Averting A Crisis In Simulation-Based Inference

Joeri Hermans*
University of Liège
joeri.hermans@doct.uliege.be

Arnaud Delaunoy*
University of Liège
a.delaunoy@uliege.be

François Rozet
University of Liège
francois.rozet@uliege.be

Antoine Wehenkel
University of Liège
antoine.wehenkel@uliege.be

Gilles Louppe
University of Liège
g.louppe@uliege.be

Abstract

We present extensive empirical evidence showing that current Bayesian simulation-based inference algorithms are inadequate for the falsificationist methodology of scientific inquiry. Our results collected through months of experimental computations show that all benchmarked algorithms – (S)NPE, (S)NRE, SNL and variants of ABC – may produce overconfident posterior approximations, which makes them demonstrably unreliable and dangerous if one's scientific goal is to constrain parameters of interest. We believe that failing to address this issue will lead to a well-founded trust crisis in simulation-based inference. For this reason, we argue that research efforts should now consider theoretical and methodological developments of conservative approximate inference algorithms and present research directions towards this objective. In this regard, we show empirical evidence that ensembles are consistently more reliable.

1 Introduction

Many scientific disciplines rely on computer simulations to study complex phenomena under various conditions. Although modern simulators can generate realistic synthetic observables through detailed descriptions of their data generating processes, they are unfortunately not suitable for statistical inference. The computer code describing the data generating processes defines the likelihood function $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ only implicitly, and its direct evaluation requires the often *intractable* integration of all stochastic execution paths. In this problem setting, statistical inference based on the likelihood becomes impractical. However, approximate inference remains possible by relying on likelihood-free *approximations* thanks to the increasingly accessible and effective suite of methods and software from the field of simulation-based inference (Cranmer et al., 2020).

While simulation-based inference targets domain sciences, advances in the field are mainly driven from a machine learning perspective. The field, therefore, inherits the quality assessments (Lueckmann et al., 2021) customary to the machine learning literature, such as the minimization of classical divergence criteria. Despite recent developments of post hoc diagnostics to inspect the quality of likelihood-free approximations (Cranmer et al., 2015; Brehmer et al., 2018, 2019; Hermans et al., 2021; Lueckmann et al., 2021; Talts et al., 2018; Dalmasso et al., 2020), assessing whether approximate inference results are sufficiently reliable for scientific inquiry remains largely unanswered whenever fitting criteria are not globally optimized or whenever the data is limited. In fact, domain sciences, and more specifically the physical sciences, are not necessarily interested in the *exactness* of an approximation. Instead, in the tradition of Popperian falsification, they often seek to **constrain parameters** of interest as much as possible at a given confidence level. Scientific examples include frequentist confidence intervals on the mass of the Higgs boson (Aad et al., 2012), Bayesian credible regions on cosmological parameters (Gilman et al., 2018; Aghanim et al., 2020), or constraints on the intrinsic parameters of binary black hole coalescences (Abbott et al., 2016). From a Bayesian perspective, this implies that statistical approximations in simulation-based inference should ideally come with *conservative* guarantees to not produce credible regions smaller than they should be. Incorrectly constraining model param-

Under review. *Equal contribution

eters would otherwise impede scientific inquiry.

In this work, we measure the quality of the credible regions computed through various Bayesian techniques in simulation-based inference. We frame our main contribution as the collection of extensive empirical evidence that required months of computation. Our results demonstrate that all benchmarked techniques can produce non-conservative credible regions, highlighting the need for a new class of conservative approximate inference algorithms. The structure of the paper is outlined as follows. Section 2 describes the statistical formalism, necessary background and includes a thorough motivation for coverage. Section 3 highlights our main results. Section 4 presents several avenues of future research to **enable drawing reliable scientific conclusions** with simulation-based inference. All code related to this manuscript is available at `github.com/montefiore-ai/crisis-in-sbi`.

2 Background

2.1 Statistical formalism

We evaluate posterior estimators that produce approximations $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ with the following semantics.

Target parameters $\boldsymbol{\vartheta}$ denote the parameters of interest of a simulation model, and are sometimes referred to as *free* or *model* parameters. The precise definition of $\boldsymbol{\vartheta}$ depends on the problem setting. We make the reasonable assumption that the prior $p(\boldsymbol{\vartheta})$ is tractable.

An **observable $\boldsymbol{x}$** denotes a synthetic realization of the simulator. Observed data $\boldsymbol{x}_o$ is the observable we would like to do inference on, under the assumption that the simulation model is correctly specified.

The **likelihood** model $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ implicitly defined by the simulator's computer code. While we cannot evaluate the density $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$, we *can* simulate samples.

The **ground truth $\boldsymbol{\vartheta}^*$** specified to the simulation model whose forward evaluation produced the observable $\boldsymbol{x}_o$, i.e., $\boldsymbol{x}_o \sim p(\boldsymbol{x} \mid \boldsymbol{\vartheta} = \boldsymbol{\vartheta}^*)$.

A **credible region** is a space Θ within the target parameter domain that satisfies

$$\int_\Theta p(\boldsymbol{\vartheta} \mid \boldsymbol{x} = \boldsymbol{x}_o) \, \mathrm{d}\boldsymbol{\vartheta} = 1 - \alpha \tag{1}$$

for some observable $\boldsymbol{x}_o$ and confidence level $1 - \alpha$. Because many such regions exist, we compute the credible region with the *smallest* volume. In the literature this credible region is known as the *highest posterior density region* (Box and Tiao, 1973; Hyndman, 1996).

2.2 Statistical quality assessment

Common metrics for evaluating the quality of a posterior surrogate include the Classifier Two-sample Test (Lehmann and Romano, 2006; Lopez-Paz and Oquab, 2017) and Maximum Mean Discrepancy (Gretton et al., 2012; Bengio et al., 2014; Dziugaite et al., 2015). The main problem with these metrics is that they assess *exactness* of an approximation through a divergence with respect to a posterior that is intractable in practice. Even if such evaluations would hypothetically be possible, there are no criteria to what constitutes an *acceptable* estimator. Moreover, it is not possible to be certain whether the classifier or kernel used to measure the divergence are expressive enough to differentiate between the true posterior and its approximation.

