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]

[1]Department of Physics, University of Montreal, Montreal, Canada
[2]MILA Quebec AI Institute, Montreal, Canada
[3]CIELA Institute, 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>

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 $1/2$ in the worst case. 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 [3, 7, 15, 27, 28] and conducting Bayesian model comparison [9, 18, 22, 29]. 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: $p(y \mid x, \mathcal{M}) \propto p(x \mid y, \mathcal{M})p(y \mid \mathcal{M})$.

Evaluating whether a posterior estimator is calibrated is crucial, particularly with the rise of implicit inference methods powered by deep learning [e.g. 4, 16, 17]. 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. [14], 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. The classical Bayesian approach relies on the computation of the model evidence $p(x \mid \mathcal{M}) = \int p(x \mid y, \mathcal{M})p(y \mid \mathcal{M})dy$ [10], which is often computationally intractable, particularly in high-dimensional parameter space or simulation-based settings [1, 25].

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 [12] and PQMass [13], Pokie quantifies the expected probability that the samples of

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Frontiers in Probabilistic Inference: Sampling Meets Learning.

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 [13] 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 $n \sim \mathrm{B}(N, \lambda_n)$, and $k \sim \mathrm{B}(1, \lambda_k)$, 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

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

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 [12], 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: $p(k \mid n, \mathcal{R})$. 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 in Appendix D.1):

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

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

$$\mathbb{P}_{\mathrm{Pokie}}(\mathcal{M}) = \mathbb{E}_{p(Z)}\left[\mathbb{E}_{p(k,n,\mathcal{R}|Z)}\left[p(k \mid n, \mathcal{R})\right]\right] \tag{3}$$

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

$$\mathbb{P}_{\mathrm{Pokie}}(\mathcal{M}) \approx \frac{1}{L} \sum_{l=1}^{L} p(k_l \mid n_l, \mathcal{R}_l), \tag{4}$$

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.

3 Experiments

For all experiments, centers c are sampled from $\mathcal{U}(0, 1)$, parameters are normalized to $[0, 1]$, and we use 100 hyperspheres per fiducial. Experiments are run on an M2 MacBook Air with 8 GB of RAM.

3.1 Linear regression

We consider a linear regression task where we infer the posterior distribution over weights $\theta = [m, b]$ of the linear model $y = mx + b + \eta$ with $\eta \sim \mathcal{N}(0, \sigma^2)$, and $\theta \sim \mathcal{N}(\mu_0, \Sigma_0)$. Because both the likelihood and prior distributions are Gaussian, we can derive an analytical form of the posterior distribution $p(\theta \mid y) = \mathcal{N}(\theta \mid \mu_{\text{post}}, \Sigma_{\text{post}})$, 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, η = $\{0.001, 0.01, 0.01, 0.10, 0.20, 0.25\}$, and choose the model with the least noise as our true model. To simulate miscalibration, we apply a fixed directional bias of magnitude $\sqrt{n}$ to the posterior mean vector, shifting it equally along both parameter dimensions. 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. The results, shown in Table 1, demonstrate that Pokie correctly identifies the least noisy posterior as the most accurate and ranks posteriors with higher noise as progressively 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 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 2 to iteratively learn the prior distribution and get posterior distributions.

We use the same lensing forward model as 2, 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 G. We then run Pokie to evaluate the 4 models. The results in Table 2, 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.

3.3 Conditional distributions

Beyond its use in Bayesian inference, Pokie is also effective for evaluating the similarity of general conditional distributions. We demonstrate this by comparing how well a Conditional Variational

Table 2: 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** $\sigma_{\boldsymbol{\eta}} = 2$	$\mathbf{0.6518 \pm 0.0369}$
$p_s(x)$ 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

Autoencoder (CVAE)[8, 23] and a conditional diffusion model [6] (see Appendix H approximate the true conditional distribution $p(\text{image} \mid \text{label})$ of MNIST digits[1].

For each model, we generate $1\,000$ samples per digit class conditioned on the respective labels (see Appendix I) and compare against 100 randomly selected MNIST test samples per class. Table 3 shows the results. The CVAE achieves scores ranging from 0.543 to 0.592, indicating only partial learning of the conditional distributions, with some digits proving especially challenging. The diffusion model performs substantially better, with scores consistently between 0.652 and 0.661, approaching the theoretical maximum of $\frac{2}{3}$ for well-specified models. We repeat this analysis using TARP (see Appendix J) and observe agreement between the two metrics. This experiment demonstrates that Pokie not only distinguishes between models of differing quality in conditional generation but also surfaces class-level weaknesses. This experiment illustrates that Pokie applies naturally beyond Bayesian posteriors: it can be used to evaluate any learned conditional distribution.

Table 3: Pokie scores for conditional generative models on MNIST.

