A Trust Crisis In Simulation-Based Inference? Your Posterior Approximations Can Be Unfaithful

Joeri Hermans[*]
Unaffiliated
joeri@peinser.com

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

Volodimir Begy
University of Vienna
volodimir.begy@univie.ac.at

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

Abstract

We present extensive empirical evidence showing that current Bayesian simulation-based inference algorithms can produce computationally unfaithful posterior approximations. Our results show that all benchmarked algorithms – (S)NPE, (S)NRE, SNL and variants of ABC – can yield overconfident posterior approximations, which makes them unreliable for scientific use cases and falsificationist inquiry. Failing to address this issue may reduce the range of applicability of simulation-based inference. For this reason, we argue that research efforts should be made towards theoretical and methodological developments of conservative approximate inference algorithms and present research directions towards this objective. In this regard, we show empirical evidence that ensembling posterior surrogates provides more reliable approximations and mitigates the issue.

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 [1].

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 [2] customary to the machine learning literature, primarily targeting the exactness of the approximation. Domain sciences, and more specifically the physical sciences, are not necessarily interested in the exactness of an approximation. 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

[*]Equal contribution

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

include frequentist confidence intervals on the mass of the Higgs boson [3], Bayesian credible regions on cosmological parameters [4, 5], or constraints on the intrinsic parameters of binary black hole coalescences [6]. Wrongly excluding plausible values could drive the scientific inquiry in the wrong direction, whereas failing to exclude implausible values because of too conservative estimations is much less detrimental. 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, even when the approximations are not faithful. Despite recent developments of post hoc diagnostics to inspect the quality of likelihood-free approximations [2, 7–12], 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 this work, we measure and discuss the quality of the credible regions produced by various algorithms for Bayesian simulation-based inference. All code related to this manuscript is available at `https://github.com/montefiore-ai/trust-crisis-in-simulation-based-inference`.

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. We make the reasonable assumption that the prior $p(\boldsymbol{\vartheta})$ is tractable. An **observable** $\boldsymbol{x}$ denotes a synthetic realization of the simulator, or the observed data $\boldsymbol{x}_o$ we would like to do inference on. We assume that the simulation model is correctly specified and hence is an accurate representation of the real data generation process. The **likelihood** model $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ is implicitly defined by the simulator's computer code. While we cannot evaluate the density $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$, we can draw samples through simulation. 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 parameters domain that satisfies $\int_\Theta p(\boldsymbol{\vartheta} \mid \boldsymbol{x} = \boldsymbol{x}_o)\, \mathrm{d}\boldsymbol{\vartheta} = 1 - \alpha$ 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.

2.2 Statistical quality assessment

Common metrics for evaluating the quality of a posterior surrogate assess exactness of an approximation through a divergence with respect to the posterior. All approximations will diverge from the posterior and there are no criteria to what constitutes an acceptable estimator. For this reason, we argue that metrics evaluating the reliability for scientific inquiry should be used alongside the divergence evaluation when evaluating estimators. This work directly assesses the quality of credible regions through the notion of expected coverage, which probes the consistency of the posterior approximations and can be used to diagnose conservative and overconfident approximations.

Definition 2.1. 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

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

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 1 can be expressed either as

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

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

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

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

Definition 2.2. A **conservative posterior estimator** is an estimator that **has coverage** at the credibility level of interest, i.e. the expected coverage probability is larger or equal to the credibility level.

