caron, Gilles Louuppe and Ben Wandelt for comments on a draft of this paper. Some of the computations shown were performed using TitanX GPUs on The Distributed ASCI Supercomputer 5 (DAS-5) [101]. DAS-5 is funded by the NWO/NCF (the Netherlands Organization for Scientific Research). The project has been partially funded by the Netherlands eScience Center and SURF under grant number ETEC.2019.018. This work is part of a project that has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant agreement No. 864035).

Software used in this work includes `numpy` [102], `scipy` [103], `matplotlib` [104], `pytorch` [105], and `jupyter` [106].

Accompanying code for the examples in this paper can be found at <https://github.com/a-e-cole/swyft-CMB>.

## A Explicit sky mask

In Sec. 3.1 we incorporated partial sky coverage by modifying the effective number of degrees of freedom with  $f_{\text{sky}}$ . In this appendix, we incorporate an explicit sky mask, which generally makes a likelihood-based analysis significantly more complicated, since different modes are coupled (see e.g. [85]). On the other hand, in the context of SBI incorporating an explicit sky mask is straightforward: we may simply replace masked pixels with some fixed value and compute the corresponding maximum likelihood power spectrum. Here we set masked pixels to 0. Note that in this context the power spectrum is no longer a sufficient statistic. However, treating CMB analysis at the pixel level at sufficiently high-resolutions will require technical progress beyond the scope of this work.

As a toy example, we infer cosmology using the  $TT$  power spectrum mimicking a single frequency channel with WMAP systematics. Namely, we compute the power spectrum up to  $\ell_{\text{max}} = 1024$ , employ a Gaussian beam with  $FWHM = 0.008$  radians and white noise with  $\sigma_T = 35 \mu\text{K}$ . We mask the data using the WMAP 7-year temperature mask<sup>23</sup>, setting these pixel values to 0, and compute the corresponding maximum likelihood power spectrum.

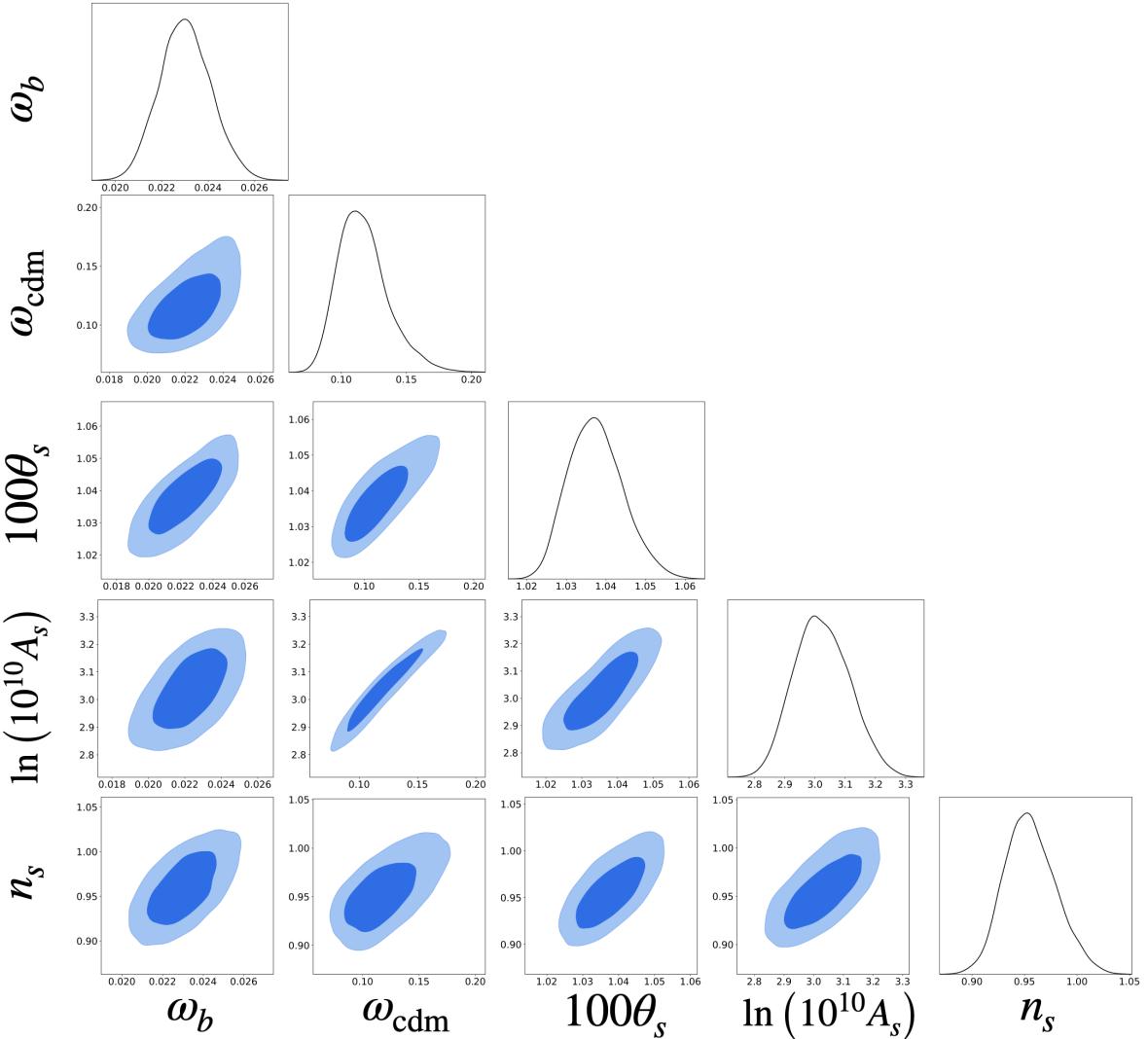
For this simulator, we infer  $\Lambda\text{CDM}$  parameters with  $\tau = 0.054$  fixed. We take as our prior for the remaining parameters