The diffusion model consistently achieves scores near the theoretical optimum of $\frac{2}{3}$, indicating it successfully learns the true conditional distribution $p(\text{image} \mid \text{label})$, while the CVAE shows substantially poorer performance.

MNIST Digit	Conditional VAE	Conditional Diffusion
0	0.5683 ± 0.0058	0.6579 ± 0.0039
1	0.5922 ± 0.0039	0.6524 ± 0.0044
2	0.5437 ± 0.0050	0.6613 ± 0.0041
3	0.5682 ± 0.0066	0.6564 ± 0.0038
4	0.5547 ± 0.0045	0.6561 ± 0.0037
5	0.5557 ± 0.0045	0.6590 ± 0.0042
6	0.5703 ± 0.0056	0.6593 ± 0.0037
7	0.5827 ± 0.0059	0.6566 ± 0.0039
8	0.5503 ± 0.0058	0.6589 ± 0.0036
9	0.5667 ± 0.0046	0.6585 ± 0.0043

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 K 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 L 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.). 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.

[1] http://yann.lecun.com/exdb/mnist/

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. G.M.B. acknowledges support from the Fonds de recherche du Québec – Nature et technologies (FRQNT) under a Doctoral Research Scholarship (doi:10.69777/368273).

References

[1] Justin Alsing, Benjamin Wandelt, and Stephen Feeney. Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology. , 477(3):2874–2885, July 2018.

[2] Gabriel Missael Barco, Alexandre Adam, Connor Stone, Yashar Hezaveh, and Laurence Perreault-Levasseur. Tackling the problem of distributional shifts: Correcting misspecified, high-dimensional data-driven priors for inverse problems. *The Astrophysical Journal*, 980(1):108, February 2025.

[3] James Carzon, Bruno Abreu, Leighton Regayre, Kenneth Carslaw, Lucia Deaconu, Philip Stier, Hamish Gordon, and Mikael Kuusela. Statistical constraints on climate model parameters using a scalable cloud-based inference framework. *Environmental Data Science*, 2:e24, 2023.

[4] Kyle Cranmer, Juan Pavez, and Gilles Louppe. Approximating likelihood ratios with calibrated discriminative classifiers, 2016.

[5] Andreas Filipp, Yashar Hezaveh, and Laurence Perreault-Levasseur. Robustness of Neural Ratio and Posterior Estimators to Distributional Shifts for Population-Level Dark Matter Analysis in Strong Gravitational Lensing. 11 2024.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Neural Information Processing Systems (NeurIPS)*, 2020.

[7] Michael F. Howland, Oliver R. A. Dunbar, and Tapio Schneider. Parameter uncertainty quantification in an idealized gcm with a seasonal cycle. *Journal of Advances in Modeling Earth Systems*, 14(3):e2021MS002735, 2022. e2021MS002735 2021MS002735.

[8] Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning, 2019.

[9] Niall Jeffrey and Benjamin D Wandelt. Evidence networks: simple losses for fast, amortized, neural bayesian model comparison. *Machine Learning: Science and Technology*, 5(1):015008, January 2024.

[10] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[12] Pablo Lemos, Adam Coogan, Yashar Hezaveh, and Laurence Perreault-Levasseur. Sampling-based accuracy testing of posterior estimators for general inference, 2023.

[13] Pablo Lemos, Sammy Sharief, Nikolay Malkin, Salma Salhi, Connor Stone, Laurence Perreault-Levasseur, and Yashar Hezaveh. Pqmass: Probabilistic assessment of the quality of generative models using probability mass estimation, 2025.

[14] Jan-Matthis Lueckmann, Jan Boelts, David S. Greenberg, Pedro J. Gonçalves, and Jakob H. Macke. Benchmarking simulation-based inference, 2021.

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

[16] George Papamakarios and Iain Murray. Fast ϵ-free inference of simulation models with bayesian conditional density estimation, 2018.

[17] George Papamakarios, David C. Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows, 2019.

[18] Juho Piironen and Aki Vehtari. Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, April 2016.

[19] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341 – 363, 1996.

[20] François Rozet, Gérôme Andry, François Lanusse, and Gilles Louppe. Learning diffusion priors from observations by expectation maximization. *Advances in Neural Information Processing Systems*, 37:87647–87682, 2024.

[21] J. L. Sérsic. 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.

[22] A. Slosar, P. Carreira, K. Cleary, R. D. Davies, R. J. Davis, C. Dickinson, R. Genova-Santos, K. Grainge, C. M. Gutierrez, Y. A. Hafez, M. P. Hobson, M. E. Jones, R. Kneissl, K. Lancaster, A. Lasenby, J. P. Leahy, K. Maisinger, P. J. Marshall, G. G. Pooley, R. Rebolo, J. A. Rubino-Martin, B. Rusholme, R. D. E. Saunders, R. Savage, P. F. Scott, P. J. Sosa Molina, A. C. Taylor, D. Titterington, E. Waldram, R. A. Watson, and A. Wilkinson. 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.