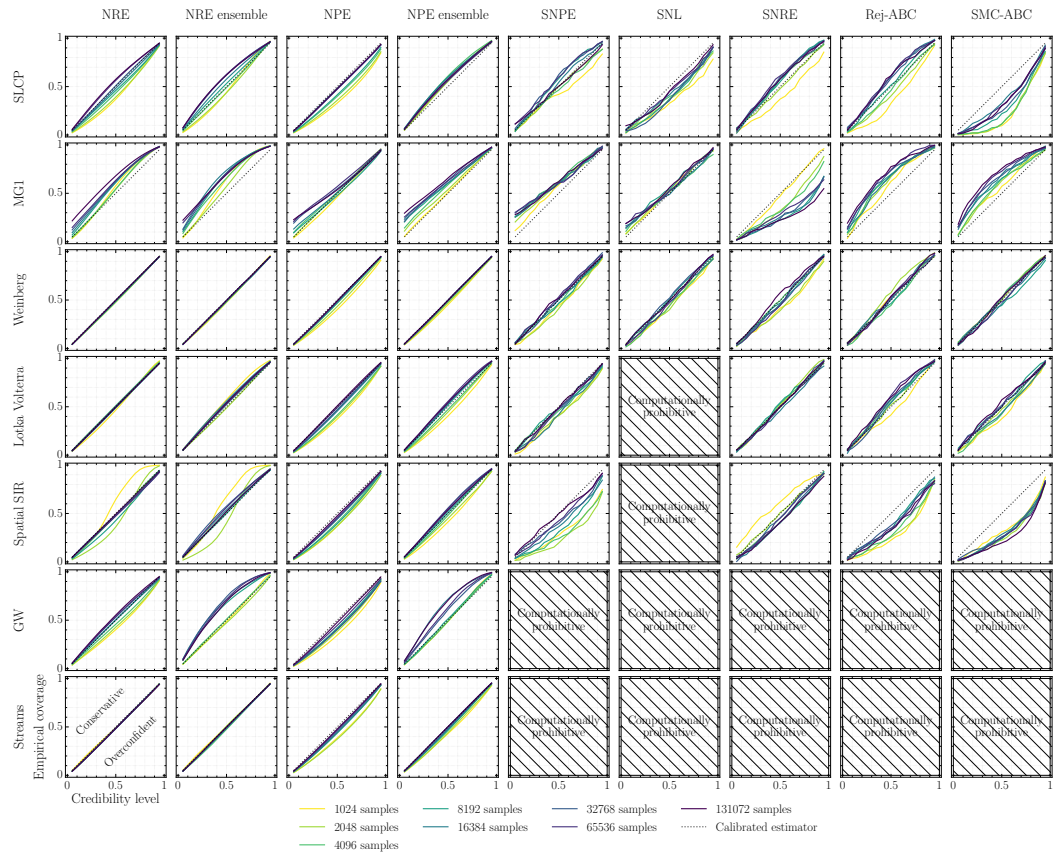

Figure 1: Evolution of the expected coverage w.r.t the simulation budget. A perfectly calibrated posterior has an expected coverage probability equal to the credibility level and produces a diagonal line. Conservative estimators on the other hand produce curves above the diagonal and overconfident models 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 computationally prohibitive results indicate that the computational requirements did not allow for a coverage analysis. In the case of SNL, this was mostly due to high dimensional observables. For the astronomy benchmarks, the simulation model was simply too expensive to reasonably evaluate coverage for non-amortized methods.

3 Empirical observations

This section covers our main contribution: the collection of empirical evidence to determine whether some simulation-based inference algorithms are conservative by nature. We are particularly interested in determining whether certain approaches should be favoured over others. We do so by estimating the expected coverage of posterior estimators produced by these approaches across a broad range of benchmarks of varying complexity, including two real problems. A description of the benchmarks can be found in Appendix A. As in real use cases, the true posteriors are effectively intractable and therefore unknown. 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. A description of the inference algorithms used, including architectures and hyperparameters, are listed in Appendix B and C. In addition, a description of the experimental protocol can be found in appendix D.

Results Figures 1 and 2 highlight our main results. Through these plots, we can directly assess whether a posterior estimator is conservative at a given confidence level and simulation budget. The figures should be interpreted as follows: a perfectly calibrated posterior has an expected coverage probability equal to the credibility level. 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 on at least one problem setting**. 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 that a posterior estimator is conservative either.

In Appendix E, we observe that the expected coverage probability of ensemble models is consistently larger than the expected coverage probability of an individual posterior estimator. **This highlights the fact that ensembling constitutes an immediately applicable and easy way to mitigate the overconfidence issue and build more reliable 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 partially capture the epistemic 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. We could potentially attribute this behaviour to the reduced effective dataset size used to train each member of the ensemble. In addition, a positive effect with respect to ensemble size is shown.