$$\begin{aligned}\omega_b &\sim \mathcal{U}[0.0116, 0.0332] \\ \omega_{\text{cdm}} &\sim \mathcal{U}[0.0206, 0.2193] \\ 100\theta_s &\sim \mathcal{U}[1.017, 1.065] \\ \ln(10^{10} A_s) &\sim \mathcal{U}[2.469, 3.681] \\ n_s &\sim \mathcal{U}[0.687, 1.243]\end{aligned}$$

which is wide for all parameters except for  $100\theta_s$ . Fiducial parameters are the same as those in Sec. 3.1. We draw 20,000 simulations from this prior and train the 1- and 2-dimensional posteriors, employing a linear embedding layer that compresses the data from 1024 to 10 features. The results are shown in Fig. 14.

---

<sup>23</sup>[http://lambda.gsfc.nasa.gov/data/map/dr4/ancillary/masks/wmap\\_temperature\\_analysis\\_mask\\_r9\\_7yr\\_v4.fits](http://lambda.gsfc.nasa.gov/data/map/dr4/ancillary/masks/wmap_temperature_analysis_mask_r9_7yr_v4.fits)



**Figure 14:** One- and two-dimensional posteriors for the power spectrum of an explicitly masked sky.

## References

- [1] E. Di Valentino et al., *Snowmass2021 - Letter of interest cosmology intertwined I: Perspectives for the next decade*, *Astropart. Phys.* **131** (2021) 102606 [[2008.11283](https://arxiv.org/abs/2008.11283)].
- [2] D.B. Rubin, *Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician*, *The Annals of Statistics* **12** (1984) 1151 .
- [3] S. Tavaré, D.J. Balding, R.C. Griffiths and P. Donnelly, *Inferring Coalescence Times From DNA Sequence Data*, *Genetics* **145** (1997) 505 [<https://academic.oup.com/genetics/article-pdf/145/2/505/35202057/genetics0505.pdf>].
- [4] P.J. Diggle and R.J. Gratton, *Monte carlo methods of inference for implicit statistical models*, *Journal of the Royal Statistical Society: Series B (Methodological)* **46** (1984) 193.
- [5] K. Cranmer, J. Brehmer and G. Louppe, *The frontier of simulation-based inference*, *Proc. Natl. Acad. Sci. U. S. A.* (2020) .