[23] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[24] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 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.

[25] A Spurio Mancini, M M Docherty, M A Price, and J D McEwen. Bayesian model comparison for simulation-based inference. *RAS Techniques and Instruments*, 2(1):710–722, January 2023.

[26] Connor Stone, Alexandre Adam, Adam Coogan, M. J. Yantovski-Barth, Andreas Filipp, Landung Setiawan, Cordero Core, Ronan Legin, Charles Wilson, Gabriel Missael Barco, Yashar Hezaveh, and Laurence Perreault-Levasseur. Caustics: A python package for accelerated strong gravitational lensing simulations, 2024.

[27] Max Tegmark, Michael A. Strauss, Michael R. Blanton, Kevork Abazajian, Scott Dodelson, Havard Sandvik, Xiaomin Wang, David H. Weinberg, Idit Zehavi, Neta A. Bahcall, Fiona Hoyle, David Schlegel, Roman Scoccimarro, Michael S. Vogeley, Andreas Berlind, Tamás Budavari, Andrew Connolly, Daniel J. Eisenstein, Douglas Finkbeiner, Joshua A. Frieman, James E. Gunn, Lam Hui, Bhuvnesh Jain, David Johnston, Stephen Kent, Huan Lin, Reiko Nakajima, Robert C. Nichol, Jeremiah P. Ostriker, Adrian Pope, Ryan Scranton, Uroš Seljak, Ravi K. Sheth, Albert Stebbins, Alexander S. Szalay, István Szapudi, Yongzhong Xu, James Annis, J. Brinkmann, Scott Burles, Francisco J. Castander, Istvan Csabai, Jon Loveday, Mamoru Doi, Masataka Fukugita, Bruce Gillespie, Greg Hennessy, David W. Hogg, Željko Ivezić, Gillian R. Knapp, Don Q. Lamb, Brian C. Lee, Robert H. Lupton, Timothy A. McKay, Peter Kunszt, Jeffrey A. Munn, Liam O'Connell, John Peoples, Jeffrey R. Pier, Michael Richmond, Constance Rockosi, Donald P. Schneider, Christopher Stoughton, Douglas L. Tucker, Daniel E. vanden Berk, Brian Yanny, and Donald G. York. Cosmological parameters from SDSS and WMAP. , 69(10):103501, May 2004.

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

[29] Ka-Veng Yuen. Recent developments of bayesian model class selection and applications in civil engineering. *Structural Safety*, 32(5):338–346, 2010. Probabilistic Methods for Modeling, Simulation and Optimization of Engineering Structures under Uncertainty in honor of Jim Beck's 60th Birthday.

A TARP

TARP [12] 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 [13] 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 $\mathcal{R} \subseteq \Omega$:

$$\mathbb{P}_p(\mathcal{R}) = \mathbb{P}_q(\mathcal{R}) \quad \forall \mathcal{R} \subseteq \Omega. \tag{5}$$

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

$$\mathbb{P}_p(\mathcal{R}) = \mathbb{E}_{y \sim p(y)}[\mathbb{1}(y \in \mathcal{R})] \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(y_i \in \mathcal{R}), \tag{6}$$

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:

$$n \sim \mathcal{B}(N, \lambda), \quad \lambda = \mathbb{P}_p(\mathcal{R}). \tag{7}$$

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 $c_{j,\ell} \sim \pi_c$
        $r_{j,\ell} \leftarrow d(c_{j,\ell}, y_{i,j})$
        $\mathcal{R}_{j,\ell} \leftarrow \{y : d(y, c_{j,\ell}) \leq r_{j,\ell}\}$
        $n \leftarrow \sum_i \mathbf{1}[y_{i,j} \in \mathcal{R}_{j,\ell}]$
        $k \leftarrow \mathbf{1}[y_j^* \in \mathcal{R}_{j,\ell}]$
        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 4, 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 region 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:*

$$p\big(k = 1 \mid n, \mathcal{R}\big) = \frac{n+1}{N+2},$$
$$p\big(k = 0 \mid n, \mathcal{R}\big) = \frac{N-n+1}{N+2}. \tag{8}$$

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

$$\begin{aligned} p(k \mid n, \mathcal{R}) &= \int p(k, \lambda \mid n, \mathcal{R}) d\lambda, \\ &= \int p(k \mid n, \lambda, \mathcal{R}) p(\lambda \mid n, \mathcal{R}) d\lambda. \end{aligned} \tag{9}$$

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

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

By applying Bayes theorem, we obtain

$$\begin{aligned} &p(k \mid n, \mathcal{R}) \\ &= \frac{1}{p(n \mid \mathcal{R})} \int p(k \mid \lambda, \mathcal{R}) p(n \mid \lambda, \mathcal{R}) p(\lambda \mid \mathcal{R}) d\lambda. \end{aligned} \tag{11}$$