Not evident from Figure 1 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 the coverage. This highlights the simulation cost associated with diagnosing non-amortized approaches. In opposition, amortized methods do not require retraining or new simulations to determine the empirical expected coverage probability of a posterior estimator. For this reason, a global coverage analysis of non-amortized approaches is computationally prohibitive and mostly impractical. More importantly, the coverage analysis of a non-amortized approach only measures the quality of the training procedure, whereas a coverage analysis of an amortized approach diagnoses the posterior estimator itself. In addition, a global coverage analysis not only serves as diagnostic but also allows to partially alleviate the issue by performing post-training calibration. A simple way for calibrating level α credible regions is to replace those by credible regions at a level that has the desired expected coverage. Finally, non-amortized sequential algorithms have to repeat the entire simulation-training pipeline whenever architectural or hyperparameter changes are made, while amortized methods reuse previously simulated datasets. All of the above lead us to conclude that while sequential methods have the benefit of being faster to train, amortized methods should be considered for sensitive applications requiring detailed statistical validation.

4 Discussion

As demonstrated empirically, simulation-based inference can be unreliable, especially whenever its approximations cannot be diagnosed. We are of the opinion that theoretical and methodological advances within the field of simulation-based inference will strengthen its reliability and thereby 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 optimality conditions, constitutes a valuable research avenue. From a Bayesian perspective, Rozet et al. [13] propose using the focal and the peripheral losses to weigh down strongly classified samples as a means to tune the conservativeness of a posterior estimator. However, the technique is empirical and requires tuning to attain the desired properties in practice. Dalmasso et al. [12] on the other hand consider the frequentist setting and introduce a theoretically-grounded algorithm for the construction of confidence intervals that are guaranteed to have calibrated coverage, regardless of the quality of the used statistic. Dalmasso et al. [14] extends this work with finite sample guarantees. Second, in light of our results that ensembles produce more conservative posteriors, model averaging constitutes another promising direction of study as a simple and directly applicable 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 toward more conservative inference. To some extent, this

has already been considered for amortized methods [7, 8, 10] and would be worth exploring further, especially for non-amortized approaches.

In summary, we show that current algorithms for simulation-based inference can produce over-confident posterior approximations, making them possibly unreliable for scientific use cases and falsificationist inquiry. Nevertheless, we remain confident and optimistic and advocate that our results are only a stepping stone toward more reliable simulation-based inference and its wider adoption in the sciences.

Broader impact

Our work constitutes an empirical demonstration of failure modes of current simulation-based algorithms. In this regard, we believe that it could only have a positive impact by preventing practitioners from applying such algorithms without performing diagnostics and hence prevent them from drawing wrong scientific conclusions.

Acknowledgments and Disclosure of Funding

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.

Do **not** include this section in the anonymized submission, only in the final paper. You can use the `ack` environment provided in the style file to automatically hide this section in the anonymized submission.

References

[1] Kyle Cranmer et al. "The frontier of simulation-based inference". In: *Proceedings of the National Academy of Sciences* (2020).

[2] Jan-Matthis Lueckmann et al. "Benchmarking Simulation-Based Inference". In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Ed. by Arindam Banerjee et al. Vol. 130. Proceedings of Machine Learning Research. PMLR, 13–15 Apr 2021, pp. 343–351.

[3] Georges Aad et al. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 1–29.

[4] Daniel Gilman et al. "Probing the nature of dark matter by forward modelling flux ratios in strong gravitational lenses". In: *Monthly Notices of the Royal Astronomical Society* 481.1 (2018), pp. 819–834.

[5] N. Aghanim et al. "Planck 2018 results. VI. Cosmological parameters". In: *Astron. Astrophys.* 641 (2020). [Erratum: Astron.Astrophys. 652, C4 (2021)], A6. DOI: 10.1051/0004-6361/201833910. arXiv: 1807.06209 [astro-ph.CO].

[6] Benjamin P Abbott et al. "GW151226: observation of gravitational waves from a 22-solar-mass binary black hole coalescence". In: *Physical review letters* 116.24 (2016), p. 241103.

[7] Kyle Cranmer et al. "Approximating likelihood ratios with calibrated discriminative classifiers". In: *arXiv preprint arXiv:1506.02169* (2015).