- [6] J. Hermans, V. Begy and G. Louppe, *Likelihood-free mcmc with amortized approximate ratio estimators*, in *International Conference on Machine Learning*, pp. 4239–4248, PMLR, 2020.
- [7] G. Papamakarios, D. Sterratt and I. Murray, *Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows*, in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 837–848, PMLR, 2019.
- [8] D. Greenberg, M. Nonnenmacher and J. Macke, *Automatic posterior transformation for likelihood-free inference*, in *International Conference on Machine Learning*, pp. 2404–2414, PMLR, 2019.
- [9] J. Alsing, T. Charnock, S. Feeney and B. Wandelt, *Fast likelihood-free cosmology with neural density estimators and active learning*, *Mon. Not. Roy. Astron. Soc.* **488** (2019) 4440 [[1903.00007](#)].
- [10] X. Zhao, Y. Mao, C. Cheng and B.D. Wandelt, *Simulation-Based Inference of Reionization Parameters From 3D Tomographic 21 cm Images*, [2105.03344](#).
- [11] T.L. Makinen, T. Charnock, J. Alsing and B.D. Wandelt, *Lossless, Scalable Implicit Likelihood Inference for Cosmological Fields*, [2107.07405](#).
- [12] F. Villaescusa-Navarro et al., *Multifield Cosmology with Artificial Intelligence*, [2109.09747](#).
- [13] F. Villaescusa-Navarro et al., *Robust marginalization of baryonic effects for cosmological inference at the field level*, [2109.10360](#).
- [14] M. Dax, S.R. Green, J. Gair, J.H. Macke, A. Buonanno and B. Schölkopf, *Real-time gravitational-wave science with neural posterior estimation*, *arXiv preprint arXiv:2106.12594* (2021) .
- [15] A. Delaunoy, A. Wehenkel, T. Hinderer, S. Nissanke, C. Weniger, A.R. Williamson et al., *Lightning-fast gravitational wave parameter inference through neural amortization*, *arXiv preprint arXiv:2010.12931* (2020) .
- [16] B.K. Miller, A. Cole, G. Louppe and C. Weniger, *Simulation-efficient marginal posterior estimation with swyft: stop wasting your precious time*, in *34th Conference on Neural Information Processing Systems*, 11, 2020 [[2011.13951](#)].
- [17] B.K. Miller, A. Cole, P. Forré, G. Louppe and C. Weniger, *Truncated Marginal Neural Ratio Estimation*, in *35th Conference on Neural Information Processing Systems*, 7, 2021, DOI [[2107.01214](#)].
- [18] J. Hermans, A. Delaunoy, F. Rozet, A. Wehenkel and G. Louppe, *Averting a crisis in simulation-based inference*, *arXiv preprint arXiv:2110.06581* (2021) .
- [19] J. Alsing, B. Wandelt and S. Feeney, *Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology*, *Mon. Not. Roy. Astron. Soc.* **477** (2018) 2874 [[1801.01497](#)].
- [20] J. Alsing and B. Wandelt, *Nuisance hardened data compression for fast likelihood-free inference*, *Mon. Not. Roy. Astron. Soc.* **488** (2019) 5093 [[1903.01473](#)].
- [21] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, *Equation of state calculations by fast computing machines*, *The journal of chemical physics* **21** (1953) 1087.
- [22] W.K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, *Biometrika* **57** (1970) 97 [<https://academic.oup.com/biomet/article-pdf/57/1/97/23940249/57-1-97.pdf>].
- [23] P.J. Green, *Reversible jump markov chain monte carlo computation and bayesian model determination*, *Biometrika* **82** (1995) 711.
- [24] J. Skilling, *Nested sampling*, in *AIP Conference Proceedings*, vol. 735, pp. 395–405, American Institute of Physics, 2004.