To clarify these points, consider the demonstration in Figure 1. A binary classifier is trained to discriminate between samples from a posterior approximation and the true posterior. The discriminative performance of the classifier is expressed through Area Under the Receiver Operating Characteristics curve (AUROC) and serves as a measure for divergence between both densi-

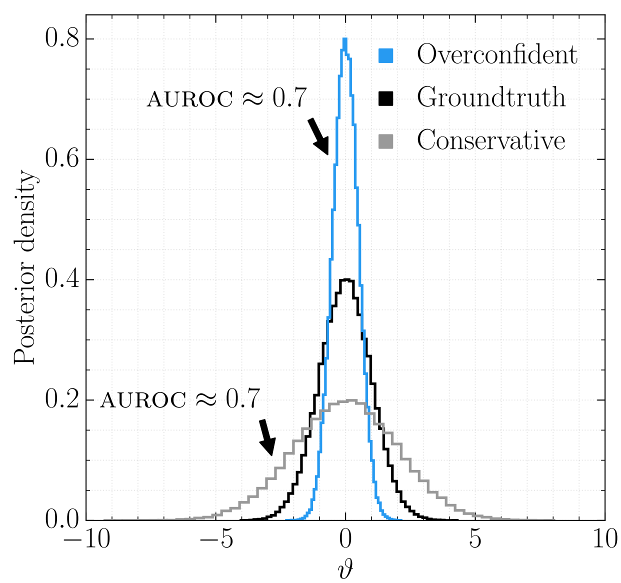

Figure 1: A classifier-based metric measures the divergence between posterior approximations and a ground truth by means of evaluating the classifier's discriminative performance through Area Under the Receiver Operating Characteristics curve (AUROC). In this case, the metric argues that both the conservative and overconfident approximations are equally accurate as it yields AUROC $= 0.7$ for both approximations. From an inference perspective however, the conservative approximation is more suitable because it produces credible regions larger than they should be.

ties. An AUROC $= 0.5$ suggests an approximation that is indistinguishable from the true posterior, while AUROC $= 1.0$ implies that both distributions do not overlap. Although both approximations in our demonstration are equally accurate according to the AUROC, the *over-confident* approximation illustrates the potential trust crisis in simulation-based inference: producing credible regions that are biased or smaller than they should be, potentially leading to erroneous scientific conclusions. For this reason, we take the position that posterior approximations should, irrespective of the available simulation budget, produce inflated credible regions. They do not have to closely match the true posterior to draw meaningful inferences.

Instead of measuring exactness of approximations with respect to an intractable posterior, this work directly probes the quality of credible regions through the notion of *expected* coverage, which determines whether posterior approximations are well-calibrated with respect to the specified prior. It is a quantity that can be estimated in practice and has a threshold to determine whether a posterior estimator is acceptable.

Definition 1. *The* ***expected coverage probability*** *of the* $1-\alpha$ *highest posterior density regions derived from the posterior estimator* $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ *is*

$$\mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\mathbb{1}\left[\boldsymbol{\vartheta} \in \Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}(1-\alpha)\right]\right], \tag{2}$$

where the function $\Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}(1-\alpha)$ *yields the* $1-\alpha$ *highest posterior density region of* $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$.

Note that Equation 2 can be expressed as either

$$\mathbb{E}_{p(\boldsymbol{\vartheta})}\mathbb{E}_{p(\boldsymbol{x} \mid \boldsymbol{\vartheta})}\left[\mathbb{1}\left[\boldsymbol{\vartheta} \in \Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}(1-\alpha)\right]\right], \tag{3}$$

which is the *expected* frequentist coverage probability, or alternatively as the expected Bayesian credibility

$$\mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\left[\mathbb{1}\left[\boldsymbol{\vartheta} \in \Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}(1-\alpha)\right]\right], \tag{4}$$

whose inner expectation reduces to $1-\alpha$ whenever the posterior estimator $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ is well-calibrated.

Definition 2. *The* ***empirical expected coverage probability*** *of the* $1-\alpha$ *highest posterior density regions derived from the posterior estimator* $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ *given a set of* n *i.i.d. samples* $(\boldsymbol{\vartheta}_i^*, \boldsymbol{x}_i) \sim p(\boldsymbol{\vartheta}, \boldsymbol{x})$ *is*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left[\boldsymbol{\vartheta}_i^* \in \Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}_i)}(1-\alpha)\right]. \tag{5}$$

Definition 3. *The* ***nominal expected coverage probability*** *is the expected coverage probability of the true posterior and is equal to the confidence level.*

Definition 4. *A posterior estimator is deemed acceptable if it* ***has coverage*** *at the confidence level of interest, i.e., whenever the empirical expected coverage probability is* ***larger or equal*** *to the nominal expected coverage probability.*

Definition 5. *A* ***conservative posterior estimator*** *has coverage for* ***all*** *confidence levels.*

While coverage is a necessary metric to assess conservativeness, it is limited in its ability to determine the information gain a posterior (approximation) has over its prior. To clarify this point, consider an estimator whose posteriors are identical to the prior. In this case, there is no gain in information and the empirical expected coverage probability is equal to the nominal expected coverage probability. For this reason, a complete analysis should be complemented with measures such as the mutual information or expected information gain $\mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\log p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) - \log p(\boldsymbol{\vartheta})\right]$. This work is however concerned with **conservative inference** and will therefore limit the analysis to the evaluation of expected coverage. Finally, it should be noted that expected coverage is a statement about the credible regions in expectation and therefore does not make any statement about the quality of an individual posterior.

3 Experimental observations

This section covers our main contribution: the collection of empirical evidence to determine whether approaches in simulation-based inference are conservative by nature. We are particularly interested in determining *whether* certain approaches should be favoured over others. We do so by measuring the coverage of posterior estimators produced by these approaches across a broad range of hyperparameters and benchmarks of varied complexity, including two *real* problems from the field of astronomy. As in real use-cases, the true posteriors associated with these benchmarks are unknown.

3.1 Methods

We make the distinction between two paradigms. Non-amortized approaches are designed to approximate a *single* posterior, while amortized methods aim to learn a general purpose estimator that attempts to approximate *all* posteriors supported by the prior. The architectures of all techniques, including hyperparameters, are listed in Appendix B.

3.1.1 Amortized

Neural Posterior Estimation (NPE) is concerned with directly learning an amortized posterior estimator $\hat{p}_\psi(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ with normalizing flows. Normalizing flows define a class of probability distributions $p_\psi(\cdot)$ built from neural network-based bijective transformations (Dinh et al., 2015, 2017) parameterized by ψ. They are usually optimized via $\arg\min_\psi \mathbb{E}_{p(\boldsymbol{x})}\left[\mathrm{KL}(p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \mid\mid \hat{p}_\psi(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right]$, which is equivalent to $\arg\max_\psi \mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\log \hat{p}_\psi(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right]$.

Once trained, the density of the modeled distribution can be evaluated *and* sampled from.

Neural Ratio Estimation (NRE) is an established approach in the simulation-based inference literature both from a frequentist (Cranmer et al., 2015) and Bayesian (Thomas et al., 2016; Hermans et al., 2020) perspective. In a Bayesian analysis, an amortized estimator $\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ of the *intractable* likelihood-to-evidence ratio $r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ can be learned by training a binary discriminator $\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})$ to distinguish between samples of the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$ with class label 1 and samples of the product of marginals $p(\boldsymbol{\vartheta})p(\boldsymbol{x})$ with class label 0 using a criterion such as the binary cross entropy. Similar to the density-ratio trick (Sugiyama et al., 2012; Goodfellow et al., 2014; Cranmer et al., 2015; Hermans et al., 2020), the Bayes optimal discriminator $d(\boldsymbol{\vartheta}, \boldsymbol{x})$ models

$$\frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta}, \boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})} = \sigma\left(\log \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\right), \qquad (6)$$

where $\sigma(\cdot)$ is the sigmoid function. Given a target parameter $\boldsymbol{\vartheta}$ and an observable $\boldsymbol{x}$ supported by $p(\boldsymbol{\vartheta})$ and $p(\boldsymbol{x})$ respectively, the learned discriminator $\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})$ approximates the log likelihood-to-evidence ratio $\log r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ through the logit function because $\text{logit}(\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})) \approx \log r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$. The approximate log posterior density function is $\log p(\boldsymbol{\vartheta}) + \log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$.