[8] Johann Brehmer et al. "A Guide to Constraining Effective Field Theories with Machine Learning". In: *Phys. Rev. D* 98.5 (2018), p. 052004. DOI: 10.1103/PhysRevD.98.052004. arXiv: 1805.00020 [hep-ph].

[9] Johann Brehmer et al. "Mining for Dark Matter Substructure: Inferring subhalo population properties from strong lenses with machine learning". In: *The Astrophysical Journal* 886.1 (2019), p. 49.

[10] Joeri Hermans et al. “Towards constraining warm dark matter with stellar streams through neural simulation-based inference”. In: *Monthly Notices of the Royal Astronomical Society* 507.2 (2021), pp. 1999–2011.

[11] Sean Talts et al. “Validating Bayesian inference algorithms with simulation-based calibration”. In: *arXiv preprint arXiv:1804.06788* (2018).

[12] Niccolò Dalmasso et al. “Confidence Sets and Hypothesis Testing in a Likelihood-Free Inference Setting”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2323–2334.

[13] François Rozet et al. “Arbitrary Marginal Neural Ratio Estimation for Likelihood-free Inference”. MA thesis. University of Liège, Belgium, 2021.

[14] Niccolò Dalmasso et al. “Likelihood-Free Frequentist Inference: Bridging Classical Statistics and Machine Learning in Simulation and Uncertainty Quantification”. In: *arXiv preprint arXiv:2107.03920* (2021).

[15] George Papamakarios et al. “Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 837–848.

[16] Kyle Cranmer et al. *“Active Sciencing” with Reusable Workflows*. `https://github.com/cranmer/active_sciencing`. 2017.

[17] Michael GB Blum et al. “Non-linear regression models for Approximate Bayesian Computation”. In: *Statistics and computing* 20.1 (2010), pp. 63–73.

[18] Alfred J Lotka. “Analytical note on certain rhythmic relations in organic systems”. In: *Proceedings of the National Academy of Sciences* 6.7 (1920), pp. 410–415.

[19] Vito Volterra. “Fluctuations in the abundance of a species considered mathematically”. In: *Nature* 118.2972 (1926), pp. 558–560.

[20] Nilanjan Banik et al. “Probing the nature of dark matter particles with stellar streams”. In: *Journal of Cosmology and Astroparticle Physics* 2018.07 (2018), p. 061.

[21] LIGO Scientific Collaboration. *LIGO Algorithm Library - LALSuite*. free software (GPL). 2018. DOI: `10.7935/GT1W-FZ16`.

[22] C. M. Biwer et al. “PyCBC Inference: A Python-based parameter estimation toolkit for compact binary coalescence signals”. In: *Publ. Astron. Soc. Pac.* 131.996 (2019), p. 024503. DOI: `10.1088/1538-3873/aaef0b`. arXiv: `1807.10312 [astro-ph.IM]`.

[23] Owen Thomas et al. “Likelihood-free inference by ratio estimation”. In: *Bayesian Analysis* (2016).

[24] Joeri Hermans et al. “Likelihood-free MCMC with Amortized Approximate Ratio Estimators”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III et al. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 4239–4248.

[25] Conor Durkan et al. “On contrastive learning for likelihood-free inference”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2771–2781.

[26] Masashi Sugiyama et al. “Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation”. In: *Annals of the Institute of Statistical Mathematics* 64.5 (2012), pp. 1009–1044.

[27] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014).

[28] Danilo Rezende et al. “Variational inference with normalizing flows”. In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538.

[29] Laurent Dinh et al. “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.

[30] Laurent Dinh et al. “Density estimation using Real NVP”. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. 2017.

[31] Donald B. Rubin. “Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician”. In: *The Annals of Statistics* 12.4 (1984), pp. 1151–1172.

[32] Jonathan K Pritchard et al. "Population growth of human Y chromosomes: a study of Y chromosome microsatellites." In: *Molecular biology and evolution* 16.12 (1999), pp. 1791–1798.

[33] Tina Toni et al. "Simulation-based model selection for dynamical systems in systems and population biology". In: *Bioinformatics* 26.1 (Oct. 2009), pp. 104–110. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp619.

