Pokie: Posterior Accuracy and Model Comparison

Sammy Sharief [1 2 3] **Justine Zeghal** [1 2 3] **Gabriel Missael Barco** [1 2 3] **Pablo Lemos** [4] **Yashar Hezaveh** [1 2 3 5 6]
Laurence Perreault-Levasseur [1 2 3 5 7 6]

Abstract

We present Pokie, a sample-based method for comparing posterior distributions. Pokie estimates the expected probability that samples from an inferred posterior match the true, unknown posterior of a probabilistic model for which only joint samples are available. This framework enables direct Bayesian model comparison by assessing how each model's posterior distribution aligns with the posterior of the true model, all while avoiding evidence computation and relying solely on simulations. We show that Pokie converges to a score of $2/3$ under well-specified models and has a lower bound of $1/2$ for misspecified models. We demonstrate its effectiveness across several toy problems and cosmological inference tasks. Code: https://github.com/SammyS15/Pokie.

1. Introduction

In probabilistic modeling, where the relationship between observations x and parameters y is described by a probabilistic model $p(x, y \mid \mathcal{M})$, two fundamental challenges arise across diverse scientific fields: quantifying posterior distribution calibration in Bayesian inference (Carzon et al., 2023; Howland et al., 2022; Orozco Valero et al., 2025; Tegmark et al., 2004; Tolley et al., 2024) and conducting Bayesian model comparison (Jeffrey & Wandelt, 2024; Piironen & Vehtari, 2016; Slosar et al., 2003; Yuen, 2010). In Bayesian inference, the posterior distribution is an update of prior beliefs about parameters y after observing data x. This update is derived using Bayes' Theorem:

$$p(y \mid x, \mathcal{M}) \propto p(x \mid y, \mathcal{M}) p(y \mid \mathcal{M}).$$

[1]Department of Physics, University of Montreal, Montreal, Canada [2]MILA Quebec AI Institute, Montreal, Canada [3]CIELA Institute, Montreal Institute for Astrophysics and Machine Learning, Montreal, Canada [4]Sandbox, California, USA [5]Center for Computational Astrophysics, Flatiron Institute, New York, USA [6]Trottier Space Institute, McGill University, Montreal, Canada [7]Perimeter Institute for Theoretical Physics, Waterloo, Canada. Correspondence to: Sammy Sharief <sammy.sharief@umontreal.ca>.

ML4Astro 2025, Vancouver, CA.

Evaluating whether a posterior estimator is calibrated is crucial, particularly with the rise of implicit inference methods powered by deep learning (e.g. Cranmer et al., 2016; Papamakarios & Murray, 2018; Papamakarios et al., 2019). The ideal quality metric would compare the inferred posterior to the true posterior distribution. In a simulation-based framework, however, only joint samples $x^*, y^* \sim p(x, y \mid \mathcal{M})$ are typically available, limiting the applicability of many existing metrics (Lueckmann et al., 2021), which often assume access to the true posterior or its density.

Bayesian model comparison aims to rank competing hypotheses based on their ability to reproduce the joint behavior of observations and parameters, effectively balancing fit and complexity. Classical Bayesian approach rely on the computation of the model evidence $p(x \mid \mathcal{M}) = \int p(x \mid y, \mathcal{M}) p(y \mid \mathcal{M}) dy$ (Kass & Raftery, 1995), which is often computationally intractable, particularly in high-dimensional parameter space or simulation-based settings (Alsing et al., 2018; Spurio Mancini et al., 2023).

To address both of these challenges, we propose Pokie (Posterior over K Inference Estimations), a likelihood-free, sample-based approach designed for probabilistic posterior comparison. Building upon TARP (Lemos et al., 2023) and PQMass (Lemos et al., 2025), Pokie quantifies the expected probability that the samples of the inferred posterior distribution match the true unknown posterior distribution, using only joint fiducial samples $x^*, y^* \sim p(x, y \mid \mathcal{M})$. Pokie operates with minimal assumptions, leveraging only the Central Limit Theorem (CLT) to produce the scaled-value calibration score, referred to as the *Pokie score*. The Pokie score allows for a Bayesian model comparison by quantifying how closely each candidate model's posterior approximates that of the reference model. As a result, by shifting the comparison from data space to parameter space, Pokie enables Bayesian model comparison without requiring explicit computation of the evidence, and remains effective even in high-dimensional parameter spaces.

In summary, our contributions are as follows. We introduce Pokie as a new framework for posterior-level model comparison that avoids likelihood evaluation. We show that Pokie provides a score that converges to $\frac{2}{3}$ for well-specified models and to $\frac{1}{2}$ for poorly specified ones in the infinite-

sample limit. Finally, we demonstrate the effectiveness of our method on several tasks.

2. Method

Let $\mathcal{M}$ be a candidate model and let $\mathcal{M}^*$ be the ground-truth model. Our objective is to evaluate whether the posterior under, $p\big(y \mid x^*, \mathcal{M}\big)$, is calibrated with respect to the true posterior, $p\big(y \mid x^*, \mathcal{M}^*\big)$. We assume we only have access to posterior samples $\{y\}_{i=1}^N \sim p(y \mid x^*, \mathcal{M})$, and joint samples $x^*, y^* \sim p(x, y \mid \mathcal{M}^*)$ from the true model, i.e. we only have one sample from $p(x \mid y, \mathcal{M}^*)$.

Given that two distributions are equal if they assign the same mass on all measurable regions $\mathcal{R}$, following PQMass (Lemos et al., 2025) framework, we compare distributions by comparing the number of samples falling into randomly constructed regions $\mathcal{R}$. Formally, we denote

$$n = \sum_{i=1}^{N} \mathbf{1}\big[y_i \in \mathcal{R}\big] \quad \text{and} \quad k = \mathbf{1}\big[y^* \in \mathcal{R}\big], \tag{1}$$

where n is the number of posterior samples that fall inside the region, and k indicates whether the ground-truth parameter y^* is contained within $\mathcal{R}$. These two random variables follow Binomial distributions

$$n \sim \mathrm{B}(N, \lambda_n) \quad \text{and} \quad k \sim \mathrm{B}(1, \lambda_k)$$

with λ_n and λ_k denoting respectively the posterior mass of $p(y \mid x^*, \mathcal{M})$ and $p(y \mid x^*, \mathcal{M}^*)$ that falls in $\mathcal{R}$, that is

$$\lambda_n = \int_{\mathcal{R}} p(y \mid x^*, \mathcal{M}) dy,$$
$$\lambda_k = \int_{\mathcal{R}} p(y \mid x^*, \mathcal{M}^*) dy.$$

We define random regions $\mathcal{R}$ by first sampling a center point from the parameter space, $c \sim \pi_c$, and selecting a random posterior y_j sample from $\{y\}_{i=1}^N$. The region $\mathcal{R}$ is then defined as a hypersphere centered at c with radius $\|c - y_j\|$. This construction, based on TARP (Lemos et al., 2023), introduces stochasticity and avoids bias from fixed regions, allowing for the exploration of the entire parameter space.