By assuming an uninformative distribution $p(\lambda \mid \mathcal{R}) = \mathcal{U}[0, 1]$, we derive

$$p(k \mid n, \mathcal{R}) = \frac{1}{p(n \mid \mathcal{R})} \int_0^1 p(k \mid \lambda, \mathcal{R}) p(n \mid \lambda, \mathcal{R}) d\lambda. \tag{12}$$

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

$$\begin{aligned} &p(k \mid n, \mathcal{R}) = \\ &\frac{\binom{1}{k}\binom{N}{n}}{\int_0^1 p(n \mid \lambda, \mathcal{R}) d\lambda} \int_0^1 \lambda^{k+n} (1-\lambda)^{1-k+N-n} d\lambda. \end{aligned} \tag{13}$$

Recognizing the Beta distribution we end up with

$$\begin{aligned} &p(k \mid n, \mathcal{R}) \\ &= \binom{1}{k+1} \frac{\beta(k+n+1, 2-k+N-n)}{\beta(n+1, N-n+1)} \\ &= \frac{\Gamma(k+n+1)\Gamma(2-k+N-n)}{\Gamma(2-k)\Gamma(k+1)\Gamma(n+1)\Gamma(N-n+1)(N+2)} \end{aligned} \tag{14}$$

Finally, we have for $k = 0$

$$p(k = 0 \mid n, \mathcal{R}) = \frac{N-n+1}{N+2}, \tag{15}$$

and for $k = 1$

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

□

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 3 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:

$$p(k,n) = \int p(k,n,\lambda_n,\lambda_k)d\lambda_n d\lambda_k \tag{17}$$

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

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

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

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

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

After simplifying this expectation, we find:

$$\mathbb{P}_{Pokie}(\mathcal{M}) = \frac{2N\,\mathbb{E}[\lambda_n\lambda_k] - N\,\mathbb{E}[\lambda_n] - N\,\mathbb{E}[\lambda_k] + N + 1}{N+2}. \tag{19}$$

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

$$\mathbb{P}_{\text{Pokie}}(\mathcal{M}) = \frac{2N\mathbb{E}_{p(\lambda_n)}[\lambda_n^2] - 2N\mathbb{E}_{p(\lambda_n)}[\lambda_n] + N + 1}{N+2}. \tag{20}$$

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

$$\lambda_n = \int p(y \mid x,\mathcal{M})\mathbb{1}\left(\|y-c\| \le \|y_j - c\|\right)dy, \tag{21}$$

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

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

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. (19):

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

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 19 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 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. 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 4 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 4: 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	$\mathbf{0.6669 \pm 0.0003}$
+3	0.5156 ± 0.0003
+6	0.5000 ± 0.0004

F Model misspecification images

We apply Pokie to the joint inference of lens and source parameters in strong gravitational lensing to test whether it can detect model misspecification in the lens mass profile or in the number of background sources.

Following [5], we generate lenses with multiple Sérsic components [21] 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` [2][26].

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 1). Each models share the same prior distribution over parameters, wee apply the same Gaussian noise $\eta \sim \mathcal{N}(0, 1^2)$, and all generated lenses are rendered on a 100×100 grid with pixel scale 0.05.

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	$\mathcal{U}[0.4, 0.8]$

Table 5 lists all parameters inferred during posterior estimation. In setups with three Sérsic sources, we independently sample the second and third source positions and intensities (x_0, y_0, I_e), for each component. We use Metropolis-adjusted Langevin sampling (MALA, 19) to get our 100 posterior distributions, each with 20 000 samples in a 13-dimensional parameter space. Specifically, we use 100 walkers, with 200 steps for burn-in and 200 steps for sampling, yielding 20,000 posterior samples per model. To get each posterior, we use 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.

After sampling, we apply Pokie to evaluate how well the posterior samples match the fiducial parameters of the observed image and perform model ranking. 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. Results are shown in Table 6. 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.

Table 6: 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	$\mathbf{0.6297 \pm 0.0054}$
SIE + 3 Sersic Sources	0.5777 ± 0.0027
EPL + 1 Sersic Sources	0.5277 ± 0.0028
SIE + 1 Sersic Sources	0.5267 ± 0.0031

Figure 1 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.