Ensembles of models constitute a standard method to improve predictive performance. In this work, we consider an ensemble model that *averages* the approximated posteriors of n independently trained posterior estimators. While this formulation is natural for NPE, averaging likelihood-to-evidence ratios is equivalent since $p(\boldsymbol{\vartheta})\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i(\boldsymbol{x} \mid \boldsymbol{\vartheta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_i(\boldsymbol{\vartheta} \mid \boldsymbol{x})$.

3.1.2 Non-amortized

Rejection **Approximate Bayesian Computation** (ABC) (Rubin, 1984; Pritchard et al., 1999) numerically estimates a *single* posterior by collecting samples $\boldsymbol{\vartheta} \sim p(\boldsymbol{\vartheta})$ whenever $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ is *similar* to $\boldsymbol{x}_o$. Similarity is expressed by means of a *distance function* ρ. For high-dimensional observables, the probability density of simulating an observable $\boldsymbol{x}$ such that $\boldsymbol{x} = \boldsymbol{x}_o$ is extremely small. For this reason, ABC uses a *summary statistic* s and an *acceptance threshold* ϵ. Using these components, ABC accepts samples into the approximate posterior whenever $\rho(s(\boldsymbol{x}), s(\boldsymbol{x}_o)) \leq \epsilon$. In our experiments we use the identity function as a sufficient summary statistic. Finally, we emphasize that ABC approximations are *only exact* whenever the summary statistic is sufficient *and* the acceptance threshold ϵ tends to 0 (Sisson et al., 2018).

Sequential methods aim to approximate a single posterior by *iteratively* improving a posterior approximation. These methods alternate between a simulation and exploitation phase. The latter being designed to take *current* knowledge into account such that subsequent simulations can be focused on parameters that are more likely to produce observables x similar to x_o.

Sequential Monte-Carlo ABC (SMC-ABC) (Toni and Stumpf, 2009; Sisson et al., 2007; Beaumont et al., 2009) iteratively updates a set of proposal states to match the posterior distribution. At each iteration, accepted proposals are ranked by distance. The rankings determine whether a proposal is propagated to the next iteration. New candidate proposals are generated by perturbing the selected ranked proposals.

Sequential Neural Posterior Estimation (SNPE) (Papamakarios and Murray, 2016; Lueckmann et al., 2017; Greenberg et al., 2019) directly models the posterior. Our evaluations will specifically use the SNPE-C (Greenberg et al., 2019) variant.

Sequential Neural Likelihood (SNL) (Papamakarios et al., 2019) models the likelihood $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$. A numerical approximation of the posterior is obtained by plugging the learned likelihood estimator into a Markov Chain Monte Carlo (MCMC) sampler as a surrogate likelihood.

Sequential Neural Ratio Estimator (SNRE) (Hermans et al., 2020; Durkan et al., 2020) iteratively improves the modelled likelihood-to-evidence ratio.

3.2 Benchmarks

Our evaluations consider 7 benchmarks, ranging from toy problems to real applications in astrophysics.

The *SLCP* simulator models a fictive problem with 5 parameters. The observable $\boldsymbol{x}$ is composed of 8 scalars which represent the 2D-coordinates of 4 points. The coordinate of each point is sampled from the same multivariate Gaussian whose mean and covariance matrix are parametrized by $\boldsymbol{\vartheta}$. We consider an alternative version of the original task (Papamakarios et al., 2019) by inferring the marginal posterior density of 2 of those parameters. In contrast to its original formulation, the likelihood is not tractable due to the marginalization.

The *Weinberg* problem (Cranmer et al., 2017) concerns a simulation of high energy particle collisions $e^+e^- \rightarrow \mu^+\mu^-$. The angular distributions of the particles can be used to measure the Weinberg angle $\boldsymbol{x}$ in the standard model of particle physics. From the scattering angle, we are interested in inferring Fermi's constant $\boldsymbol{\vartheta}$.

The *Spatial SIR* model involves a grid-world of susceptible, infected, and recovered individuals. Based on initial conditions and the infection and recovery rate $\boldsymbol{\vartheta}$, the model describes the spatial evolution of an infection. The observable $\boldsymbol{x}$ is a snapshot of the grid-world after some fixed amount of time.

M/G/1, originally introduced by Papamakarios et al. (2019), models a processing and arrival queue. The problem is described by 3 parameters $\boldsymbol{\vartheta}$ that influence the time it takes to serve a customer, and the time between their arrivals. The observable $\boldsymbol{x}$ is composed of 5 equally spaced quantiles of inter-departure times.

The *Lotka-Volterra* population model (Lotka, 1920; Volterra, 1926) describes a process of interactions between a predator and a prey species. The model is conditioned on 4 parameters $\boldsymbol{\vartheta}$ which influence the reproduction and mortality rate of the predator and prey species. We infer the marginal posterior of the predator parameters from time series representing the evolution of both populations over time.

Stellar *Streams* form due to the disruption of spherically packed clusters of stars by the Milky Way. Because of their distance from the galactic center and other visible matter, *distant* stellar streams are considered to be ideal probes to detect gravitational interactions with dark matter. The model (Banik et al., 2018) evolves the stellar density $\boldsymbol{x}$ of a stream over several billion years and perturbs the stream over its evolution through gravitational interactions with dark matter subhaloes parameterized by the dark matter mass $\boldsymbol{\vartheta}$.

Gravitational Waves (GW) are ripples in space-time emitted during events such as the collision of two black-holes. They can be detected through interferometry measurements $\boldsymbol{x}$ and convey information about celestial bodies, unlocking new ways to study the universe. We consider inferring the masses $\boldsymbol{\vartheta}$ of two black-holes colliding through the observation of the gravitational wave as measured by LIGO's dual detectors (LIGO Scientific Collaboration, 2018; Biwer et al., 2019).

3.3 Setup

Our evaluations consider simulation budgets ranging from 2^{10} up to 2^{17} samples and confidence levels from 0.05 up to 0.95. Within the *amortized* setting we train, for every simulation budget, 5 posterior estimators for 100 epochs. The empirical expected coverage probability is computed on at least 5,000 unseen samples from the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$ and for all confidence levels under consideration. In addition, we repeat the expected coverage evaluation for ensembles of 5 estimators as well. Special care for *non-amortized* approaches is necessary because they only approximate a *single* posterior and can therefore not reasonably evaluate expected coverage. Our experiments estimate expected coverage of these methods *by proxy* by repeating the inference procedure on 300 distinct observables for a given simulation budget. The empirical expected coverage probabilities are estimated using the 300 approximated posteriors. Our experiments with NPE, SNPE, SNL, SNRE, REJ-ABC and SMC-ABC rely on the implementation in the `sbi` package (Tejero-Cantero et al., 2020), while we use a custom implementation for NRE.

Computational cost We would like to emphasize the computational requirements necessary to generate our main contribution: the experimental observations, whose generation took months of computation. The average CPU time for evaluating an amortized procedure on all non-astrophysical benchmark problems is in order of 200 CPU days, while for a non-amortized approach it increases to 2800 CPU days. The bulk of the cost was associated with the repeated optimization procedure and the constant resampling of the simulator.