- [25] J. Skilling, *Nested sampling for general bayesian computation*, *Bayesian analysis* **1** (2006) 833.
- [26] F. Feroz, M.P. Hobson and M. Bridges, *Multinest: an efficient and robust bayesian inference tool for cosmology and particle physics*, *Monthly Notices of the Royal Astronomical Society* **398** (2009) 1601–1614.
- [27] W.J. Handley, M.P. Hobson and A.N. Lasenby, *polychord: next-generation nested sampling*, *Monthly Notices of the Royal Astronomical Society* **453** (2015) 4385–4399.
- [28] E. Higson, W. Handley, M. Hobson and A. Lasenby, *Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation*, *Statistics and Computing* **29** (2019) 891.
- [29] S.A. Sisson, Y. Fan and M. Beaumont, *Handbook of approximate Bayesian computation*, CRC Press (2018).
- [30] D.B. Rubin, *Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician*, *The Annals of Statistics* **12** (1984) 1151 .
- [31] S. Tavaré, D.J. Balding, R.C. Griffiths and P. Donnelly, *Inferring Coalescence Times From DNA Sequence Data*, *Genetics* **145** (1997) 505  
[\[https://academic.oup.com/genetics/article-pdf/145/2/505/35202057/genetics0505.pdf\]](https://academic.oup.com/genetics/article-pdf/145/2/505/35202057/genetics0505.pdf).
- [32] J.-M. Lueckmann, J. Boelts, D. Greenberg, P. Goncalves and J. Macke, *Benchmarking simulation-based inference*, in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, A. Banerjee and K. Fukumizu, eds., vol. 130 of *Proceedings of Machine Learning Research*, pp. 343–351, PMLR, 13–15 Apr, 2021,  
<http://proceedings.mlr.press/v130/lueckmann21a.html>.
- [33] G. Papamakarios and I. Murray, *Fast  $\epsilon$ -free inference of simulation models with bayesian conditional density estimation*, in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett, eds., vol. 29, Curran Associates, Inc., 2016, <https://proceedings.neurips.cc/paper/2016/file/6aca97005c68f1206823815f66102863-Paper.pdf>.
- [34] J.-M. Lueckmann, P.J. Gonçalves, G. Bassetto, K. Öcal, M. Nonnenmacher and J.H. Macke, *Flexible statistical inference for mechanistic models of neural dynamics*, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1289–1299, 2017.
- [35] C. Durkan, I. Murray and G. Papamakarios, *On contrastive learning for likelihood-free inference*, in *International Conference on Machine Learning*, pp. 2771–2781, PMLR, 2020.
- [36] N. Jeffrey and B.D. Wandelt, *Solving high-dimensional parameter inference: marginal posterior densities & Moment Networks*, in *34th Conference on Neural Information Processing Systems*, 11, 2020 [[2011.05991](https://arxiv.org/abs/2011.05991)].
- [37] C.M. Bishop, *Pattern recognition and machine learning (information science and statistics)*, 2006.
- [38] G. Papamakarios, T. Pavlakou and I. Murray, *Masked autoregressive flow for density estimation*, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 2335–2344, 2017.
- [39] G. Papamakarios, E. Nalisnick, D.J. Rezende, S. Mohamed and B. Lakshminarayanan, *Normalizing flows for probabilistic modeling and inference*, *Journal of Machine Learning Research* **22** (2021) 1.
- [40] K. Cranmer, J. Pavez and G. Louppe, *Approximating likelihood ratios with calibrated discriminative classifiers*, *arXiv preprint arXiv:1506.02169* (2015) .
- [41] K.P. Murphy, *Machine learning: a probabilistic perspective*, MIT press (2012).