The fact that only a single sample is available from the true posterior motivates us to cast the comparison in a Bayesian framework. Specifically, we derive the probability that y^* falls in the region $\mathcal{R}$ given that n samples from $\{y\}_{i=1}^N$ falls in $\mathcal{R}$, that is: $p(k|n, \mathcal{R})$. Under the null hypothesis:

$$y^* \sim p(y \mid x^*, \mathcal{M}), \tag{2}$$

(i.e., $\lambda_n = \lambda_k$, $\forall \mathcal{R}$), we derive the analytic posterior predictive probability (proof can be found in Appendix D.1):

$$p\big(k = 1 \mid n, \mathcal{R}\big) = \frac{n+1}{N+2},$$
$$p\big(k = 0 \mid n, \mathcal{R}\big) = \frac{N-n+1}{N+2}. \tag{3}$$

Averaging these probabilities across all fiducial draws and simulations defines the **Pokie score**:

$$\mathbb{P}_{\text{Pokie}}(\mathcal{M}) = \mathbb{E}_{p(Z)}\left[\mathbb{E}_{p(k,n,\mathcal{R}|Z)}\left[p(k \mid n, \mathcal{R})\right]\right] \tag{4}$$

with $Z = (y^*, x^*, \{y\}_{i=1}^N)$. The score is approximated using Monte Carlo integration as

$$\mathbb{P}_{\text{Pokie}}(\mathcal{M}) \approx \frac{1}{L} \sum_{l=1}^{L} p(k_l \mid n_l, \mathcal{R}_l), \tag{5}$$

with L the number of fiducial values. In practice, we generate several hyperspheres per fiducial to mitigate both the limited number of fiducials and the single posterior sample from the true posterior distribution. In Appendix C, we provide algorithm 1, the pseudocode for estimating the Pokie score. We demonstrate in Appendix D.2 and Appendix D.3, that within the limit of infinite samples, $P_{\text{Pokie}}(\mathcal{M})$ converges to $\frac{2}{3}$ for well-specified models ($\mathcal{M} = \mathcal{M}^*$) and to $\frac{1}{2}$ for poorly-specified models ($\mathcal{M} \neq \mathcal{M}^*$).

Pokie offers a direct approach to Bayesian model comparison by evaluating how closely the posterior distributions of candidate models match the posterior of the true model. Unlike the Bayes factor, Pokie operates directly in parameter space, thus avoiding the need for evidence computation. In contrast to the often hard-to-interpret and computationally-expensive Bayes factor, the Pokie score is a probabilistic metric with theoretical bounds between $\frac{1}{2}$ and $\frac{2}{3}$, enabling consistent interpretability across tasks. Moreover, Pokie is computationally efficient and works well even in high-dimensional parameter settings.

However, Pokie shares a fundamental limitation with the Bayes factor: the equality of posterior distributions does not imply model equivalence. However, Pokie and the Bayes factor are complementary. While the Bayes factor assesses models based on evidence, Pokie focuses on the agreement of posterior distributions. This distinction is crucial when models yield similar evidence but diverge in their posterior structure, in which case the Bayes factor may be misleading. For instance, a failure of the Bayes factor can arise when using Expectation Maximization to update the prior (Barco et al., 2025; Rozet et al., 2024). While this method can maximize the evidence and yield a favorable Bayes factor, it may simultaneously produce inaccurate posteriors. A more comprehensive comparison is left for future work.

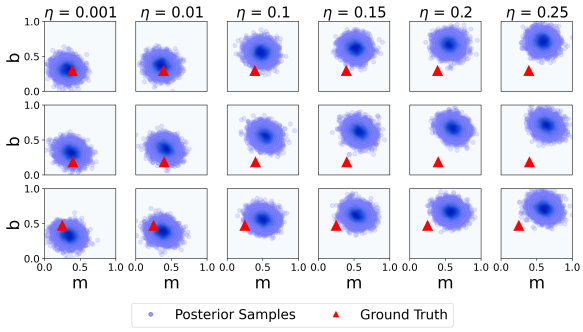

Figure 1. We display a visualization of posterior distributions under varying noise levels η. Each column corresponds to a different noise level, while each row shows one of three representative ground-truth draws (red triangles). For each case, we plot the posterior samples (blue dots and density contours) from the analytically computed posterior.

3. Experiments

We apply Pokie (Section 2) to the following experiments. Parameters are normalized to $[0, 1]$, and centers c_j are sampled from $\mathcal{U}(0, 1)$. Distances are calculated using the L2 Distance, and unless otherwise noted, all experiments are run on an M2 MacBook Air with 8 GB of RAM. For all experiments, we generate 100 hyperspheres per fiducial.

3.1. Linear regression

We consider a linear regression task where we aim to infer the posterior distribution over weights $\theta = [m, b]$ of the linear model $y = mx + b + \eta$ with $\eta \sim \mathcal{N}(0, \sigma^2)$. We chose a Gaussian prior $p(\theta) = \mathcal{N}(\mu_0, \Sigma_0)$ to infer the posterior distribution $p(\theta|y)$. Because both the likelihood and prior distributions are Gaussian, we can derive an analytical form of the posterior distribution $p(\theta|y) = \mathcal{N}(\theta|\mu_{\text{post}}, \Sigma_{\text{post}})$, where

$$\Sigma_{\text{post}} = \left(\Sigma_0^{-1} + A^T \Sigma_n^{-1} A\right)^{-1},$$
$$\mu_{\text{post}} = \Sigma_{\text{post}}\left(A^T \Sigma_n^{-1} y + \Sigma_0^{-1} \mu_0\right).$$

and $\Sigma_n = \sigma^2 I$ denotes the observation noise covariance matrix.

We define five models where we perturb the mean vector with increasing noise levels, $\eta = \{0.001, 0.01, 0.01, 0.10, 0.20, 0.25\}$, and choose the model with the least noise as our true model. We consider 5 000 fiducial samples, i.e. we have 5 000 posterior distributions. From each posterior, we draw 5 000 samples from our analytic posterior distribution to evaluate the model's sensitivity and determine which posterior is the best calibrated. Figure 1 visualizes a subset of these posteriors alongside their corresponding ground-truth parameters.

The results, shown in Table 1, demonstrate that across the different posteriors with varying levels of noise, Pokie can detect that the posterior with the least noise is the most accurate, as well as determine that the posteriors with increasing levels of noise are less accurate. This result showcases Pokie's ability to identify the most in-distribution posterior.

Table 1. Pokie score with 68% bootstrap confidence intervals for each noise level. We demonstrate that Pokie assigns higher calibration and probability to the model with the lowest noise.

Noise Level	Pokie Score (68% CI)
0.001	**0.6670 ± 0.0011**
0.010	0.6417 ± 0.0020
0.100	0.5669 ± 0.0005
0.150	0.5589 ± 0.0009
0.200	0.5548 ± 0.0009
0.250	0.5517 ± 0.0009

3.2. Analyzing distribution shifts

By definition, Pokie is the expectation of the probability $p(k \mid n, \mathcal{R})$ over fiducial values, and we demonstrated in subsection D.2 and subsection D.3 its upper and lower bounds. In this experiment, we aim to test the distributional shift of a unique distribution, i.e., we aim to test if Pokie can detect misspecification using only one fiducial value. For this, we use Gaussian Mixture Models (GMMs) of 100 dimensions and 20 mixture components as our unique posterior distribution without performing Bayesian inference. The means and variances of each component of the true model are randomly selected. The other posteriors are built by introducing a shift of the vector of ones multiplied by l, along the diagonal direction (a 2-d version of the GMMs can be found in Figure 2). This setup simulates a scenario with generative models that are either in- or out-of-distribution. From each GMM, both truth and shifted, we generate 5 000 samples. We then run Pokie to test if it can detect a distribution shift using only one fiducial. Our result in Table 2 shows that Pokie correctly identifies the model with $l = 0$ as the best-calibrated, while classifying the others as out-of-distribution. This highlights Pokie's ability to detect shifts using a single posterior distribution.

Table 2. Pokie score with 68% bootstrap confidence intervals for each shift level. We demonstrate that Pokie assigns higher calibration and probability to the true posterior, validating its ability to detect shifts using a single posterior.

Model Shift	Pokie Score (68% CI)
-6	0.5002 ± 0.0003
-3	0.5115 ± 0.0006
0	**0.6669 ± 0.0003**
+3	0.5156 ± 0.0003
+6	0.5000 ± 0.0004

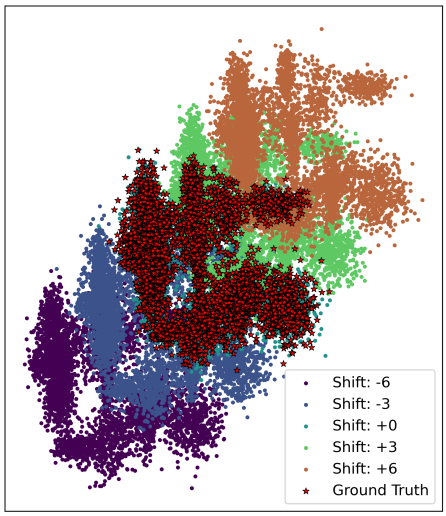

Figure 2. 2-dimensional version of the GMM experiment. We show the true GMM distribution (red star) with the 4 different GMM posteriors with shifts, l, in the mean vector.