3.4 Results

Figures 2 and 3 highlight our main results. Through these plots we can directly assess the *conservativeness* at a given confidence level and simulation budget. The figures should be interpreted as follows: a *perfectly calibrated* posterior has an empirical expected coverage probability equal to the nominal expected coverage probability. Plotting this relation produces a diagonal line. Conservative estimators on the other hand produce curves *above* the diagonal and overconfident models underneath. The plots highlight an unsettling observation: **all** benchmarked approaches produce non-conservative posterior approximations. In general, this pathology is especially prominent in non-amortized approaches with a small simulation budget; a regime they have been specifically designed for. A large simulation budget does not guarantee conservativeness either.

In sequential approaches, this behaviour could be explained by the alternating exploitation and simulation phase. One potential failure mode is that a non-conservative posterior approximation at a previous iteration forces the next simulation phase to not produce observables that should be associated with a higher posterior density, causing the estimator to increase its non-conservativeness at each iteration.

Despite the fact that all ABC approaches use a sufficient summary statistic (the identity function), our results demonstrate that this alone is no guarantee for conservative posterior approximations. In fact, using a sufficient summary statistic with $\epsilon > 0$ does not always correspond to conservative approximations. In such cases, ABC accepts samples with larger distances, permitting the procedure to shift the mass of the approximated posterior elsewhere. In addition, a limited number of posterior samples can negatively affect the quality of the credible regions, e.g., when approximating the posterior density function with kernel density estimation. Both cases could cause the observed be-

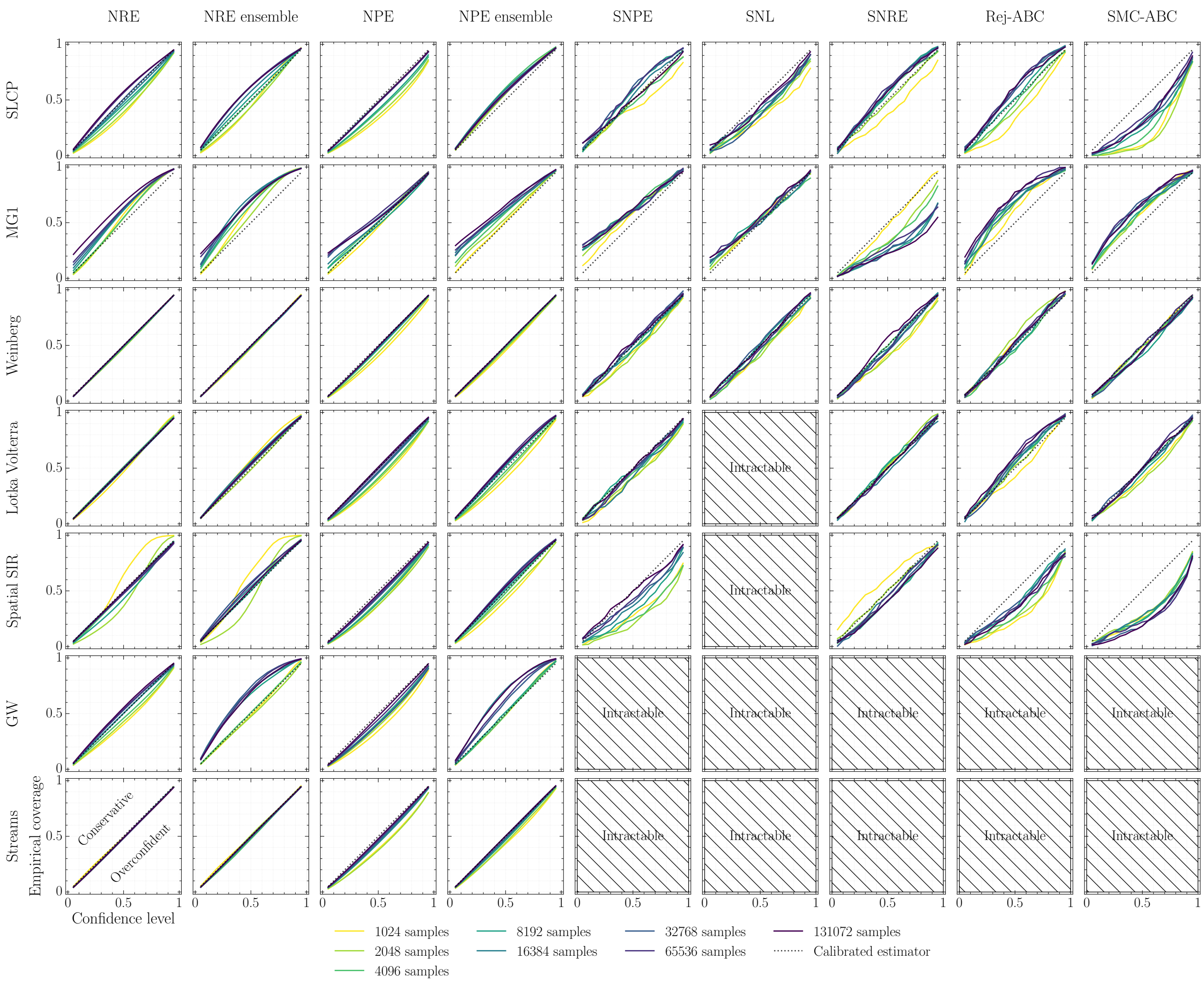

Figure 2: Evolution of expected coverage w.r.t simulation budget. A perfectly calibrated posterior has an empirical expected coverage probability equal to the nominal expected coverage probability and produces a diagonal line. Conservative estimators on the other hand produce curves *above* the diagonal and overconfident models underneath. All algorithms can lead to non-conservative estimators. This pathology tends to be accentuated for small simulation budgets and non-amortized methods. Finally, the intractable results indicate that the computational requirements did not allow for a coverage analysis. In the case of SNL, this was mostly due to the high dimensional observables. We did not train an embedding network as that is outside of the scope of this work. For the astronomy applications, the simulation model was simply too expensive to reasonably evaluate coverage for non-amortized methods.

haviour. Scientific applications should therefore be cautious. Even though a handcrafted, albeit sufficient, summary statistic provides some insight into the approximated posterior, it does not imply that ABC approximations are conservative whenever $\epsilon > 0$.