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

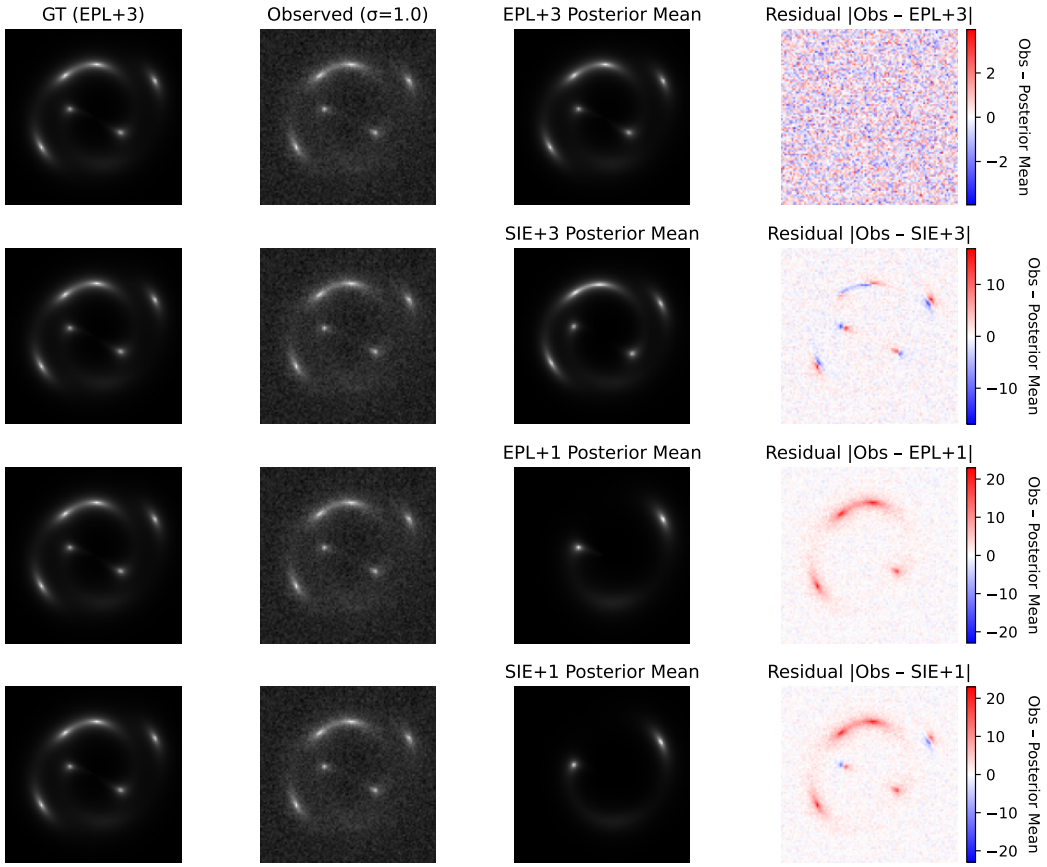

Figure 1: Left to right: clean ground truth image (EPL + 3 Sérsic sources), observed data with Gaussian noise ($\sigma = 1$), posterior means from four candidate models, and corresponding residuals (observation minus posterior mean). Only the correctly specified model (top row: EPL + 3 Sérsic source)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

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.

G Lensed galaxy images

In Figure 2, 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 3.2.

Here, we describe the model and training hyperparameters of the SBM priors, $p_s(x)$ and $p_e(x)$, taken from [2] 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`[3] package, and follow a NCSN++ architecture [24]. 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"
```

[3] github.com/AlexandreAdam/score_models

The SBMs were trained with an Adam optimizer [11], 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.

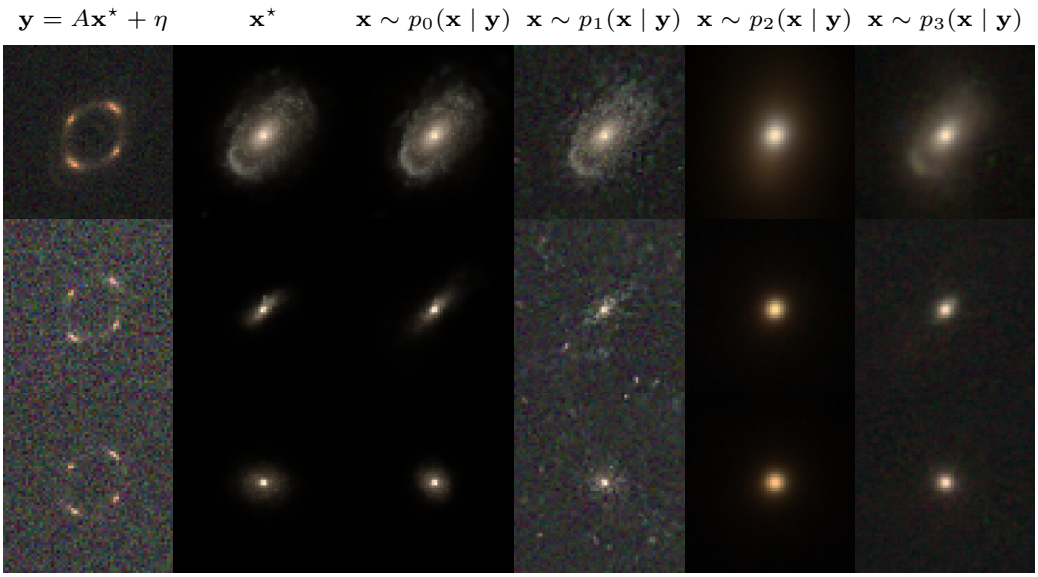

Figure 2: Plotted in order of left to right is the result of the forward model noised up. The 2nd column is the ground truth, next is the first posterior model with elliptical galaxy prior and $\sigma_n = 2.0$, the next column is the posterior given elliptical galaxy prior and $\sigma_n = 2$, the 5th 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 [2], which is a predictor-corrector solver [24] 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.

H Generative model hyperparameters

For the conditional distribution experiments in Section 3.3, the diffusion model was implemented using the `score-models` package. It utilizes an NCSN++ architecture [24] with a variance exploding noising process. The key hyperparameters within the `score-models` package were:

```
"\sigma_{\min}": 0.1,
"\sigma_{\max}": 10.0,
"channels": 1,
"nf": 64,
"ch_mult": [1, 1],
```

The model was trained with the Adam optimizer [11] ($lr = 1 \times 10^{-4}$, batch size 256, `ema_decay` $= 0.999$). All other hyperparameters followed the `score-models` defaults. Training was performed on an A100 GPU for $\sim$2 hours (wall time) using 32 GB of VRAM.

The conditional VAE used a convolutional encoder with layer structure $(28 \times 28) \to (32 \times 14 \times 14) \to (64 \times 7 \times 7) \to 32$, and a symmetric decoder with transposed convolutions. Class conditioning was introduced through learned label embeddings concatenated to both encoder and decoder inputs. The model was trained for 500 epochs with batch size 512 using Adam ($lr = 3 \times 10^{-4}$, weight decay 10^{-5}) and a cosine annealing learning rate schedule, with KL warmup.

I Conditional model samples

In Section 3.3, we evaluated conditional generative models on MNIST using Pokie. To complement the quantitative results in Table 3, we show representative samples generated from the conditional VAE and the conditional diffusion model.

Figure 3 illustrates samples from the conditional VAE. While the model captures the overall digit structure, several classes remain poorly learned, producing blurry or distorted digits. In contrast, Figure 4 shows samples from the conditional diffusion model, which produces consistently sharp and class-accurate digits across all labels. These qualitative differences mirror the Pokie scores: the cVAE achieves substantially lower values, while the diffusion model approaches $\frac{2}{3}$, confirming its stronger ability to approximate the true conditional distribution.

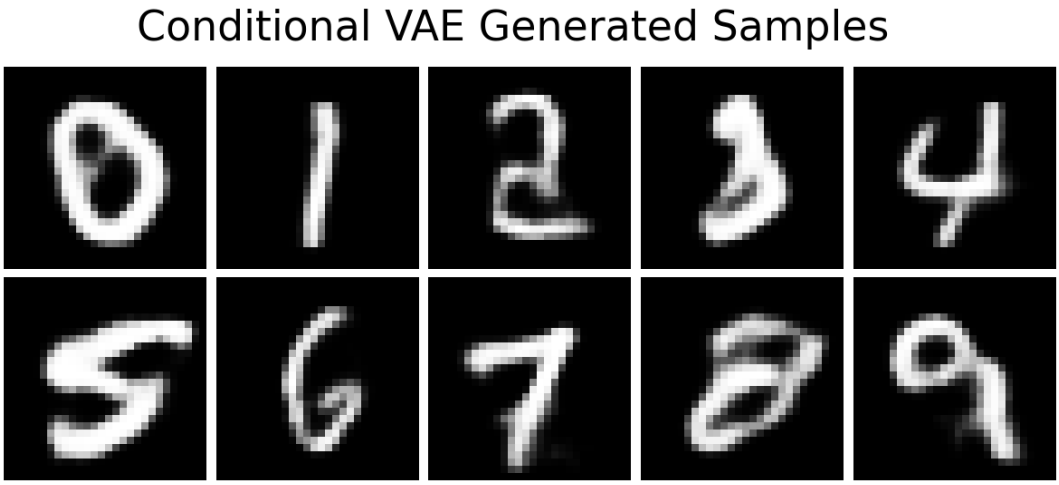

Figure 3: Samples from the conditional VAE where classes like 0, 2, 3, and 9 are distorted, showcasing the VAE's inability to learn the conditional distribution.

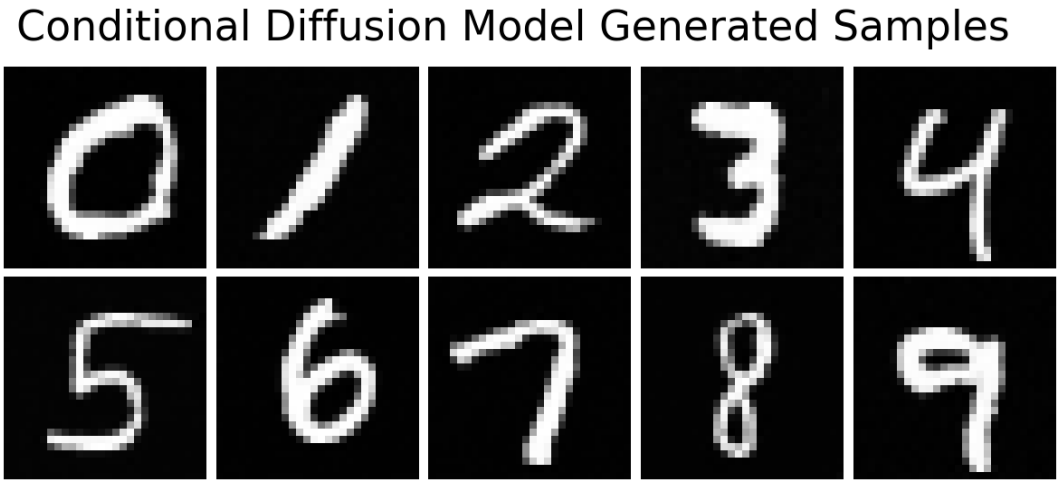

Figure 4: Samples from the conditional diffusion model where all digits are sharp and accurate, showcasing the diffusion model's ability to learn the conditional distribution.