[34] Scott A Sisson et al. "Sequential monte carlo without likelihoods". In: *Proceedings of the National Academy of Sciences* 104.6 (2007), pp. 1760–1765.

[35] Mark A Beaumont et al. "Adaptive approximate Bayesian computation". In: *Biometrika* 96.4 (2009), pp. 983–990.

[36] George Papamakarios et al. "Fast ε-free inference of simulation models with bayesian conditional density estimation". In: *Advances in neural information processing systems*. 2016, pp. 1028–1036.

[37] Jan-Matthis Lueckmann et al. "Flexible statistical inference for mechanistic models of neural dynamics". In: *Advances in Neural Information Processing Systems* 30 (2017).

[38] David Greenberg et al. "Automatic posterior transformation for likelihood-free inference". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2404–2414.

[39] Günter Klambauer et al. "Self-normalizing neural networks". In: *Proceedings of the 31st international conference on neural information processing systems*. 2017, pp. 972–981.

[40] Diederick P Kingma et al. "Adam: A method for stochastic optimization". In: *International Conference on Learning Representations (ICLR)*. 2015.

[41] Ilya Loshchilov et al. "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations*. 2019.

[42] Alvaro Tejero-Cantero et al. "sbi: A toolkit for simulation-based inference". In: *The Journal of Open Source Software* 5.52 (2020), p. 2505.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] There are none.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] There are none.
   (b) Did you include complete proofs of all theoretical results? [N/A] There are none.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We did so to preserve anonymity. Those will be available after the review process.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] see Appendix C
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Multiple runs have been performed and we report the mean.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [No]

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] There are none.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] There are none.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] There are none.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] There are none.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] There are none.

A Benchmarks

A.1 Description

We consider 7 benchmarks, ranging from a toy problem to real scientific use cases covering various disciplines. All benchmarks and priors are available in the codebase.

The **SLCP** simulator models a fictive problem with 5 parameters. The observable $\boldsymbol{x} \in \mathbb{R}^8$ represents 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 [15] 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 [16] concerns a simulation of high energy particle collisions $e^+e^- \to \mu^+\mu^-$. The angular distribution 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 [17] 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 [18, 19] 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 predators and preys. 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 [20] 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 [21, 22].

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 ± 0.96	9.18 ± 0.28	545.13 ± 23.63	$39,369 \pm 584$

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.

A.2 Expected simulation times

B 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 used for each inference algorithm, including hyperparameters, are listed in Appendix C.

B.1 Amortized

Neural Ratio Estimation (NRE) is an established approach in the simulation-based inference literature both from a frequentist [7] and Bayesian [23–25] 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 classifier $\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})$ 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, with equal label marginal probability. Similar to the density-ratio trick [7, 24, 26, 27], the Bayes optimal classifier $d(\boldsymbol{\vartheta}, \boldsymbol{x})$ for the cross-entropy loss is

$$d(\boldsymbol{\vartheta}, \boldsymbol{x}) = \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta}, \boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})} = \sigma\left(\log \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\right), \tag{4}$$

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 classifier $\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})$ approximates the log likelihood-to-evidence ratio $\log r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ through the logit function because $\operatorname{logit}(\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})) \approx \log r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$. The approximate log posterior density function is $\log p(\boldsymbol{\vartheta}) + \log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$.

Neural Posterior Estimation (NPE) [28] 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 [28–30] parameterized by ψ. They are usually optimized using variational inference, by solving $\arg\min_\psi \mathbb{E}_{p(\boldsymbol{x})}\left[\mathrm{KL}(p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \mid\mid \hat{p}_\psi(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right]$, 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 directly be evaluated and sampled from.

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 $\frac{1}{n}\sum_{i=1}^{n} \hat{p}_i(\boldsymbol{\vartheta} \mid \boldsymbol{x}) = p(\boldsymbol{\vartheta})\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i(\boldsymbol{x} \mid \boldsymbol{\vartheta})$.

B.2 Non-amortized

Rejection Approximate Bayesian Computation (REJ-ABC) [31, 32] 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.