In Figure 3 we observe that the empirical expected coverage probability of ensemble models is consistently larger than the empirical expected coverage probability of the expected individual posterior estimator. Current applications of simulation-based inference can therefore rely on ensembling to build more conservative posterior estimators. However, the ensemble model can still be non-conservative. We hypothesize that the increase in coverage is linked to the added uncertainty captured by the ensemble model, leading to inflated credible regions. In fact, individual estimators only capture data uncertainty, while an ensemble is expected to capture *part* of the model uncertainty as well. Surprisingly, we find that ensembles built using bagging do not always produce higher coverage than individual models while they should also capture part of the epistemic uncertainty. Although a deeper understanding of this effect is required, this behaviour could be explained by

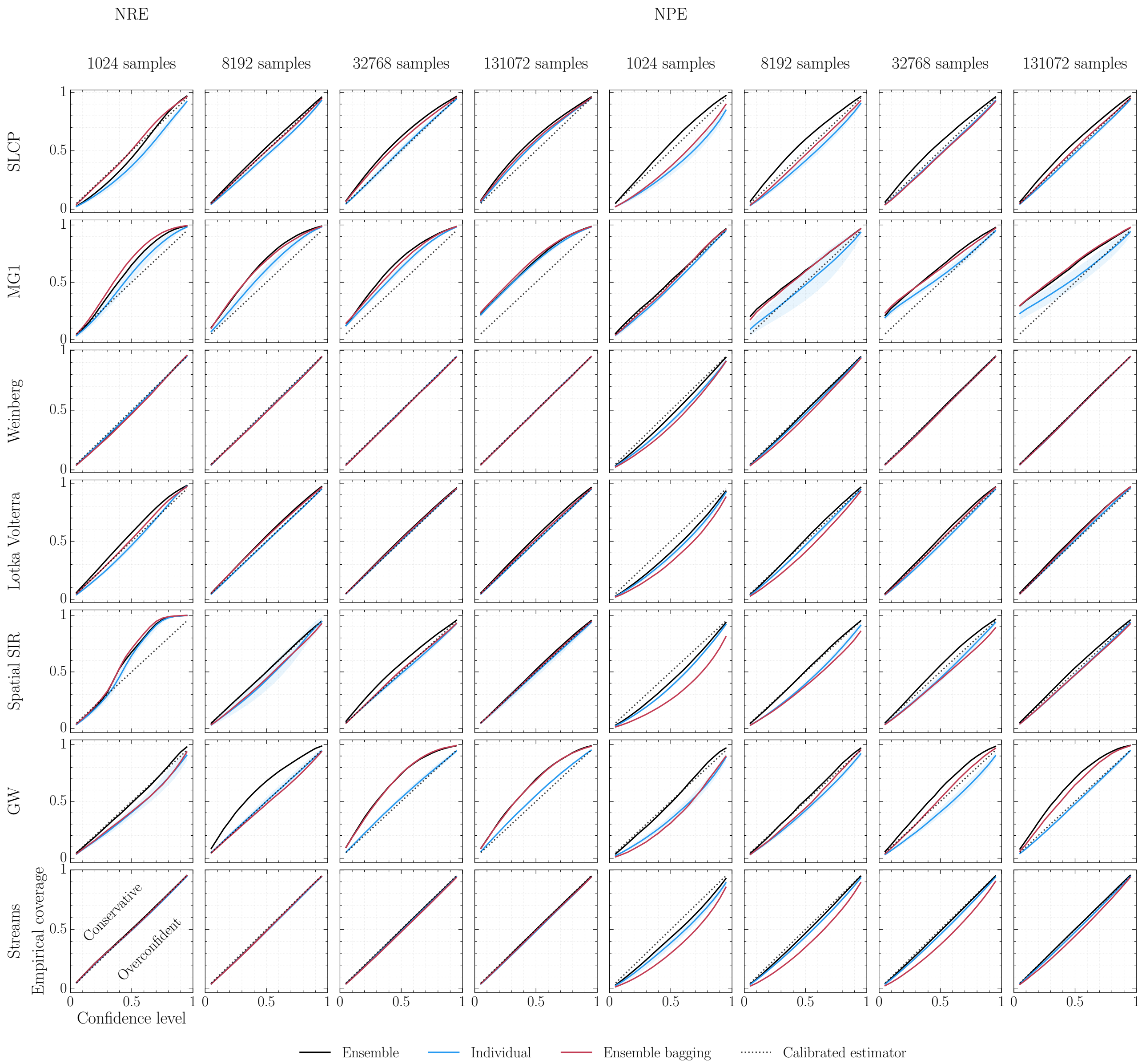

Figure 3: Analysis of coverage between ensemble and individual models w.r.t various simulation budgets. The blue line represents the mean empirical expected coverage of individual models over 5 runs, the shaded area represents its standard deviation. The black line represents the empirical expected coverage of a single ensemble composed of 5 models. We observe that ensembles consistently have a higher empirical expected coverage probability compared to the average individual model. A similar effect is not always observed with bagging, indicated by the red line. Ensembles are only considered on amortized approached.

the fact that bagging reduces the effective dataset size used to train each member of the ensemble. Appendix C shows a positive effect with respect to ensemble size.

Not evident from Figures 2 and 3 are the computational consequences of a coverage analysis on non-amortized methods. Although the figures mention a certain simulation budget, the *total* number of simulations for non-amortized methods should be multiplied by the number of approximated posteriors (300) to estimate coverage, highlighting the simulation cost associated with diagnosing non-amortized approaches. This issue is not limited to coverage. Simulation-Based Calibration (SBC) (Talts et al., 2018) relies on samples of arbitrary posterior approximations. Diagnosing non-amortized estimators with SBC therefore requires a similar approach as we have taken in our coverage analyses. In fact, Lueckmann et al. (2021) also mention that SBC is computationally prohibitive for non-amortized approaches and therefore restrain from evaluating it.

Our results illustrate a clear distinction between the amortized and non-amortized paradigms. Amortized methods do not require retraining or new simulations to determine the empirical expected coverage probability of a posterior estimator, while non-amortized methods do. A coverage analysis of non-amortized approaches therefore only measures the quality of the training procedure. In contrast to amortized approaches, where the posterior estimator is diagnosed. This has severe implications on the applicability of non-amortized methods, because their reliability cannot be practically determined. In addition, non-amortized approaches have to repeat the approximation procedure whenever architectural or hyperparameter changes are made, while amortized methods reuse previously simulated datasets. In particular, sequential methods cannot do this as new simulations depend on the posterior approximation at a previous state. This is often overlooked in studies on simulation efficiency and raises questions about whether sequential approaches should still be considered simulation efficient over their amortized counterparts, considering that amortized approaches produce trustworthier posterior approximations in expectation.

All of the above leads us to conclude that *currently*, amortization should be favoured over non-amortized approaches because their reliability cannot practically be determined. Our results further suggest that even for small simulation budgets amortized methods, on average, produce more conservative estimators; a striking result, given that non-amortized, and sequential methods in particular, dedicate the available simulation budget to accurately approximate a single posterior.

Observation 1 All benchmarked algorithms produce non-conservative posterior approximations. This pathology tends to be accentuated with small simulation-budgets in both paradigms.

Observation 2 Amortized approaches have the tendency to be more conservative in comparison to non-amortized approaches.

Observation 3 The empirical expected coverage probability of an ensemble model is larger than the average individual model. The ensemble size positively affects the empirical expected coverage probability as well.

Observation 4 Amortized methods are simulation-efficient, especially when taking hyper-parameter tuning and the evaluation of the expected coverage diagnostic into account.

4 Discussion

As demonstrated empirically, simulation-based inference can be unreliable, especially when its approximations cannot be diagnosed. The problem of determining whether a posterior approximation is correct is in fact not restricted to simulation-based inference specifically, the concern occurs in all of approximate Bayesian inference. The MCMC literature deals with this exact same problem in the form of determining whether a set of Markov chain samples have converged to the target distribution (Lin, 2014; Hogg and Foreman-Mackey, 2018). In this regard, empirical diagnostic tools have been proposed over the years (Geweke et al., 1991; Gelman and Rubin, 1992; Raftery and Lewis, 1991; Dixit and Roy, 2017; Talts et al., 2018) and have helped practitioners using MCMC properly. Nonetheless, there is currently no clear solution to determine convergence with absolute certainty (Dixit, 2018; Roy, 2020), even if the likelihood function is here tractable.