3.3. Astrophysics parameter inference: detecting physical model shifts

We apply Pokie to the joint inference of lens and source parameters in strong gravitational lensing. Here, the goal is to evaluate whether Pokie can detect when the underlying physical model is misspecified, either in the lens mass profile or the number of background sources.

Following Filipp et al. (2024), we generate lenses with multiple Sérsic components (Sérsic, 1963) to model complex background source morphologies. For our background sources, we define each model to contain either one or three Sérsic profiles. All lenses are generated using `caustics` [1](Stone et al., 2024). Details on lens simulations can be found in Appendix E.

We generate 100 synthetic observations using an elliptical power-law (EPL) profile applied to three Sérsic sources. We define four candidate models that vary in lens type, EPL vs. singular isothermal ellipsoid (SIE) and source count (one vs. three Sérsic components): EPL + 3 (correct), SIE + 3 (incorrect lens), EPL + 1 (incorrect source count), and SIE + 1 (both incorrect) (see Appendix E). Each models share the same prior distribution over parameters.

We use Metropolis-adjusted Langevin sampling (MALA, Roberts & Tweedie, 1996) to get our 100 posterior distributions, each with 20 000 samples in a 13-dimensional parameter space. After sampling, we apply Pokie to evaluate how well the posterior samples match the fiducial parameters of the observed image and perform model ranking. Results are shown in Table 3.3. We see that EPL + 3 Sérsic sources obtain the highest Pokie score, which makes intuitive sense as it follows the correct data generation process. We note that having the correct number of sources is more important than having the correct lens profile. These results show that Pokie reliably detects model misspecification.

[1]https://github.com/Ciela-Institute/caustics

Table 3. Pokie score with 68% bootstrap confidence intervals. We demonstrate that Pokie identifies the correct lensing model in a likelihood misspecification problem.

Likelihood	Pokie Score (68% CI)
EPL + 3 Sersic Sources	**0.6297 ± 0.0054**
SIE + 3 Sersic Sources	0.5777 ± 0.0027
EPL + 1 Sersic Sources	0.5277 ± 0.0028
SIE + 1 Sersic Sources	0.5267 ± 0.0031

3.4. Strong lensing background source reconstruction: detecting prior distribution shifts

We apply Pokie to the inference of pixelated background sources in strong gravitational lensing. For this, we use the score-based models (SBM) method from Barco et al., 2025 to iteratively learn the prior distribution and get posterior distributions.

We use the same lensing forward model as Barco et al., 2025, and consider 4 different simulation models, by using different prior and different Gaussian additive noise $y = Ax + \eta$, with $\eta \sim \mathcal{N}(0, \sigma_\eta)$: (1) spiral galaxies $p_s(x)$ and $\sigma_{\boldsymbol{\eta}} = 2$, (2) spiral galaxies $p_s(x)$ and $\sigma_{\boldsymbol{\eta}} = 0.5$, (3) elliptical galaxies $p_e(x)$ and $\sigma_{\boldsymbol{\eta}} = 2$, and (4) elliptical galaxies $p_e(x)$ and $\sigma_{\boldsymbol{\eta}} = 0.5$. The true model is the one with spiral galaxy prior $x \sim p_s(x)$ and, $\sigma_{\boldsymbol{\eta}} = 2$. We consider 16 observations and corresponding fiducial parameters. For each observation, we generate 64 posterior samples from our SBM. Some observations y, ground truths x^*, and posterior samples under each configuration are shown in Appendix F. We then run Pokie to evaluate the 4 models.

Table 4. Pokie score with 68% bootstrap confidence intervals. We demonstrate that Pokie assigns a higher score to the lensing model with the correct prior and noise level.

Prior and Noise Level	Pokie Score (68% CI)
$p_s(x)$ **and** $\sigma_{\boldsymbol{\eta}} = 2$	**0.6518 ± 0.0369**
$p_s(x)$ and $\sigma_{\boldsymbol{\eta}} = 0.5$	0.5728 ± 0.0089
$p_e(x)$ and $\sigma_{\boldsymbol{\eta}} = 2$	0.5214 ± 0.0168
$p_e(x)$ and $\sigma_{\boldsymbol{\eta}} = 0.5$	0.5085 ± 0.0069

The results in Table 4, show that Pokie favors the best model (first row), demonstrating Pokie ability to scale well with dimensionality, ($3 \times 64 \times 64$ pixels), as well as its sensitivity to detecting distribution shift in the prior regime for complex astrophysical data, even in low samples and fiducial regime.

4. Discussions and Conclusion

We introduce Pokie, a sample-based metric for evaluating posterior calibration and model comparison. Pokie quantifies the expected probability that samples from an inferred posterior distribution match those from the true, unknown posterior, using only joint samples from the probabilistic model. This framework allows for model comparisons directly in parameters, bypassing the need for analytical likelihood and computation of model evidence. We showed that Pokie has well-defined theoretical bounds: it converges to $\frac{2}{3}$ for well-calibrated models and has a lower bound of $\frac{1}{2}$ for misspecified ones. Our experiments demonstrate Pokie's ability to identify out-of-distribution posteriors, model misspecification, and out-of-distribution priors, and rank multiple models effectively. Its scalability, interpretability, and reliability make Pokie a practical and principled alternative to existing methods for assessing posterior calibration and model comparison.

We conducted two additional studies to support these claims. First, in Appendix G we compare the Pokie score to the Bayes Factor (BF) for experiments where the BF is tractable. For these experiments, we observe agreement between the two metrics. Then, in Appendix H we present a sensitivity analysis, varying the model dimensionality, the number of posterior samples, the number of hyperspheres per fiducial, and the number of distinct posterior distributions. We demonstrate that Pokie is robust across these variations.

While we have demonstrated that Pokie performs well across multiple experiments, there are important limitations to consider. First, Pokie relies on a sufficient number of fiducial and posterior samples to produce reliable estimates; otherwise, Pokie may become noisy or uninformative. Second, similarly to PQMass, Pokie assumes that the samples are independent and identically distributed (i.i.d.); violations of this assumption will render Pokie unusable. Pokie requires access to the ground-truth parameters, limiting its applicability to real data. Finally, Pokie is only a necessary condition for correctness. In future work, we will explore the sufficiency condition as well as delve deeper into the complementarity of the Bayes Factor and Pokie.

Acknowledgements

This work is partially supported by Schmidt Futures, a philanthropic initiative founded by Eric and Wendy Schmidt as part of the Virtual Institute for Astrophysics (VIA). The work is in part supported by computational resources provided by Calcul Quebec and the Digital Research Alliance of Canada. Y.H. and L.P. acknowledge support from the Canada Research Chairs Program, the National Sciences and Engineering Council of Canada through grants RGPIN-2020-05073 and 05102.

References

Alsing, J., Wandelt, B., and Feeney, S. Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology. , 477(3):2874–2885, July 2018. doi: 10.1093/mnras/sty819.

Barco, G. M., Adam, A., Stone, C., Hezaveh, Y., and Perreault-Levasseur, L. Tackling the problem of distributional shifts: Correcting misspecified, high-dimensional data-driven priors for inverse problems. *The Astrophysical Journal*, 980(1):108, February 2025. ISSN 1538-4357. doi: 10.3847/1538-4357/ad9b92. URL http://dx.doi.org/10.3847/1538-4357/ad9b92.

Carzon, J., Abreu, B., Regayre, L., Carslaw, K., Deaconu, L., Stier, P., Gordon, H., and Kuusela, M. Statistical constraints on climate model parameters using a scalable cloud-based inference framework. *Environmental Data Science*, 2:e24, 2023. doi: 10.1017/eds.2023.12.