- [42] F. Rozet and G. Louppe, *Arbitrary marginal neural ratio estimation for simulation-based inference*, arXiv preprint arXiv:2110.00449 (2021) .
- [43] <https://github.com/undark-lab/swyft>.
- [44] J.S. Speagle, *A Conceptual Introduction to Markov Chain Monte Carlo Methods*, arXiv e-prints (2019) arXiv:1909.12313 [[1909.12313](#)].
- [45] V. Roy, *Convergence diagnostics for markov chain monte carlo*, 2019.
- [46] M. Betancourt, *The convergence of markov chain monte carlo methods: From the metropolis method to hamiltonian monte carlo*, 2018.
- [47] L.D. Brown, T.T. Cai and A. DasGupta, *Interval Estimation for a Binomial Proportion*, *Statistical Science* **16** (2001) 101 .
- [48] N. Jeffrey, F. Boulanger, B.D. Wandelt, B.R.-S. Blancard, E. Ally and F. Levrier, *Single frequency CMB B-mode inference with realistic foregrounds from a single training image*, *Mon. Not. Roy. Astron. Soc.* **510** (2022) L1 [[2111.01138](#)].
- [49] J. Akeret, A. Refregier, A. Amara, S. Seehars and C. Hasner, *Approximate Bayesian Computation for Forward Modeling in Cosmology*, *JCAP* **08** (2015) 043 [[1504.07245](#)].
- [50] COIN collaboration, *COSMOABC: Likelihood-free inference via Population Monte Carlo Approximate Bayesian Computation*, *Astron. Comput.* **13** (2015) 1 [[1504.06129](#)].
- [51] E. Jennings and M. Madigan, *astroabc: an approximate bayesian computation sequential monte carlo sampler for cosmological parameter estimation*, *Astronomy and computing* **19** (2017) 16.
- [52] A. Mootoovaloo, A.F. Heavens, A.H. Jaffe and F. Leclercq, *Parameter Inference for Weak Lensing using Gaussian Processes and MOPED*, *Mon. Not. Roy. Astron. Soc.* **497** (2020) 2213 [[2005.06551](#)].
- [53] A. Heavens, E. Sellentin and A. Jaffe, *Extreme data compression while searching for new physics*, *Mon. Not. Roy. Astron. Soc.* **498** (2020) 3440 [[2006.06706](#)].
- [54] F. Leclercq, *Bayesian optimization for likelihood-free cosmological inference*, *Phys. Rev. D* **98** (2018) 063511 [[1805.07152](#)].
- [55] K.K. Rogers, H.V. Peiris, A. Pontzen, S. Bird, L. Verde and A. Font-Ribera, *Bayesian emulator optimisation for cosmology: application to the Lyman-alpha forest*, *JCAP* **02** (2019) 031 [[1812.04631](#)].
- [56] T. Takhtaganov, Z. Lukic, J. Müller and D. Morozov, *Cosmic Inference: Constraining Parameters with Observations and a Highly Limited Number of Simulations*, *Astrophys. J.* **906** (2021) 74 [[1905.07410](#)].
- [57] E. Sellentin, M. Quartin and L. Amendola, *Breaking the spell of Gaussianity: forecasting with higher order Fisher matrices*, *Mon. Not. Roy. Astron. Soc.* **441** (2014) 1831 [[1401.6892](#)].
- [58] F. Leclercq, W. Enzi, J. Jasche and A. Heavens, *Primordial power spectrum and cosmology from black-box galaxy surveys*, *Mon. Not. Roy. Astron. Soc.* **490** (2019) 4237 [[1902.10149](#)].
- [59] T. Charnock, G. Lavaux and B.D. Wandelt, *Automatic physical inference with information maximizing neural networks*, *Physical Review D* **97** (2018) .
- [60] A. Spurio Mancini, D. Piras, J. Alsing, B. Joachimi and M.P. Hobson, *CosmoPower : emulating cosmological power spectra for accelerated Bayesian inference from next-generation surveys*, [2106.03846](#).
- [61] A. Coogan, K. Karchev and C. Weniger, *Targeted Likelihood-Free Inference of Dark Matter Substructure in Strongly-Lensed Galaxies*, in *34th Conference on Neural Information Processing Systems*, 10, 2020 [[2010.07032](#)].