We are of the opinion that theoretical and methodological advances within the field of simulation-based inference will strengthen its reliability and promote its applicability in sciences. First, although all benchmarked algorithms recover the true posterior under specific optimal conditions, it is generally not possible to know whether those conditions are satisfied in practice. Therefore, the study of new objective functions that would force posterior estimators to always be conservative, regardless of their optimal conditions, is worth investigating. From a Bayesian perspective, Rozet and Louppe (2021) propose using the focal and the peripheral losses to weigh down easily classified samples as a means to tune the conservativeness of a posterior estimator. Dalmasso et al. (2021) consider the frequentist setting and introduce a theoretically-grounded algorithm for the construction of confidence intervals that are guaranteed to have perfect coverage, regardless of the quality of the used statistic. Second, in light of our results that ensembles produce more conservative posteriors, model averaging constitutes another promising direction of study, as a simple and efficient method to produce reliable posterior estimators. However, a deeper understanding of the behaviour we observe is certainly first required to further develop these methods. Third, post-training calibration can be used to improve the reliability of posterior estimators and should certainly be considered as a way towards more conservative inference. To some extent, this has already been considered for amortized methods (Cranmer et al., 2015; Brehmer et al., 2018; Hermans et al., 2021) and would be worth exploring further, especially for non-amortized approaches.

In summary, we show that current algorithms for simulation-based inference may all produce overconfident posterior approximations, making them demonstrably unreliable if one's scientific goal is to constrain parameters of interest or reject theoretical models. Nevertheless, we remain confident and optimistic and advocate that this result is only a stepping stone towards more reliable simulation-based inference, its wider adoption, and eventually better science.

Acknowledgements

The authors would like to thank Christoph Weniger, Kyle Cranmer and Maxime Vandegar for their insightful comments and feedback. Antoine Wehenkel, Arnaud Delaunoy, and Joeri Hermans would like to thank the National Fund for Scientific Research (F.R.S.-FNRS) for their scholarships. Gilles Louppe is recipient of the ULiège - NRB Chair on Big Data and is thankful for the support of the NRB. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the National Fund for Scientific Research (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

References

Georges Aad, Tatevik Abajyan, B Abbott, J Abdallah, S Abdel Khalek, Ahmed Ali Abdelalim, R Aben, B Abi, M Abolins, OS AbouZeid, et al. Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Physics Letters B*, 716(1):1–29, 2012.

Benjamin P Abbott, R Abbott, TD Abbott, MR Abernathy, F Acernese, K Ackley, C Adams, T Adams, P Addesso, RX Adhikari, et al. Gw151226: observation of gravitational waves from a 22-solar-mass binary black hole coalescence. *Physical review letters*, 116(24):241103, 2016.

N. Aghanim et al. Planck 2018 results. VI. Cosmological parameters. *Astron. Astrophys.*, 641:A6, 2020. doi: 10.1051/0004-6361/201833910. [Erratum: Astron.Astrophys. 652, C4 (2021)].

Nilanjan Banik, Gianfranco Bertone, Jo Bovy, and Nassim Bozorgnia. Probing the nature of dark matter particles with stellar streams. *Journal of Cosmology and Astroparticle Physics*, 2018(07):061, 2018.

Mark A Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P Robert. Adaptive approximate bayesian computation. *Biometrika*, 96(4):983–990, 2009.

Yoshua Bengio, Li Yao, and Kyunghyun Cho. Bounding the test log-likelihood of generative models. In *International Conference on Learning Representations*, 2014.

C. M. Biwer, Collin D. Capano, Soumi De, Miriam Cabero, Duncan A. Brown, Alexander H. Nitz, and V. Raymond. PyCBC Inference: A Python-based parameter estimation toolkit for compact binary coalescence signals. *Publ. Astron. Soc. Pac.*, 131(996):024503, 2019. doi: 10.1088/1538-3873/aaef0b.

Jo Bovy. galpy: A python library for galactic dynamics. *The Astrophysical Journal Supplement Series*, 216 (2):29, 2015.

George EP Box and George C Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 1973.

Johann Brehmer, Kyle Cranmer, Gilles Louppe, and Juan Pavez. A Guide to Constraining Effective Field Theories with Machine Learning. *Phys. Rev. D*, 98 (5):052004, 2018. doi: 10.1103/PhysRevD.98.052004.

Johann Brehmer, Siddharth Mishra-Sharma, Joeri Hermans, Gilles Louppe, and Kyle Cranmer. Mining for dark matter substructure: Inferring subhalo population properties from strong lenses with machine learning. *The Astrophysical Journal*, 886(1):49, 2019.

Kyle Cranmer, Juan Pavez, and Gilles Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*, 2015.

Kyle Cranmer, Lukas Heinrich, Tim Head, and Gilles Louppe. "Active Sciencing" with Reusable Workflows. https://github.com/cranmer/active_sciencing, 2017.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 2020.

Niccolò Dalmasso, Rafael Izbicki, and Ann Lee. Confidence sets and hypothesis testing in a likelihood-free inference setting. In *International Conference on Machine Learning*, pages 2323–2334. PMLR, 2020.

Niccolò Dalmasso, David Zhao, Rafael Izbicki, and Ann B Lee. Likelihood-free frequentist inference: Bridging classical statistics and machine learning in simulation and uncertainty quantification. *arXiv preprint arXiv:2107.03920*, 2021.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Anand Dixit and Vivekananda Roy. Mcmc diagnostics for higher dimensions using kullback leibler divergence. *Journal of Statistical Computation and Simulation*, 87(13):2622–2638, 2017.

Anand Ulhas Dixit. *Developments in MCMC diagnostics and sparse Bayesian learning models*. PhD thesis, Iowa State University, 2018.

Conor Durkan, Iain Murray, and George Papamakarios. On contrastive learning for likelihood-free inference. In *International Conference on Machine Learning*, pages 2771–2781. PMLR, 2020.

Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 258–267, 2015.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

John F Geweke et al. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Technical report, Federal Reserve Bank of Minneapolis, 1991.

Daniel Gilman, Simon Birrer, Tommaso Treu, Charles R Keeton, and Anna Nierenberg. Probing the nature of dark matter by forward modelling flux ratios in strong gravitational lenses. *Monthly Notices of the Royal Astronomical Society*, 481(1):819–834, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, pages 2404–2414. PMLR, 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free MCMC with amortized approximate ratio estimators. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4239–4248. PMLR, 13–18 Jul 2020.

Joeri Hermans, Nilanjan Banik, Christoph Weniger, Gianfranco Bertone, and Gilles Louppe. Towards constraining warm dark matter with stellar streams through neural simulation-based inference. *Monthly Notices of the Royal Astronomical Society*, 507(2): 1999–2011, 2021.

David W Hogg and Daniel Foreman-Mackey. Data analysis recipes: Using markov chain monte carlo. *The Astrophysical Journal Supplement Series*, 236 (1):11, 2018.

Rob J Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2): 120–126, 1996.

Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Proceedings of the 31st international conference on neural information processing systems*, pages 972–981, 2017.

Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

LIGO Scientific Collaboration. LIGO Algorithm Library - LALSuite. free software (GPL), 2018.

Jing Lin. An integrated procedure for bayesian reliability inference using mcmc. *Journal of Quality and Reliability Engineering*, 2014, 2014.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Alfred J Lotka. Analytical note on certain rhythmic relations in organic systems. *Proceedings of the National Academy of Sciences*, 6(7):410–415, 1920.

Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in Neural Information Processing Systems*, 30, 2017.

Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th*

International Conference on Artificial Intelligence and Statistics, volume 130 of *Proceedings of Machine Learning Research*, pages 343–351. PMLR, 13–15 Apr 2021.

George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. In *Advances in neural information processing systems*, pages 1028–1036, 2016.

George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.

Jonathan K Pritchard, Mark T Seielstad, Anna Perez-Lezaun, and Marcus W Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16 (12):1791–1798, 1999.

Adrian E Raftery and Steven Lewis. How many iterations in the gibbs sampler? Technical report, 1991.

Vivekananda Roy. Convergence diagnostics for markov chain monte carlo. *Annual Review of Statistics and Its Application*, 7:387–412, 2020.

François Rozet and Gilles Louppe. Arbitrary marginal neural ratio estimation for likelihood-free inference. Master's thesis, University of Liège, Belgium, 2021.

Donald B. Rubin. Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, 12(4):1151 – 1172, 1984.

Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

Scott A Sisson, Yanan Fan, and Mark A Beaumont. Overview of abc. In *Handbook of approximate Bayesian computation*, pages 3–54. Chapman and Hall/CRC, 2018.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

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

Owen Thomas, Ritabrata Dutta, Jukka Corander, Samuel Kaski, Michael U Gutmann, et al. Likelihood-free inference by ratio estimation. *Bayesian Analysis*, 2016.

Tina Toni and Michael P. H. Stumpf. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110, 10 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp619.

Vito Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118 (2972):558–560, 1926.

Appendix

A Expected simulation times

SLCP	M/G/1	Weinberg	Lotka-V.	Spatial SIR	GW	Streams
0.22 ± 0.002	0.20 ± 0.002	0.20 ± 0.002	19.08 ± 0.964	9.18 ± 0.280	545.13 ± 23.631	$39,369 \pm 584.43$

Table 1: Expected simulation time to produce 1000 simulations for all benchmark problems on a single CPU core. The expected time and standard deviation are reported in seconds.

B Architectures & hyperparameters

In this section we describe the neural architectures and hyperparameters associated with our experiments. Our descriptions are complemented with the actual number of coverage evaluations. As evident from the tables describing both amortized and non-amortized approaches, the number of coverage evaluations for amortized approaches is substantially larger. It should be noted that, a coverage analysis consisting of 300 posteriors of the non-amortized approaches took *months* on these relatively simple problems. While for the amortized methods, a coverage analysis of 100,000 samples was a matter of hours to a few days depending on the dimensionality of $\boldsymbol{\vartheta}$.

B.1 Amortized

B.1.1 Neural Posterior Estimation

The MLP embeddings are 3 layer MLP's with 64 hidden units and a final latent space of 10, which is fed to the normalizing flow. The CNN architecture in the Gravitational Waves benchmark consists of a 13-layer deep convolutional head of 1D convolutions with a dilation factor of 2^d. Where d corresponds to the depth of the convolutional head. The SELU (Klambauer et al., 2017) function is used as an activation function.

	SLCP	M/G/1	Weinberg	Lotka-V.	Spatial SIR	GW	Streams
Embedding	MLP	MLP	MLP	MLP	MLP	CNN	MLP
Batch-size	128	128	128	128	128	64	128
Coverage samples individual	100,000	5,000	100,000	100,000	100,000	10,000	100,000
Coverage samples ensemble	20,000	5,000	20,000	20,000	20,000	5,000	20,000
Epochs	100	100	100	100	100	100	100
Model	NSF	NSF	NSF	NSF	NSF	NSF	NSF
Transforms	3	3	1	3	3	3	3
Learning-rate	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Table 2: Architectures and hyperparameters associated with Neural Posterior Estimation.

B.1.2 Neural Ratio Estimation

Our experiments use the ADAMW (Kingma and Ba, 2015; Loshchilov and Hutter, 2019) optimizer. Accross all benchmarks, the MLP architectures constitute of 3 hidden layers with 128 units and SELU (Klambauer et al., 2017) activations. The Gravitational Waves benchmark uses the same convolutional architecture as in NPE. The resulting embedding is flattened and fed to a MLP in which the dependence on the target parameter $\boldsymbol{\vartheta}$ is added. As before, the MLP consists of 3 hidden layers with 128 units.

B.2 Non-amortized

All our implementations of non-amortized approaches rely on the reference implementation in `sbi` (Tejero-Cantero et al., 2020). We use the recommended defaults unless stated otherwise. Whenever available, the same MLP

	SLCP	M/G/1	Weinberg	Lotka-V.	Spatial SIR	GW	Streams
Architecture	MLP	MLP	MLP	MLP	MLP	CNN	MLP
Batch-size	128	128	128	128	128	64	128
Coverage samples individual	100,000	100,000	100,000	100,000	100,000	10,000	100,000
Coverage samples ensemble	20,000	20,000	20,000	20,000	20,000	10,000	20,000
Epochs	100	100	100	100	100	100	100
Learning-rate	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Table 3: Architectures and hyperparameters associated with Neural Ratio Estimation.

embedding network is used. It consists of 3 hidden layers with 64 units and SELU (Klambauer et al., 2017) activations. The latent space has a dimensionality of 10 features. For all sequential methods, we use 10 rounds to iteratively improve the posterior approximation.

B.2.1 SNPE

Our evaluations with SNPE specifically use the SNPE-C (Greenberg et al., 2019) variant, as suggested by `sbi` (Tejero-Cantero et al., 2020). To minimize inconsistencies between experiments, we use the defaults suggested by the `sbi` authors unless states otherwise. Specific changes are highlighted in Table 4.

	SLCP	M/G/1	Weinberg	Lotka-V.	Spatial SIR	GW	Streams
Batch-size	128	128	128	128	32	Intractable	Intractable
Coverage samples	300	300	300	300	300	Intractable	Intractable
Embedding	MLP	MLP	MLP	MLP	MLP	Intractable	Intractable
Epochs	100	100	100	100	100	Intractable	Intractable
Features	64	64	64	64	64	Intractable	Intractable
Model	NSF	NSF	NSF	NSF	NSF	Intractable	Intractable
Transforms	3	3	1	3	3	Intractable	Intractable
Rounds	10	10	10	10	10	Intractable	Intractable
Learning-rate	0.001	0.001	0.001	0.001	0.001	Intractable	Intractable

Table 4: Architectures and hyperparameters associated with Sequential Neural Posterior Estimation.

B.2.2 SNL

In contrast to other sequential methods, our evaluations with SNL (Papamakarios et al., 2019) add two additional intractable benchmarks. At the root of this issue lies the dimensionality of the observable. In both cases, the dimensionality of observables caused memory issues in SNL. In addition, training a seperate embedding model (that requires additional simulations) is outside of the scope of this work. For this reason, we consider the Lotka-Volterra en Spatial SIR benchmark to be intractable.