Cranmer, K., Pavez, J., and Louppe, G. Approximating likelihood ratios with calibrated discriminative classifiers, 2016. URL https://arxiv.org/abs/1506.02169.

Filipp, A., Hezaveh, Y., and Perreault-Levasseur, L. Robustness of neural ratio and posterior estimators to distributional shifts for population-level dark matter analysis in strong gravitational lensing, 2024. URL https://arxiv.org/abs/2411.05905.

Howland, M. F., Dunbar, O. R. A., and Schneider, T. Parameter uncertainty quantification in an idealized gcm with a seasonal cycle. *Journal of Advances in Modeling Earth Systems*, 14(3):e2021MS002735, 2022. doi: https://doi.org/10.1029/2021MS002735. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002735. e2021MS002735 2021MS002735.

Jeffrey, N. and Wandelt, B. D. Evidence networks: simple losses for fast, amortized, neural bayesian model comparison. *Machine Learning: Science and Technology*, 5(1):015008, January 2024. ISSN 2632-2153. doi: 10.1088/2632-2153/ad1a4d. URL http://dx.doi.org/10.1088/2632-2153/ad1a4d.

Kass, R. E. and Raftery, A. E. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Lemos, P., Coogan, A., Hezaveh, Y., and Perreault-Levasseur, L. Sampling-based accuracy testing of posterior estimators for general inference, 2023. URL https://arxiv.org/abs/2302.03026.

Lemos, P., Sharief, S., Malkin, N., Salhi, S., Stone, C., Perreault-Levasseur, L., and Hezaveh, Y. Pqmass: Probabilistic assessment of the quality of generative models using probability mass estimation, 2025. URL https://arxiv.org/abs/2402.04355.

Lueckmann, J.-M., Boelts, J., Greenberg, D. S., Gonçalves, P. J., and Macke, J. H. Benchmarking simulation-based inference, 2021. URL https://arxiv.org/abs/2101.04653.

Orozco Valero, A., Rodríguez-González, V., Montobbio, N., Casal, M. A., Tlaie, A., Pelayo, F., Morillas, C., Poza, J., Gómez, C., and Martínez-Cañada, P. A python toolbox for neural circuit parameter inference. *NPJ Systems Biology and Applications*, 11(1):45, May 2025. doi: 10.1038/s41540-025-00527-9.

Papamakarios, G. and Murray, I. Fast ϵ-free inference of simulation models with bayesian conditional density estimation, 2018. URL https://arxiv.org/abs/1605.06376.

Papamakarios, G., Sterratt, D. C., and Murray, I. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows, 2019. URL https://arxiv.org/abs/1805.07226.

Piironen, J. and Vehtari, A. Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, April 2016. ISSN 1573-1375. doi: 10.1007/s11222-016-9649-y. URL http://dx.doi.org/10.1007/s11222-016-9649-y.

Roberts, G. O. and Tweedie, R. L. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341 – 363, 1996.

Rozet, F., Andry, G., Lanusse, F., and Louppe, G. Learning diffusion priors from observations by expectation maximization. *Advances in Neural Information Processing Systems*, 37:87647–87682, 2024.

Sérsic, J. L. Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Boletin de la Asociacion Argentina de Astronomia La Plata Argentina*, 6:41–43, February 1963.

Slosar, A., Carreira, P., Cleary, K., Davies, R. D., Davis, R. J., Dickinson, C., Genova-Santos, R., Grainge, K., Gutierrez, C. M., Hafez, Y. A., Hobson, M. P., Jones, M. E., Kneissl, R., Lancaster, K., Lasenby, A., Leahy, J. P., Maisinger, K., Marshall, P. J., Pooley, G. G., Rebolo, R., Rubino-Martin, J. A., Rusholme, B., Saunders, R. D. E., Savage, R., Scott, P. F., Sosa Molina, P. J., Taylor, A. C., Titterington, D., Waldram, E., Watson, R. A., and Wilkinson, A. Cosmological parameter estimation and bayesian model comparison using very small array data. *Monthly Notices of the Royal Astronomical Society*, 341 (4):L29–L34, June 2003. ISSN 1365-2966. doi: 10.1046/j.1365-8711.2003.06564.x. URL http://dx.doi.org/10.1046/j.1365-8711.2003.06564.x.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Spurio Mancini, A., Docherty, M. M., Price, M. A., and McEwen, J. D. Bayesian model comparison for simulation-based inference. *RAS Techniques and Instruments*, 2(1):710–722, January 2023. ISSN 2752-8200. doi: 10.1093/rasti/rzad051. URL http://dx.doi.org/10.1093/rasti/rzad051.

Stone, C., Adam, A., Coogan, A., Yantovski-Barth, M. J., Filipp, A., Setiawan, L., Core, C., Legin, R., Wilson, C., Barco, G. M., Hezaveh, Y., and Perreault-Levasseur, L. Caustics: A python package for accelerated strong gravitational lensing simulations, 2024. URL https://arxiv.org/abs/2406.15542.

Tegmark, M., Strauss, M. A., Blanton, M. R., Abazajian, K., Dodelson, S., Sandvik, H., Wang, X., Weinberg, D. H., Zehavi, I., Bahcall, N. A., Hoyle, F., Schlegel, D., Scoccimarro, R., Vogeley, M. S., Berlind, A., Budavari, T., Connolly, A., Eisenstein, D. J., Finkbeiner, D., Frieman, J. A., Gunn, J. E., Hui, L., Jain, B., Johnston, D., Kent, S., Lin, H., Nakajima, R., Nichol, R. C., Ostriker, J. P., Pope, A., Scranton, R., Seljak, U., Sheth, R. K., Stebbins, A., Szalay, A. S., Szapudi, I., Xu, Y., Annis, J., Brinkmann, J., Burles, S., Castander, F. J., Csabai, I., Loveday, J., Doi, M., Fukugita, M., Gillespie, B., Hennessy, G., Hogg, D. W., Ivezić, Ž., Knapp, G. R., Lamb, D. Q., Lee, B. C., Lupton, R. H., McKay, T. A., Kunszt, P., Munn, J. A., O'Connell, L., Peoples, J., Pier, J. R., Richmond, M., Rockosi, C., Schneider, D. P., Stoughton, C., Tucker, D. L., vanden Berk, D. E., Yanny, B., and York, D. G. Cosmological parameters from SDSS and WMAP. , 69(10): 103501, May 2004. doi: 10.1103/PhysRevD.69.103501.

Tolley, N., Rodrigues, P. L. C., Gramfort, A., and Jones, S. R. Methods and considerations for estimating parameters in biophysically detailed neural models with simulation based inference. *PLoS Computational Biology*, 20(2): e1011108, Feb 2024. doi: 10.1371/journal.pcbi.1011108.

Yuen, K.-V. Recent developments of bayesian model class selection and applications in civil engineering. *Structural Safety*, 32(5):338–346, 2010. ISSN 0167-4730. doi: https://doi.org/10.1016/j.strusafe.2010.03.011. URL https://www.sciencedirect.com/science/article/pii/S0167473010000305. Probabilistic Methods for Modeling, Simulation and Optimization of Engineering Structures under Uncertainty in honor of Jim Beck's 60th Birthday.

A. TARP

TARP (Lemos et al., 2023) is a method for estimating coverage probabilities of generative posterior estimators using only posterior samples, without requiring access to explicit posterior densities. This is particularly valuable in high-dimensional inference problems where density evaluations are unavailable or computationally prohibitive. TARP provides a way to validate whether posterior samples accurately reflect the true distribution, even in simulation-based settings where traditional methods fail.

TARP constructs coverage regions in parameter space. Specifically, given a true parameter θ^*, TARP defines a hypersphere centered at a randomly chosen reference point, θ_r, with a radius defined to be $d(\theta^*, \theta_r)$. The coverage is then estimated as the proportion of posterior samples that fall within this region:

$$f_i = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}[d(\theta_{ij}, \theta_r) < d(\theta_i^*, \theta_r)]$$

where $\theta_{ij} \sim \hat{p}(\theta \mid x)$ are posterior samples. TARP's key theoretical insight is that if this expected coverage holds uniformly over random choices of θ_r, then the posterior samples are guaranteed to be calibrated. This setup allows TARP to validate the accuracy of posterior inference without requiring likelihood evaluations or explicit density functions.