Sequential methods for simulation-based inference aim to approximate a single posterior by iteratively improving a posterior approximation. These methods alternate between a simulation and an exploitation phase. The latter is 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) [33–35] 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 candidates are generated by perturbing the selected ranked proposals.

Sequential Neural Posterior Estimation (SNPE) [36–38] iteratively improves a normalizing flow that models the posterior. Our evaluations will specifically use the SNPE-C [38] variant.

Sequential Neural Likelihood (SNL) [15] 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 Estimation (SNRE) [24, 25] iteratively improves the modelled likelihood-to-evidence ratio.

C 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 $\boldsymbol{\vartheta}$.

C.1 Amortized

C.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 [39] 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.

C.1.2 Neural Ratio Estimation

Our experiments use the ADAMW [40, 41] optimizer. Accross all benchmarks, the MLP architectures constitute of 3 hidden layers with 128 units and SELU [39] 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.

C.2 Non-amortized

All our implementations of non-amortized approaches rely on the reference implementation in `sbi` [42]. We use the recommended defaults unless stated otherwise. Whenever available, the same MLP embedding network is used. It consists of 3 hidden layers with 64 units and SELU [39] activations.

	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.

The latent space has a dimensionality of 10 features. For all sequential methods, we use 10 rounds to iteratively improve the posterior approximation. Tasks that are tagged with the **prohibitive** keyword are computationally prohibitive, but technically not intractable because the computational cost is tied to the learning of the posterior approximation and not to the underlying (intractable) likelihood model.

C.2.1 SNPE

Our evaluations with SNPE specifically use the SNPE-C [38] variant, as suggested by `sbi` [42]. 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	Prohibitive	Prohibitive
Coverage samples	300	300	300	300	300	Prohibitive	Prohibitive
Embedding	MLP	MLP	MLP	MLP	MLP	Prohibitive	Prohibitive
Epochs	100	100	100	100	100	Prohibitive	Prohibitive
Features	64	64	64	64	64	Prohibitive	Prohibitive
Model	NSF	NSF	NSF	NSF	NSF	Prohibitive	Prohibitive
Transforms	3	3	1	3	3	Prohibitive	Prohibitive
Rounds	10	10	10	10	10	Prohibitive	Prohibitive
Learning-rate	0.001	0.001	0.001	0.001	0.001	Prohibitive	Prohibitive

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

C.2.2 SNL

In contrast to other sequential methods, our evaluations with SNL [15] add two additional computationally prohibitive or Prohibitive 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 Prohibitive.

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

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

C.2.3 SNRE

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

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

C.2.4 Approximate Bayesian Computation

Our ABC implementation relies on the `MCABC` and `SMCABC` classes in the `sbi` [42] 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	Prohibitive	Prohibitive
Quantile	0.01	0.01	0.01	0.01	0.01	Prohibitive	Prohibitive

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	Prohibitive	Prohibitive
ϵ *decay*	0.5	0.5	0.5	0.5	0.5	Prohibitive	Prohibitive
Quantile	0.01	0.01	0.01	0.01	0.01	Prohibitive	Prohibitive

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

D Experimental setup

The expected coverage probability is estimated as

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

We consider n test simulations $(\boldsymbol{\vartheta}_i^*, \boldsymbol{x}_i) \sim p(\boldsymbol{\vartheta})p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$ and compute their associated approximate posteriors $\hat{p}(\boldsymbol{\vartheta}\,|\,\boldsymbol{x}_i)$ in a discretized and empirically normalized grid of the parameter space. The associated credible region is the highest density credible region, i.e. a credible region of the form

$$\Theta_{\hat{p}(\boldsymbol{\vartheta}\,|\,\boldsymbol{x}_i)}(1-\alpha) = \{\boldsymbol{\vartheta} : \hat{p}(\boldsymbol{\vartheta}\,|\,\boldsymbol{x}_i) \geq \gamma\}. \tag{6}$$

The threshold γ is computed using a dichotomic search to produce a credible region of level $1-\alpha$. We then estimate the empirical expected coverage probability by the proportion of nominal parameters $\boldsymbol{\vartheta}_i^*$ that falls in their associated credible region $\Theta_{\hat{p}(\boldsymbol{\vartheta}\,|\,\boldsymbol{x}_i)}(1-\alpha)$.