	SLCP	M/G/1	Weinberg	Lotka-V.	Spatial SIR	GW	Streams
Batch-size	128	128	128	Intractable	Intractable	Intractable	Intractable
Coverage samples	300	300	300	Intractable	Intractable	Intractable	Intractable
Epochs	100	100	100	Intractable	Intractable	Intractable	Intractable
Features	64	64	64	Intractable	Intractable	Intractable	Intractable
Model	NSF	NSF	NSF	Intractable	Intractable	Intractable	Intractable
Transforms	3	3	1	Intractable	Intractable	Intractable	Intractable
Rounds	10	10	10	Intractable	Intractable	Intractable	Intractable
Learning-rate	0.001	0.001	0.001	Intractable	Intractable	Intractable	Intractable

Table 5: Architectures and hyperparameters associated with Sequential Neural Likelihood.

B.2.3 SNRE

	SLCP	M/G/1	Weinberg	Lotka-V.	Spatial SIR	GW	Streams
Architecture	MLP	MLP	MLP	MLP	MLP	Intractable	Intractable
Batch-size	128	128	128	128	128	Intractable	Intractable
Coverage samples	300	300	300	300	300	Intractable	Intractable
Epochs	100	100	100	100	100	Intractable	Intractable
Features	64	64	64	64	64	Intractable	Intractable
Rounds	10	10	10	10	10	Intractable	Intractable
Learning-rate	0.001	0.001	0.001	0.001	0.001	Intractable	Intractable

Table 6: Architectures and hyperparameters associated with Sequential Neural Ratio Estimation.

B.2.4 Approximate Bayesian Computation

Our ABC implementation relies on the `MCABC` and `SMCABC` classes in the `sbi` (Tejero-Cantero et al., 2020) package. The specific settings from Rejection ABC and SMC-ABC are described in Tables 7 and 8 respectively. The quantile specifically refers to the proportion of closest samples that were kept in the final posterior. Because our specific implementation of coverage requires the ability to describe the posterior density function, we relied on Kernel Density Estimation to estimate the posterior density from the accepted samples.

	SLCP	M/G/1	Weinberg	Lotka-V.	Spatial SIR	GW	Streams
Coverage samples	300	300	300	300	300	Intractable	Intractable
Quantile	0.01	0.01	0.01	0.01	0.01	Intractable	Intractable

Table 7: Hyperparameters associated with Rejection Approximate Bayesian Computation.

	SLCP	M/G/1	Weinberg	Lotka-V.	Spatial SIR	GW	Streams
Coverage samples	300	300	300	300	300	Intractable	Intractable
ϵ *decay*	0.5	0.5	0.5	0.5	0.5	Intractable	Intractable
Quantile	0.01	0.01	0.01	0.01	0.01	Intractable	Intractable

Table 8: Hyperparameters associated with Sequential Monte Carlo Approximate Bayesian Computation.

C Additional results

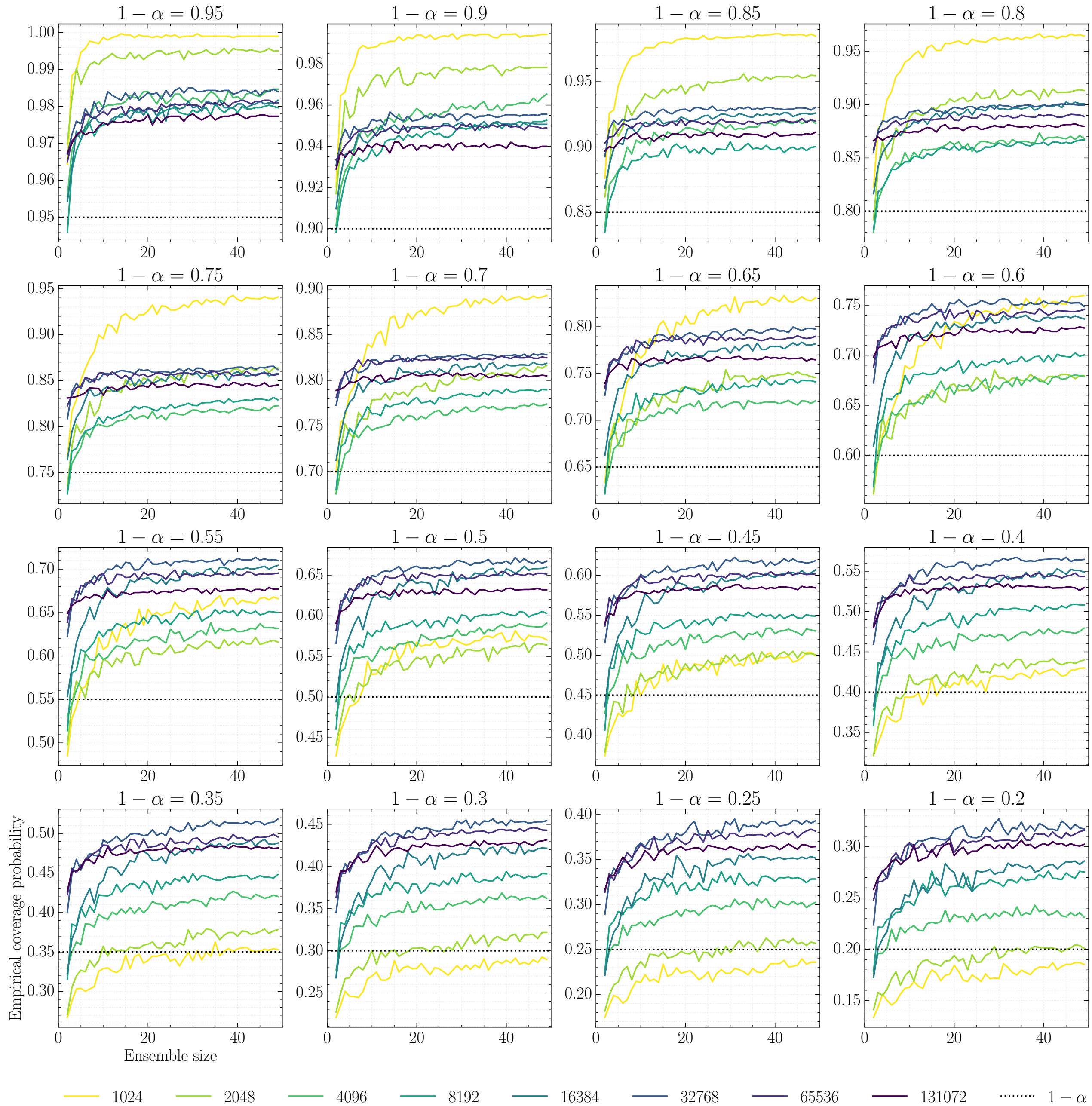

Figure 4: Evolution of the empirical expected coverage probability with respect to ensemble size for various confidence levels. The results are obtained by training 100 ratio estimators (NRE) on the SLCP benchmark. A positive effect is observed in terms of empirical expected coverage probability and ensemble size, i.e., a larger ensemble size correlates with a larger empirical expected coverage probability. This is unsurprising, because a larger ensemble is expected to capture more of the uncertainty that stems from the training procedure.