Pokie adopts the TARP's framework of working in the parameter space and utilizing the hypersphere setup to perform sample-based analysis, but modifies it in two key ways. First, instead of defining the region radius via the distance to the true parameter θ^*, Pokie defines the radius from θ_r to a randomly chosen posterior sample. This allows Pokie to define posterior quantile-like regions without needing knowledge of θ^* when defining the region itself. Second, Pokie introduces k, a Bernoulli variable, which records whether θ^* falls inside the randomly defined region. This formulation enables a probabilistic scoring mechanism that aggregates over many such randomized comparisons, yielding a theoretically bounded metric that discriminates between well and poorly calibrated models.

B. PQMass

PQMass (Lemos et al., 2025) is a sample-based method designed to assess whether two sets of samples originate from the same underlying distribution. The fundamental idea is to compare probability masses over multiple regions of the sample space, leveraging the properties of multinomial distributions.

Formally, given two distributions p and q, they are considered equal if their probability measures coincide over all

measurable sets $\mathcal{R} \subseteq \Omega$:

$$\mathbb{P}_p(\mathcal{R}) = \mathbb{P}_q(\mathcal{R}) \quad \forall \mathcal{R} \subseteq \Omega. \tag{6}$$

The probability mass of a region $\mathcal{R}$ under p can be unbiasedly estimated as:

$$\mathbb{P}_p(\mathcal{R}) = \mathbb{E}_{y\sim p(y)}[\mathbb{1}(y \in \mathcal{R})] \approx \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(y_i \in \mathcal{R}), \tag{7}$$

where $y_i \sim p(y)$ are N independent samples. Furthermore, given a set of N samples $y_i \sim p(y)$, and one region, the number of samples falling within $\mathcal{R}$ follows a binomial distribution:

$$n \sim \mathcal{B}(N, \lambda), \quad \lambda = \mathbb{P}_p(\mathcal{R}). \tag{8}$$

This property allows for the comparison of two distributions by comparing the binomial distributions over multiple chosen regions $\mathcal{R}$.

C. Pokie algorithm

Algorithm 1 Computing Pokie score for model $\mathcal{M}$

Input: Number of fiducial draws L, region samples L_r, posterior samples N
Output: Pokie score $P_{\text{Pokie}}(\mathcal{M})$
Initialize $\texttt{score} \leftarrow 0$
for $j = 1$ to L **do**
  Draw $y_j^* \sim p(y \mid x^*, \mathcal{M}^*)$
  Draw $\{y_{i,j}\}_{i=1}^N \sim p(y \mid x^*, \mathcal{M})$
  for $\ell = 1$ to L_r **do**
    Sample $c_{j,\ell} \sim \pi_c$
    $r_{j,\ell} \leftarrow d(c_{j,\ell}, y_{i,j})$
    $\mathcal{R}_{j,\ell} \leftarrow \{y : d(y, c_{j,\ell}) \le r_{j,\ell}\}$
    $n \leftarrow \sum_i \mathbf{1}[y_{i,j} \in \mathcal{R}_{j,\ell}]$
    $k \leftarrow \mathbf{1}[y_j^* \in \mathcal{R}_{j,\ell}]$
    Update $\texttt{score} \mathrel{+}= \frac{n+1}{N+2}$ if $k = 1$, else $\mathrel{+}= \frac{N-n+1}{N+2}$
  end for
end for
return $P_{\text{Pokie}}(\mathcal{M}) = \frac{\texttt{score}}{L \cdot L_r}$

When $Z = (y^*, x^*, \{y\}_{i=1}^N)$ is limited, one can run Pokie with $L = 1$; however, this provides limited information about how well the posterior model is calibrated to the true posterior. In this scenario, we outline two practical strategies to improve the estimation of calibration.

To mitigate the limited number of fiducial samples, we adopt a Monte Carlo approximation of the Pokie score by fixing the fiducial sample y^*, x^* across draws L, while independently resampling posterior samples $\{y_j\}_{j=1}^N \sim p(y \mid x^*, \mathcal{M})$. This approach, aligned with Equation 5, marginalizes over posterior variability while preserving the i.i.d. assumptions required for Pokie. Note that this approach can be considerably more computationally intensive.

An alternative approach is to consider reusing y^* and $\{y_j\}_{j=1}^N$ across Monte Carlo iterations when it is too computationally intensive to resample Z. In this case, we rerun Pokie by generating new regions $\mathcal{R}$, defined by keeping c the same, drawing new y_j, and recomputing the distances $\|c - y_j\|$, while holding the posterior samples fixed. While this reuse violates independence assumptions, it offers substantial computational savings. Empirically, and similarly to PQMass, we find that this approximation can still yield useful assessments of posterior calibration. We leave the choice to the user to determine if this approach is sufficient for their use case.

D. Proofs

D.1. Pokie statistic derivation

Proposition D.1 (Pokie statistic)**.** *Let $\mathcal{R}$ be a hypersphere centered at c with radius $\|c - y_j\|$ where $y_j \sim p(y)$. Let $n \sim \mathcal{B}(N, \lambda_n)$ and $k \sim \mathcal{B}(1, \lambda_k)$ with $\lambda_n = \int p(y)\mathbb{1}(y \in \mathcal{R})dy$ and $\lambda_k = \int q(y)\mathbb{1}(y \in \mathcal{R})dy$, be two random variables. Then, under the null hypothesis "$\lambda_n = \lambda_k$, $\forall \mathcal{R}$", the conditional distribution of k given n and $\mathcal{R}$ is given by:*

$$\begin{aligned} p\big(k = 1 \mid n, \mathcal{R}\big) &= \frac{n+1}{N+2}, \\ p\big(k = 0 \mid n, \mathcal{R}\big) &= \frac{N-n+1}{N+2}. \end{aligned} \tag{9}$$

Proof. We begin by marginalizing over the shared parameter λ, where we defined $\lambda = \lambda_n = \lambda_k$ under the null hypothesis:

$$\begin{aligned} p(k \mid n, \mathcal{R}) &= \int p(k, \lambda \mid n, \mathcal{R})d\lambda, \\ &= \int p(k \mid n, \lambda, \mathcal{R})p(\lambda \mid n, \mathcal{R})d\lambda. \end{aligned} \tag{10}$$

Since k is independent of n given λ and $\mathcal{R}$ the probability becomes

$$p(k \mid n, \mathcal{R}) = \int p(k \mid \lambda, \mathcal{R})p(\lambda \mid n, \mathcal{R})d\lambda. \tag{11}$$

By applying Bayes theorem we obtain

$$\begin{aligned} &p(k \mid n, \mathcal{R}) \\ &= \frac{1}{p(n \mid \mathcal{R})} \int p(k \mid \lambda, \mathcal{R})p(n \mid \lambda, \mathcal{R})p(\lambda \mid \mathcal{R})d\lambda. \end{aligned} \tag{12}$$

By assuming an uninformative distribution $p(\lambda \mid \mathcal{R}) =$

$\mathcal{U}[0,1]$, we derive

$$p(k \mid n, \mathcal{R}) = \frac{1}{p(n \mid \mathcal{R})} \int_0^1 p(k \mid \lambda, \mathcal{R}) p(n \mid \lambda, \mathcal{R}) d\lambda. \tag{13}$$

Recalling that $p(k \mid \lambda, \mathcal{R}) = p(k \mid \lambda) = \mathcal{B}(1, \lambda)$ and $p(n \mid \lambda, \mathcal{R}) = p(n \mid \lambda) = \mathcal{B}(N, \lambda)$ we have

$$p(k \mid n, \mathcal{R}) = \frac{\binom{1}{k}\binom{N}{n}}{\int_0^1 p(n \mid \lambda, \mathcal{R}) d\lambda} \int_0^1 \lambda^{k+n}(1-\lambda)^{1-k+N-n} d\lambda. \tag{14}$$