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 expected coverage probability is estimated on at least

$n = 5,000$ 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 approximate a single posterior and can therefore not reasonably evaluate expected coverage in the same way. Our experiments for non-amortized approaches estimate the expected coverage by repeating the inference procedure on 300 distinct observables for every simulation budget. The expected coverage probabilities are subsequently estimated based on the resulting posterior approximations. Our experiments with NPE, SNPE, SNL, SNRE, REJ-ABC and SMC-ABC rely on the implementation in the `sbi` package [42], while a custom implementation for NRE is used.

E Ensembles

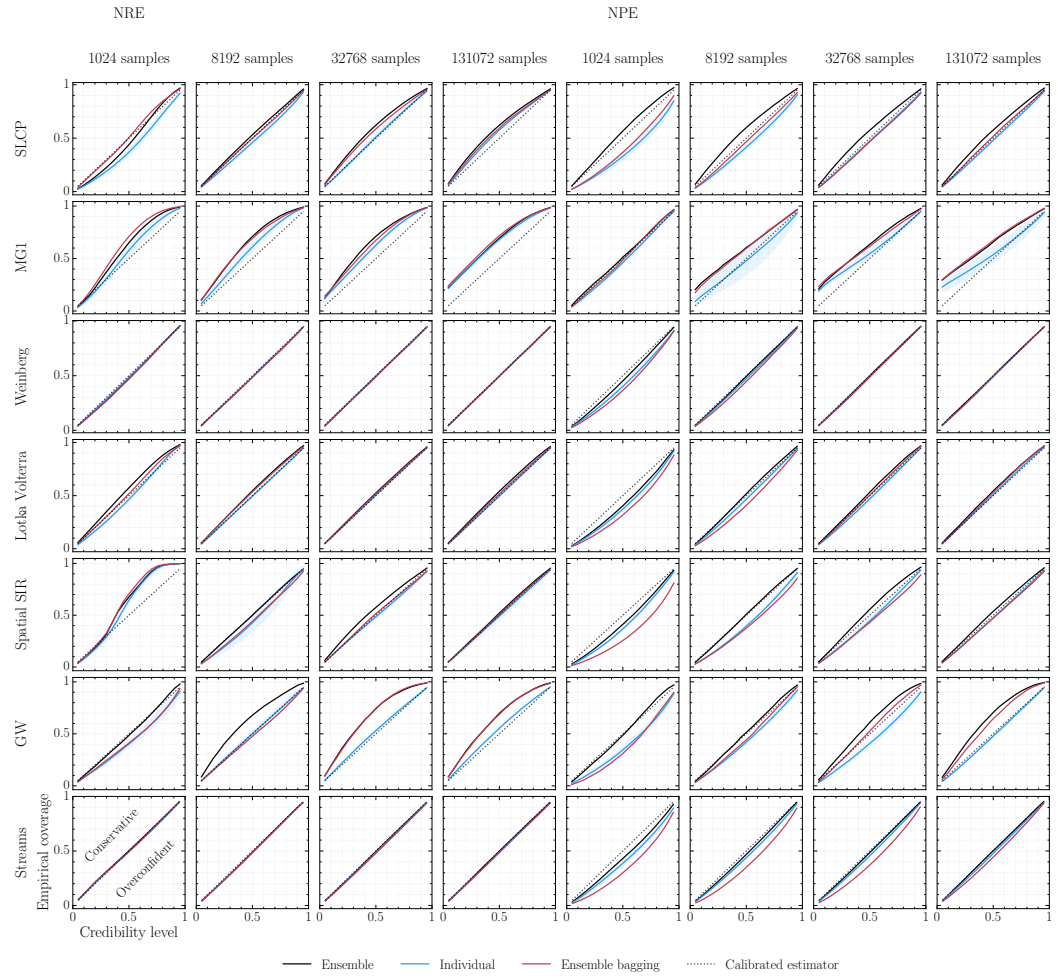

Figure 2: Analysis of coverage between ensemble and individual models w.r.t the various simulation budgets. The blue line represents the mean expected coverage of individual models over 5 runs, the shaded area represents its standard deviation. The black line represents the expected coverage of a single ensemble composed of 5 models. We observe that ensembles consistently have a higher 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 evaluated for amortized approaches such as NPE and NRE.