J Evaluating conditional distribution with TARP

We complement the analysis in Section 3.3 by evaluating the calibration of the conditional generative models using TARP [12]. TARP assesses whether the posterior is well-calibrated by comparing the empirical coverage of credible intervals to their nominal levels. Well-calibrated models should yield curves that closely follow the identity line.

Figure 5 presents TARP curves for each digit class in MNIST, computed using 1000 posterior samples per condition and plotted with 1σ confidence intervals. The conditional diffusion model exhibits excellent calibration, with expected coverage matching the nominal credibility across all classes. In contrast, the CVAE's curves deviate substantially from the ideal diagonal, indicating systematic miscalibration and a failure to accurately capture the true conditional distributions.

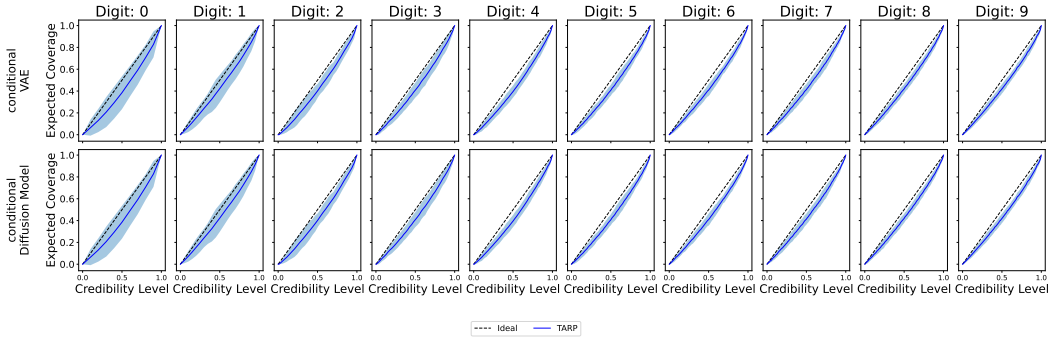

Figure 5: TARP plots for the conditional VAE (top) and conditional diffusion model (bottom). The shaded regions denote 1σ uncertainty bands. While the diffusion model tracks the identity line across all digits, the CVAE systematically undercovers.

These results align with the Pokie scores reported in Table 3, reinforcing that the diffusion model more accurately captures the true conditional distribution.

K Bayes Factor comparison

In this appendix, we provide a comparison between Pokie and the Bayes Factor. Unlike the Bayes factor, which relies on marginal likelihoods, Pokie operates directly in parameter space and does not require evidence computation. The Pokie score is bounded between $\frac{1}{2}$ and $\frac{2}{3}$, providing a probabilistic metric that is both interpretable and computationally efficient, even in high-dimensional parameter settings.