Recognizing the Beta distribution we end up with

$$\begin{aligned} &p(k \mid n, \mathcal{R}) \\ &= \binom{1}{k+1} \frac{\beta(k+n+1, 2-k+N-n)}{\beta(n+1, N-n+1)} \\ &= \frac{\Gamma(k+n+1)\Gamma(2-k+N-n)}{\Gamma(2-k)\Gamma(k+1)\Gamma(n+1)\Gamma(N-n+1)(N+2)} \end{aligned} \tag{15}$$

Finally, we have for $k = 0$

$$p(k = 0 \mid n, \mathcal{R}) = \frac{N-n+1}{N+2}, \tag{16}$$

and for $k = 1$

$$p(k = 1 \mid n, \mathcal{R}) = \frac{n+1}{N+2}. \tag{17}$$

□

D.2. Pokie calibration convergence

Theorem D.2 (Pokie Calibration Convergence)**.** *Let $\mathcal{M}^*$ denote the true model, and let $\mathcal{M}$ be a candidate model. Suppose that for all simulation $x \sim p(x \mid \mathcal{M}^*)$, the posterior distributions agree: $p(y \mid x, \mathcal{M}) = p(y \mid x, \mathcal{M}^*)$. Then, the Pokie score of $\mathcal{M}$ satisfies*

$$\mathbb{P}_{\text{Pokie}}(\mathcal{M}) = \tfrac{2}{3}.$$

Proof. Using the law of total expectation and the fact that the expression of $p(k \mid n, \mathcal{R})$ (Equation 3) does not explicitly depend on $\mathcal{R}$, Equation 4 becomes:

$$\begin{aligned} \mathbb{P}_{Pokie}(\mathcal{M}) &= \mathbb{E}_{p(k,n,\mathcal{R})}\left[p(k|n,\mathcal{R})\right] \\ &= \mathbb{E}_{p(k,n)}\left[p(k|n,\mathcal{R})\right]. \end{aligned}$$

Given that $k|\lambda_k$ and $n|\lambda_n$ are independent we write:

$$p(k,n) = \int p(k,n,\lambda_n,\lambda_k) d\lambda_n d\lambda_k \tag{18}$$

$$= \int p(k|\lambda_k) p(n|\lambda_n) p(\lambda_n,\lambda_k) d\lambda_n d\lambda_k. \tag{19}$$

Replacing $p(k,n)$ into the expectation, we derive

$$\begin{aligned} &\mathbb{E}_{p(k,n,\mathcal{R})}\left[p(k|n,\mathcal{R})\right] \\ &= \mathbb{E}_{p(\lambda_k,\lambda_n)}\left[\mathbb{E}_{p(k|\lambda_k)p(n|\lambda_n)}\left[p(k|n,\mathcal{R}\right]\right]. \end{aligned}$$

By substituting the explicit expressions of the Bernoulli and Binomial distributions, we derive

$$\begin{aligned} &\mathbb{P}_{Pokie}(\mathcal{M}) \\ &= \mathbb{E}_{p(\lambda_k,\lambda_n)p(k|\lambda_k)p(n|\lambda_n)} \\ &\left[\frac{n+1}{N+2} \cdot \mathbb{1}(k=1) + \frac{N-n+1}{N+2} \cdot \mathbb{1}(k=0)\right] \\ &= \mathbb{E}_{p(\lambda_k,\lambda_n)} \\ &\left[\frac{N\lambda_n+1}{N+2} \cdot \lambda_k + \frac{N-N\lambda_n+1}{N+2} \cdot (1-\lambda_k)\right]. \end{aligned}$$

After simplifying this expectation, we find:

$$\begin{aligned} &\mathbb{P}_{Pokie}(\mathcal{M}) \\ &= \frac{2N\,\mathbb{E}[\lambda_n\lambda_k] - N\,\mathbb{E}[\lambda_n] - N\,\mathbb{E}[\lambda_k] + N + 1}{N+2}. \end{aligned} \tag{20}$$

Under equality of posterior distributions for all observations, we have that $\lambda_n = \lambda_k$ for all $\mathcal{R}$ and $p(\lambda_n, \lambda_k) = \delta(\lambda_n - \lambda_k)p(\lambda_n)$. Substituting into the Pokie expectation formula we get

$$\mathbb{P}_{\text{Pokie}}(\mathcal{M}) = \frac{2N\mathbb{E}_{p(\lambda_n)}[\lambda_n^2] - 2N\mathbb{E}_{p(\lambda_n)}[\lambda_n] + N + 1}{N+2}. \tag{21}$$

According to the Probability Integral Transform theorem, the continuous random variable

$$\lambda_n = \int p(y \mid x, \mathcal{M}) \mathbb{1}\left(\|y - c\| \le \|y_j - c\|\right) dy, \tag{22}$$

with $y_j \sim p(y \mid x, \mathcal{M})$, follows a uniform distribution on $[0,1]$. Hence, by using $\lambda_n \sim \mathcal{U}[0,1]$, we derive

$$\begin{aligned} \mathbb{P}_{\text{Pokie}}(\mathcal{M}) &= \frac{2N \cdot \frac{1}{3} - 2N \cdot \frac{1}{2} + N + 1}{N+2} \\ &= \frac{2N+3}{3(N+2)} \to \frac{2}{3} \quad \text{as } N \to \infty. \end{aligned}$$

We note that this result is a necessary condition. Determining whether this condition is also sufficient is left as future work.

□

D.3. Pokie score lower-bound

Proposition D.3 (Pokie score lower-bound)**.** *Let $\mathcal{M}^*$ be the true model and $\mathcal{M}$ a candidate model. Suppose that for every $x \sim p(x|\mathcal{M}^*)$, $y_{\mathcal{M}} \sim p(y|x,\mathcal{M})$ and $y_{\mathcal{M}^*} \sim p(y|x,\mathcal{M}^*)$ satisfies $y_{\mathcal{M}} \perp\!\!\!\perp y_{\mathcal{M}^*}$, then*

$$\mathbb{P}_{\text{Pokie}}(\mathcal{M}) = \tfrac{1}{2}.$$

Proof. Recall the Pokie score from Eq. (20):

$$\mathbb{P}_{\text{Pokie}}(\mathcal{M}) = \frac{2N\,\mathbb{E}[\lambda_n\lambda_k] - N\,\mathbb{E}[\lambda_n] - N\,\mathbb{E}[\lambda_k] + N + 1}{N+2}.$$

Under independence of posterior distributions, we have that $\lambda_n|\mathcal{R} \perp\!\!\!\perp \lambda_k|\mathcal{R}$. Additionally, by choosing an uninformative uniform distribution on $[0,1]$ for $p(\lambda_n|\mathcal{R})$ (as used in the derivation of the Pokie statistic), and recalling that $p(\lambda_n)$ is uniform on $[0,1]$, we can simplify the expression of Equation 20 as follow

$$\begin{aligned}
&\mathbb{P}_{\text{Pokie}}(\mathcal{M}) \\
&= \frac{2N\,\mathbb{E}_{p(\mathcal{R})}\left[\mathbb{E}_{p(\lambda_n,\lambda_k|\mathcal{R})}\left[\lambda_n\lambda_k\right]\right] - N\,\mathbb{E}[\lambda_k] + \frac{N}{2} + 1}{N+2} \\
&= \frac{2N\,\mathbb{E}_{p(\mathcal{R})}\left[\frac{1}{2}\mathbb{E}_{p(\lambda_k|\mathcal{R})}\left[\lambda_k\right]\right] - N\,\mathbb{E}[\lambda_k] + \frac{N}{2} + 1}{N+2} \\
&= \frac{N\,\mathbb{E}\left[\lambda_k\right] - N\,\mathbb{E}[\lambda_k] + \frac{N}{2} + 1}{N+2} \\
&= \frac{\frac{N}{2}+1}{N+2} \\
&= \frac{1}{2}.
\end{aligned}$$

□

E. Model Misspecification Images

Here we provide additional details about the data setup in 3.3. The true distribution, $p(y|x^*, M)$, is defined to be an EPL with 3 Sérsic sources. We then define our observed images as y^* with Gaussian noise is added with $\eta \sim \mathcal{N}(0, 1^2)$, yielding I_{obs}. All generated lenses are rendered on a 100×100 grid with pixel scale 0.05.