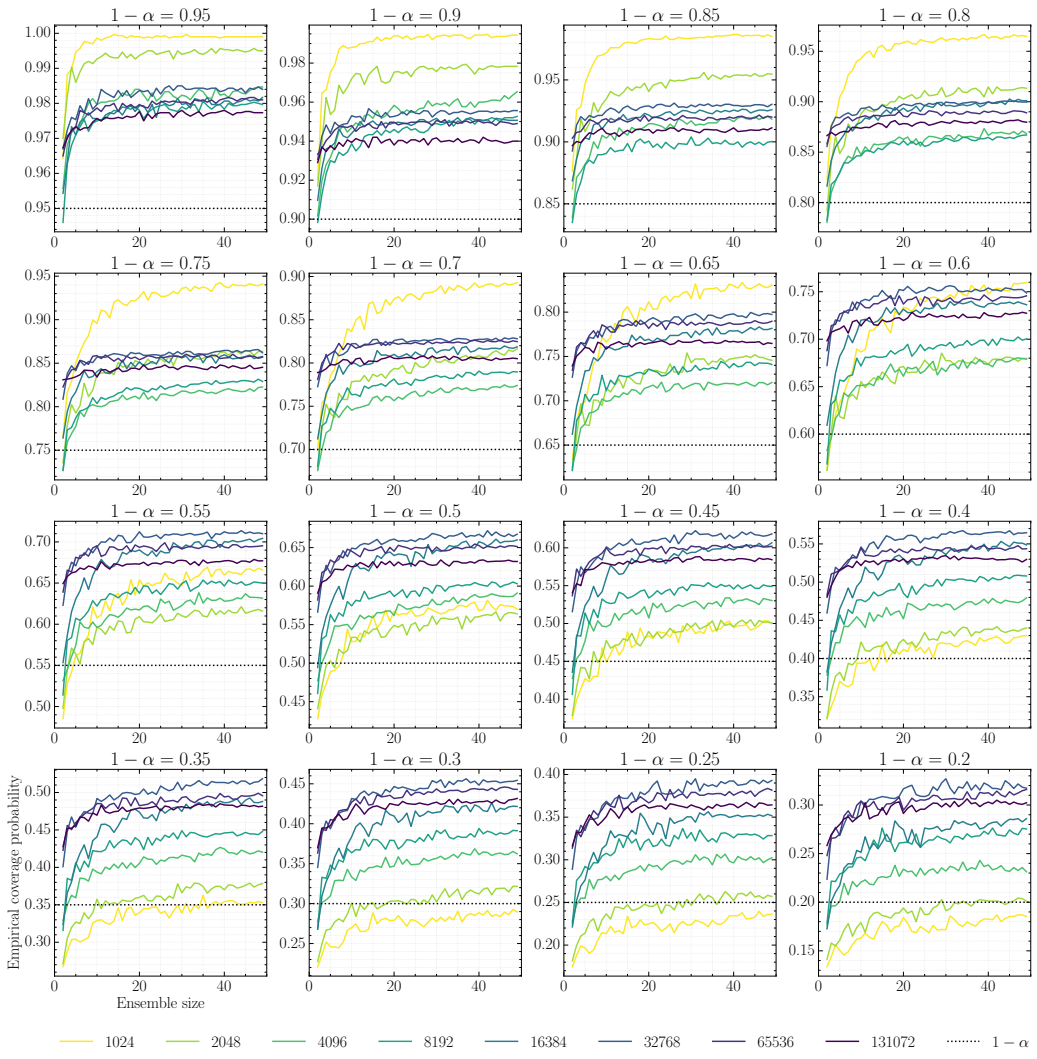

Figure 3: Evolution of the empirical expected coverage probability with respect to ensemble size for various confidence levels. The results are obtained by training 100 ratio estimators (NRE) on the SLCP benchmark. A positive effect is observed in terms of empirical expected coverage probability and 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.