- [62] T. Brinckmann, D.C. Hooper, M. Archidiacono, J. Lesgourgues and T. Sprenger, *The promising future of a robust cosmological neutrino mass measurement*, *JCAP* **01** (2019) 059 [[1808.05955](#)].
- [63] L. Perotto, J. Lesgourgues, S. Hannestad, H. Tu and Y.Y.Y. Wong, *Probing cosmological parameters with the CMB: Forecasts from full Monte Carlo simulations*, *JCAP* **10** (2006) 013 [[astro-ph/0606227](#)].
- [64] D. Blas, J. Lesgourgues and T. Tram, *The Cosmic Linear Anisotropy Solving System (CLASS) II: Approximation schemes*, *JCAP* **07** (2011) 034 [[1104.2933](#)].
- [65] CORE collaboration, *Exploring cosmic origins with CORE: Cosmological parameters*, *JCAP* **04** (2018) 017 [[1612.00021](#)].
- [66] W.J. Percival and M.L. Brown, *Likelihood methods for the combined analysis of CMB temperature and polarisation power spectra*, *Mon. Not. Roy. Astron. Soc.* **372** (2006) 1104 [[astro-ph/0604547](#)].
- [67] J. Alsing and B. Wandelt, *Generalized massive optimal data compression*, *Mon. Not. Roy. Astron. Soc.* **476** (2018) L60 [[1712.00012](#)].
- [68] A. Heavens, R. Jimenez and O. Lahav, *Massive lossless data compression and multiple parameter estimation from galaxy spectra*, *Mon. Not. Roy. Astron. Soc.* **317** (2000) 965 [[astro-ph/9911102](#)].
- [69] M. Tegmark, A. Taylor and A. Heavens, *Karhunen-Loeve eigenvalue problems in cosmology: How should we tackle large data sets?*, *Astrophys. J.* **480** (1997) 22 [[astro-ph/9603021](#)].
- [70] A. Gelman and D.B. Rubin, *Inference from iterative simulation using multiple sequences*, *Statistical science* **7** (1992) 457.
- [71] W.A. Fendt and B.D. Wandelt, *Pico: Parameters for the Impatient Cosmologist*, *Astrophys. J.* **654** (2006) 2 [[astro-ph/0606709](#)].
- [72] EUCLID collaboration, *Euclid preparation: IX. EuclidEmulator2 – Power spectrum emulation with massive neutrinos and self-consistent dark energy perturbations*, [2010.11288](#).
- [73] A. Mootoovaloo, A.H. Jaffe, A.F. Heavens and F. Leclercq, *Kernel-based emulator for the 3D matter power spectrum from CLASS*, *Astron. Comput.* **38** (2022) 100508 [[2105.02256](#)].
- [74] A. Hall and A. Taylor, *A Bayesian method for combining theoretical and simulated covariance matrices for large-scale structure surveys*, *Mon. Not. Roy. Astron. Soc.* **483** (2019) 189 [[1807.06875](#)].
- [75] N. Chartier, B. Wandelt, Y. Akrami and F. Villaescusa-Navarro, *CARPool: fast, accurate computation of large-scale structure statistics by pairing costly and cheap cosmological simulations*, *Mon. Not. Roy. Astron. Soc.* **503** (2021) 1897 [[2009.08970](#)].
- [76] S. He, Y. Li, Y. Feng, S. Ho, S. Ravanbakhsh, W. Chen et al., *Learning to Predict the Cosmological Structure Formation*, *Proc. Nat. Acad. Sci.* **116** (2019) 13825 [[1811.06533](#)].
- [77] R. Alves de Oliveira, Y. Li, F. Villaescusa-Navarro, S. Ho and D.N. Spergel, *Fast and Accurate Non-Linear Predictions of Universes with Deep Learning*, in *34th Conference on Neural Information Processing Systems*, 11, 2020 [[2012.00240](#)].
- [78] N. Kaushal, F. Villaescusa-Navarro, E. Giusarma, Y. Li, C. Hawry and M. Reyes, *NECOLA: Towards a Universal Field-level Cosmological Emulator*, [2111.02441](#).
- [79] F. Beutler, C. Blake, M. Colless, D.H. Jones, L. Staveley-Smith, L. Campbell et al., *The 6df galaxy survey: baryon acoustic oscillations and the local hubble constant*, *Monthly Notices of the Royal Astronomical Society* **416** (2011) 3017–3032.
- [80] L. Anderson, E. Aubourg, S. Bailey, F. Beutler, V. Bhardwaj, M. Blanton et al., *The clustering of galaxies in the sdss-iii baryon oscillation spectroscopic survey: baryon acoustic*

*oscillations in the data releases 10 and 11 galaxy samples, *Monthly Notices of the Royal Astronomical Society* **441** (2014) 24–62.*