Table 5 lists all parameters inferred during posterior estimation. In setups with three Sérsic sources, we independently sample source-specific parameters, x_0, y_0, and I_e, for each component. MALA inference is run using 100 walkers, with 200 steps for burn-in and 200 steps for sampling, yielding 20,000 posterior samples per model. When running Pokie, all models are evaluated in the 13-dimensional parameter space. For single-source configurations, the second and third source positions and intensities (x_0, y_0, I_e) are fixed to zero, ensuring a consistent parameter structure across models and allowing for fair posterior comparisons.

All posterior inferences were run using MALA sampling on a single AMD Milan CPU core for approximately 8 minutes (wall-time) per configuration, using up to 10 GB of memory. Across 100 synthetic observations, the total inference cost was approximately 13.33 CPU-hours.

Figure 3 shows one example of the clean ground truth image (EPL + 3 Sérsic sources), the corresponding noisy observation ($\sigma = 1$), and posterior means from each candidate model: EPL+3, SIE+3, EPL+1, and SIE+1. Each posterior mean is computed by averaging 100 MALA samples. The final column shows residuals between the observation and posterior mean. Only the correctly specified model (EPL+3, top row) produces residuals consistent with Gaussian noise. All others exhibit structured residuals, revealing mismatches due to incorrect lens profile, source count, or both.

Table 5. Parameter ranges for lens and source parameters that are inferred. All other parameters are held constant across models. SIE models implicitly fix $\gamma = 2.0$.

Parameter	Distribution
EPL Lens	
Einstein radius b	$\mathcal{U}[1.0, 1.5]$
Axis ratio q	$\mathcal{U}[0.5, 0.9]$
Orientation angle ϕ	$\mathcal{U}[0.0, \pi]$
Power-law slope γ	$\mathcal{U}[1.75, 2.25]$
SIE Lens	
Einstein radius b	$\mathcal{U}[1.0, 1.5]$
Axis ratio q	$\mathcal{U}[0.5, 0.9]$
Orientation angle ϕ	$\mathcal{U}[0.0, \pi]$
Sérsic Source	
Source center $\hat{x}_{\text{src}}$	$\mathcal{U}[-0.5, 0.5]$
Source center $\hat{y}_{\text{src}}$	$\mathcal{U}[0.05, 0.10]$
Effective intensity I_e	$\mathcal{U}[0.4, 0.8]$

F. Lensed galaxy images

In Figure 4, we showcase some ground truths $x^\star$ and posterior samples $x \sim p_i(x \mid y)$ for the 4 different posterior sampling configurations explained in 3.4.

Here, we describe the model and training hyperparameters of the SBM priors, $p_s(x)$ and $p_e(x)$, taken from (Barco et al., 2025) for reproducibility. We also refer the reader to the corresponding work for details of the training datasets. The models were created using the `score-models`[2] package, and follow a NCSN++ architecture (Song et al., 2021). The model hyperparameters, within the `score-models` package, are:

```
"beta_min": 0.01,
"beta_max": 20.0,
"channels": 3,
"nf": 64,
"ch_mult": [1, 2, 2, 2],
"sde": "vp"
```

The SBMs were trained with an Adam optimizer (Kingma

[2] github.com/AlexandreAdam/score_models

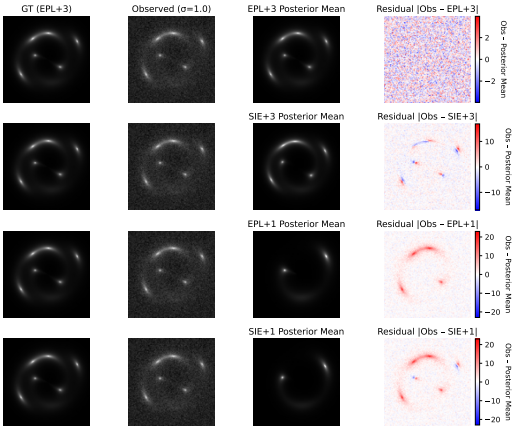

Figure 3. Left to right: clean ground truth image (EPL + 3 Sérsic sources), observed data with Gaussian noise ($\sigma = 1$), posterior means from four candidate models, and corresponding residuals (observation minus posterior mean). Only the correctly specified model (top row: EPL + 3 Sérsic source) produces residuals consistent with Gaussian noise. Other models show structured residuals, revealing mismatches due to incorrect lens type and/or source count.

& Ba, 2015), with $lr = 1e^{-4}$, batch size of 256, and `ema_decay` $= 0.999$, for approximately 2.5×10^5 optimization steps. All hyperparameters not specified here were left to the `score-models` default values. Each SBM was trained on an A100 GPU for 20 hours (wall-time) and with 32Gb of VRAM allocated.

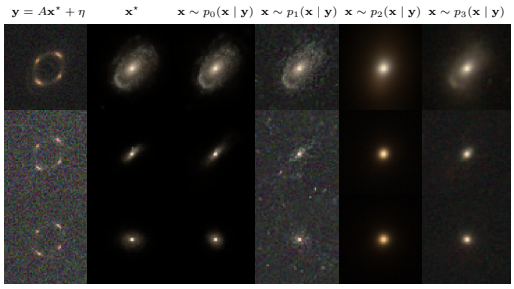

Figure 4. Plotted in order of left to right is the result of the forward model noised up. The 2nd column is the ground truth, next is the first posterior model with elliptical galaxy prior and $\sigma_n = 2.0$, the next column is the posterior given elliptical galaxy prior and $\sigma_n = 2$, the 5th column is the posterior model with a spiral galaxy prior and $\sigma_n = 0.5$, and lately the last column is the posterior model given a spiral galaxy prior and $\sigma_n = 0.5$.

Finally, we use the same SDE solver setup for prior and posterior sampling as (Barco et al., 2025), which is a predictor-corrector solver (Song et al., 2021) with 1024 solver steps. We obtain 16 prior samples to simulate the ground truths $x^\star$, and get 64 posterior samples per observation per configuration. Inference of these 4 112 samples was carried out in a single A100 GPU for 4 hours (wall-time) and 40Gb of VRAM allocated.

G. Bayes Factor Comparison

In this appendix, we compare Pokie with the Bayes Factor across two settings: the linear regression experiment from Section 3.1 and the distributional shift experiment from Section 3.2. In both cases, the goal is to compare Pokie and the Bayes Factor in identifying well-calibrated posteriors and assessing models that are poorly calibrated.

G.1. Linear Regression

We compute Bayes Factors for the linear regression setup described in Section 3.1, where models are perturbed by varying the posterior mean ($\eta = \{0.001, 0.01, 0.1, 0.15, 0.2, 0.25\}$) while keeping the covariance fixed. We have 5 000 fiducial samples, and from each model we draw 5 000 samples. We compute Bayes Factors using:

$$\mathrm{BF}(\eta) = \frac{p(y \mid \mathcal{M}_\eta)}{p(y \mid \mathcal{M}^*)},$$

where $\mathcal{M}_\eta$ is the model with noise η, and $\mathcal{M}^*$ denotes the model with the least noise ($\eta = 0.001$), which we treat as the ground truth.

Table 6. Comparison of Pokie and Bayes Factor scores in linear regression. Pokie and BF both rank models consistently with increasing levels of misspecification. Pokie scores range from $1/2$ (poorly specified model) to $2/3$ (well specified model). Bayes Factors near 1 indicate models that are nearly as plausible as the reference model $\mathcal{M}^*$; lower values indicate less support.

Noise Level	Pokie Score	BF
0.001	**0.6646**	**0.999**
0.01	0.6412	0.990
0.1	0.5656	0.906
0.15	0.5573	0.863
0.2	0.5525	0.821
0.25	0.5493	0.782

As shown in Table 6, both Bayes Factor and Pokie consistently assign higher scores to the models with less noise, with the ranking degrading smoothly as model misspecification increases. This demonstrates that both Pokie and Bayes Factor can identify, in this experiment, calibrated and poorly calibrated models.

G.2. Gaussian Mixture Model Shifts

We compute Bayes Factor scores for the GMM shift experiment described in Section 3.2. As this experiment does not involve Bayesian inference, there is no likelihood or prior. Instead, we consider the GMM probability density function (PDF) as our evidence to compute a Bayes Factor score. Specifically, we evaluate the probability of 5 000 samples

from the true (unshifted) GMM, under each shifted model's PDF. This allows us to compute a density ratio between the shifted and unshifted GMMs, serving as a proxy for the Bayes Factor in this likelihood-free setting:

$$\mathrm{BF}(\ell) = \frac{p(y \mid \mathcal{M}_\ell)}{p(y \mid \mathcal{M})},$$

where $\mathcal{M}_\ell$ is the GMM with shift magnitude ℓ, and $\mathcal{M}^*$ is the true GMM with no shift ($\ell = 0$).

Table 7. Comparison of Pokie and Bayes Factor scores across GMM shift magnitudes. Pokie reliably favors the in-distribution model and penalizes shifted ones. Pokie scores range from $1/2$ (poorly specified model) to $2/3$ (well specified model). Bayes Factors near 1 indicate models that are nearly as plausible as the reference model $\mathcal{M}^*$; lower values indicate less support.

Shift Magnitude	Pokie Score	Bayes Factor
-6	0.5048	0.000
-3	0.5865	4.55×10^{-145}
0	**0.6661**	**1.000**
3	0.5959	5.45×10^{-162}
6	0.5045	0.000

As shown in Table 7, both Pokie and the Bayes Factor correctly identify the unshifted model ($\ell = 0$) as the most accurate and identify increasingly shifted models as less probable. The Pokie score transitions from the theoretical maximum of $2/3$ for the well-calibrated model toward the lower bound of $1/2$ as the shift increases, while the log Bayes Factor becomes increasingly smaller. Here we note that both Pokie and Bayes Factor are sensitive to the fact that the shifted models are poorly calibrated.

Across both experiments, we find strong agreement between Pokie and Bayes Factor rankings. This demonstrates that Pokie can serve as a viable, sample-based metric for model assessment and comparison, especially valuable in scenarios where likelihood evaluation or evidence computation is intractable or unavailable. While Bayes Factors operate through the marginal likelihood in data space, Pokie evaluates posterior alignment directly in parameter space, making it applicable in a broader range of settings.

H. Sensitivity Analysis

We conduct a sensitivity analysis of the Pokie score to evaluate how it responds to variations in key experimental parameters: (i) the dimensionality of the parameter space, (ii) the number of hyperspheres per fiducial used for estimation, (iii) the number of posterior samples per model, and (iv) the number of distinct ground-truth posteriors. These experiments characterize the robustness of Pokie under practical constraints.

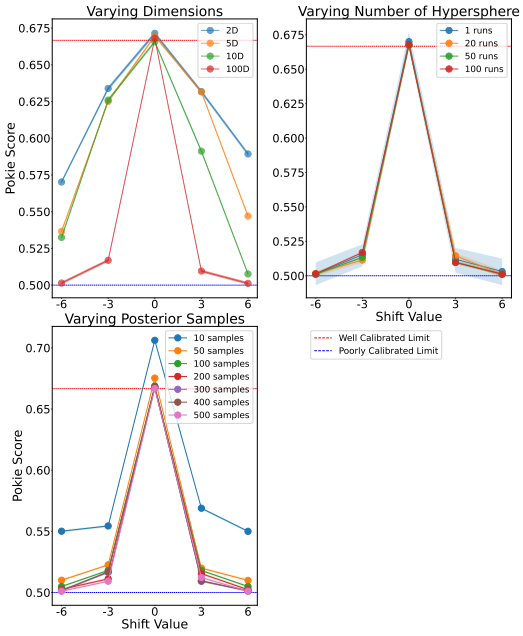

Figure 5. Pokie score sensitivity under varying experimental conditions. Top-left: effect of dimension on score; Top-right: number of hyperspheres per fiducial; Bottom-left: number of posterior samples. Across settings, Pokie scores peak for the well-calibrated ($\ell = 0$) model and fall to the poorly calibrated limit with increasing shift.

To assess sensitivity to posterior characteristics, we use the experiment introduced in Section 3.2, and vary the following: dimensionality, number of hyperspheres per fiducial, and the number of posterior samples. Results are summarized in Figure 5.

First, we vary the dimensionality of the problem, evaluating GMMs in 2, 5, 10, and 100 dimensions while fixing the number of posterior samples to 400 samples and using 100 hyperspheres per fiducial. We find that Pokie scores remain well-behaved across all tested dimensions: well-calibrated posteriors score near the theoretical maximum of $2/3$, while miscalibrated posteriors approach the lower bound of $1/2$. Next, we fix the problem to 100D and 400 posterior samples and vary the number of hyperspheres per fiducial from 1 to 100. The Pokie score stabilizes rapidly after approximately 10 runs. Finally, we fix the problem to 100D with 100 hyperspheres per fiducial and vary the number of posterior samples per model between 10 and 500. We see that with only 10 samples, the model correctly identifies the best and worst models; however, the Pokie scores are shifted upwards due to the lack of samples. The 1/2 to 2/3 Pokie score bounds are defined in the limit as the number of posterior samples approaches infinity, so with only 10 samples, these bounds no longer constrain the score, even if the model rankings remain correct. Inversely, the results

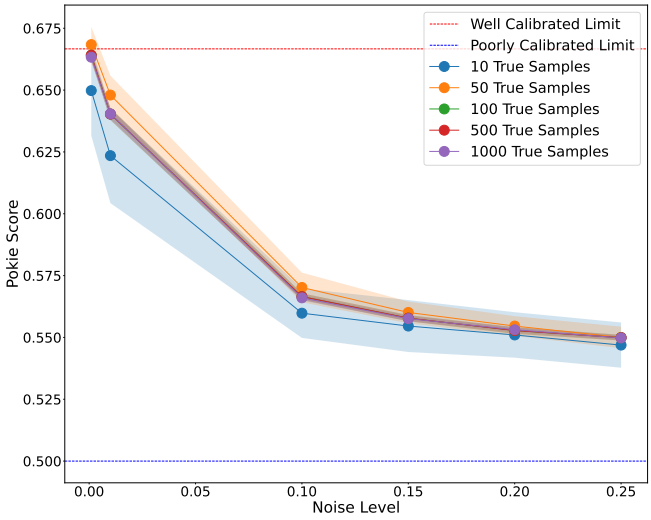

Figure 6. Pokie score vs noise level for varying counts of true posteriors. As the number of true posteriors increases, confidence in the Pokie estimate improves, and the distinction between well- and poorly-calibrated models becomes more pronounced.

indicate that even 100 posterior samples are sufficient to produce consistent scores.

We further investigate how the number of true posterior distributions affects Pokie. Using the linear regression setup from Section 3.1, we fix the number of posterior samples per model to 5 000, and perform 100 hyperspheres per fiducial. We vary the number of distinct ground-truth draws (θ^*) from 10 to 1 000, and evaluate Pokie scores at different noise levels. Results are shown in Figure 6.

As the number of true posteriors increases, the Pokie becomes sharper and confidence intervals narrow. Even with relatively few posteriors, Pokie can still correctly assess the quality of the posterior and rank models.

Overall, Pokie remains robust given the sensitivity test. It produces accurate scores with relatively small sample sizes and maintains consistency across high-dimensional spaces. The number of samples, whether fiducials or posterior samples, has the largest influence on stability, though even modest values yield usable estimates. These properties make Pokie well-suited for practical use in SBI tasks under practical constraints.