- [81] A.J. Ross, L. Samushia, C. Howlett, W.J. Percival, A. Burden and M. Manera, *The clustering of the sdss dr7 main galaxy sample – i. a 4 per cent distance measure at  $z = 0.15$ , *Monthly Notices of the Royal Astronomical Society* **449** (2015) 835–847.*
- [82] O.H.E. Philcox, M.M. Ivanov, M. Simonović and M. Zaldarriaga, *Combining Full-Shape and BAO Analyses of Galaxy Power Spectra: A 1.6% CMB-independent constraint on  $H_0$ , *JCAP* **05** (2020) 032 [2002.04035].*
- [83] A. Chudaykin, M.M. Ivanov, O.H.E. Philcox and M. Simonović, *Nonlinear perturbation theory extension of the Boltzmann code CLASS, *Phys. Rev. D* **102** (2020) 063533* [2004.10607].
- [84] F. Couchot, S. Henrot-Versillé, O. Perdereau, S. Plaszczynski, B. Rouillé d’Orfeuil, M. Spinelli et al., *Cosmology with the cosmic microwave background temperature-polarization correlation, *Astron. Astrophys.* **602** (2017) A41* [1609.09730].
- [85] PLANCK collaboration, *Planck 2018 results. V. CMB power spectra and likelihoods, *Astron. Astrophys.* **641** (2020) A5* [1907.12875].
- [86] J. Torrado and A. Lewis, *Cobaya: Code for Bayesian Analysis of hierarchical physical models, *JCAP* **05** (2021) 057* [2005.05290].
- [87] H. Prince and J. Dunkley, *Data compression in cosmology: A compressed likelihood for Planck data, *Phys. Rev. D* **100** (2019) 083502* [1909.05869].
- [88] PLANCK collaboration, *Planck 2018 results. VI. Cosmological parameters, *Astron. Astrophys.* **641** (2020) A6* [1807.06209].
- [89] H. Prince and J. Dunkley, *A Python compressed low- $\ell$  Planck likelihood for temperature and polarization, 2104.05715.*
- [90] I. Esteban, M.C. Gonzalez-Garcia, A. Hernandez-Cabezudo, M. Maltoni and T. Schwetz, *Global analysis of three-flavour neutrino oscillations: synergies and tensions in the determination of  $\theta_{23}$ ,  $\delta_{CP}$ , and the mass ordering, *JHEP* **01** (2019) 106* [1811.05487].
- [91] P.F. de Salas, D.V. Forero, S. Gariazzo, P. Martínez-Miravé, O. Mena, C.A. Ternes et al., *2020 global reassessment of the neutrino oscillation picture, *JHEP* **02** (2021) 071* [2006.11237].
- [92] F. Capozzi, E. Di Valentino, E. Lisi, A. Marrone, A. Melchiorri and A. Palazzo, *Global constraints on absolute neutrino masses and their ordering, *Phys. Rev. D* **95** (2017) 096014* [2003.08511].
- [93] J. Lesgourgues and S. Pastor, *Massive neutrinos and cosmology, *Physics Reports* **429** (2006) 307.*
- [94] Y.Y. Wong, *Neutrino mass in cosmology: status and prospects, *Annual Review of Nuclear and Particle Science* **61** (2011) 69.*
- [95] J. Lesgourgues and S. Pastor, *Neutrino mass from cosmology, *Advances in High Energy Physics* **2012** (2012) .*
- [96] J. Lesgourgues and S. Pastor, *Neutrino cosmology and planck, *New Journal of Physics* **16** (2014) 065002.*
- [97] M. Aker, A. Beglarian, J. Behrens, A. Berlev, U. Besserer, B. Bieringer et al., *Direct neutrino-mass measurement with sub-electronvolt sensitivity, *Nature Physics* **18** (2022) 160.*
- [98] KATRIN collaboration, *Improved Upper Limit on the Neutrino Mass from a Direct Kinematic Method by KATRIN, *Phys. Rev. Lett.* **123** (2019) 221802* [1909.06048].
- [99] A. Oddo, F. Rizzo, E. Sefusatti, C. Porciani and P. Monaco, *Cosmological parameters from the likelihood analysis of the galaxy power spectrum and bispectrum in real space, 2108.03204.*

- [100] F. Leclercq and A. Heavens, *On the accuracy and precision of correlation functions and field-level inference in cosmology*, *Mon. Not. Roy. Astron. Soc.* **506** (2021) L85 [[2103.04158](https://arxiv.org/abs/2103.04158)].
- [101] H. Bal, D. Epema, C. de Laat, R. van Nieuwpoort, J. Romein, F. Seinstra et al., *A medium-scale distributed system for computer science research: Infrastructure for the long term*, *Computer* **49** (2016) 54.
- [102] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau et al., *Array programming with NumPy*, *Nature* **585** (2020) 357.
- [103] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau et al., *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*, *Nature Methods* **17** (2020) 261.
- [104] J.D. Hunter, *Matplotlib: A 2d graphics environment*, *Computing in Science & Engineering* **9** (2007) 90.
- [105] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan et al., *Pytorch: An imperative style, high-performance deep learning library*, in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox and R. Garnett, eds., pp. 8024–8035, Curran Associates, Inc. (2019), <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- [106] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic et al., *Jupyter notebooks - a publishing format for reproducible computational workflows*, in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides and B. Schmidt, eds., (Netherlands), pp. 87–90, IOS Press, 2016, <https://eprints.soton.ac.uk/403913/>.