Despite these advantages, Pokie shares a fundamental limitation with the Bayes factor: equality of posterior distributions does not imply model equivalence. The two approaches are therefore complementary. The Bayes factor evaluates models through evidence, whereas Pokie evaluates agreement between posterior distributions. This distinction is relevant in cases where models yield similar evidence but differ in their posterior structure, such as when Expectation Maximization is used to update priors [2, 20], which may increase evidence while producing inaccurate posteriors.

We now compare these Pokie and the Bayes Factor approaches in two settings: the linear regression experiment from Section 3.1 and the distributional shift experiment from Appendix E. These comparisons evaluate the ability of Pokie and the Bayes Factor to identify well-calibrated posteriors and to detect models that are poorly calibrated.

K.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.

As shown in Table 7, 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.

Table 7: 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

K.2 Gaussian Mixture Model Shifts

We compute Bayes Factor scores for the GMM shift experiment described in Appendix E. 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:

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

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

Table 8: 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 8, 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.

L 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.

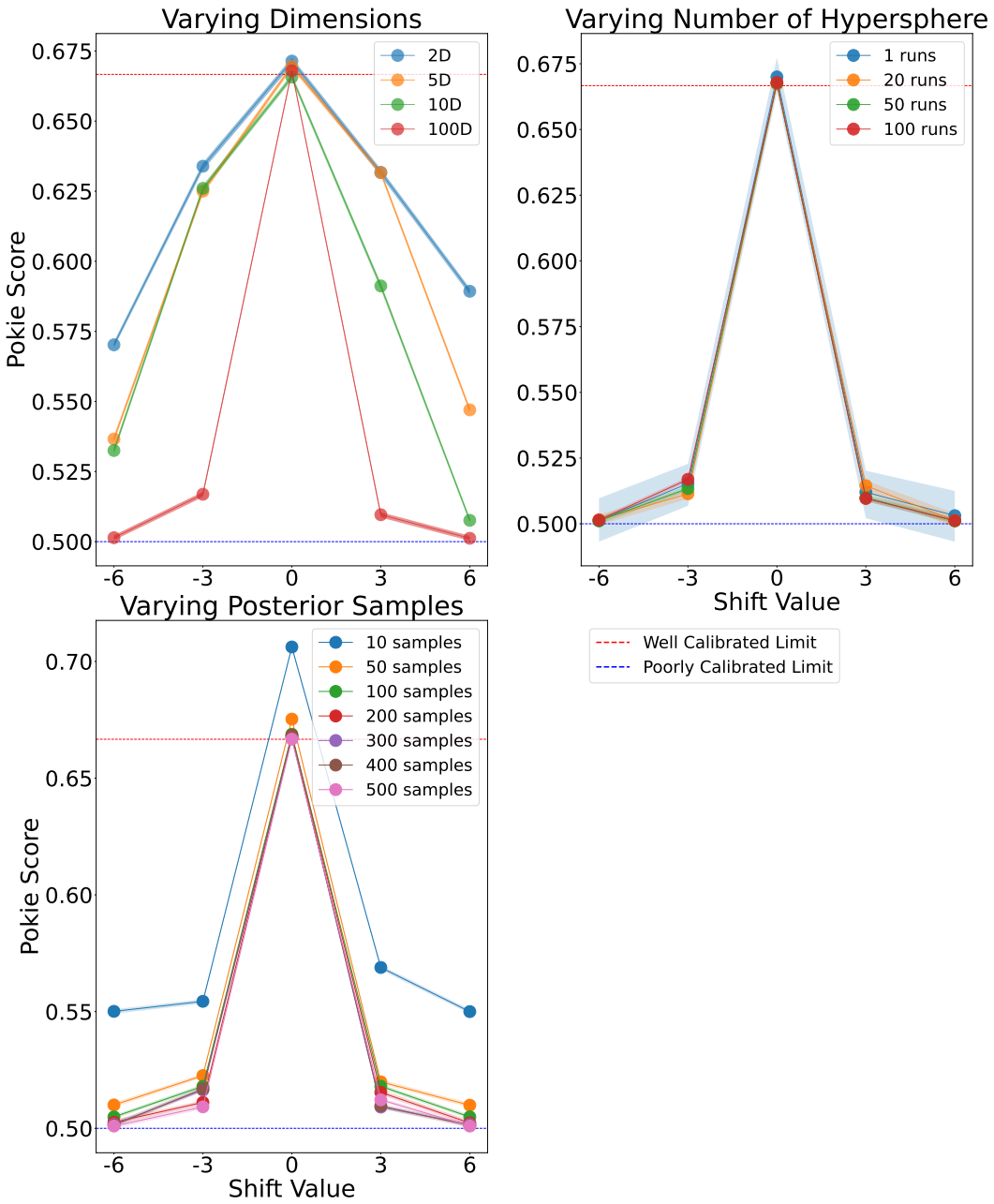

Figure 6: Pokie score sensitivity under varying experimental conditions. Top-left: effect of dimension on score; Top-right: number of hyperspheres per fiducial; Bottom-left: number of posterior samples. Across settings, Pokie scores peak for the well-calibrated ($\ell = 0$) model and fall to the poorly calibrated limit with increasing shift.

To assess sensitivity to posterior characteristics, we use the experiment introduced in Appendix E, and vary the following: dimensionality, number of hyperspheres per fiducial, and the number of posterior samples. Results are summarized in Figure 6.

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

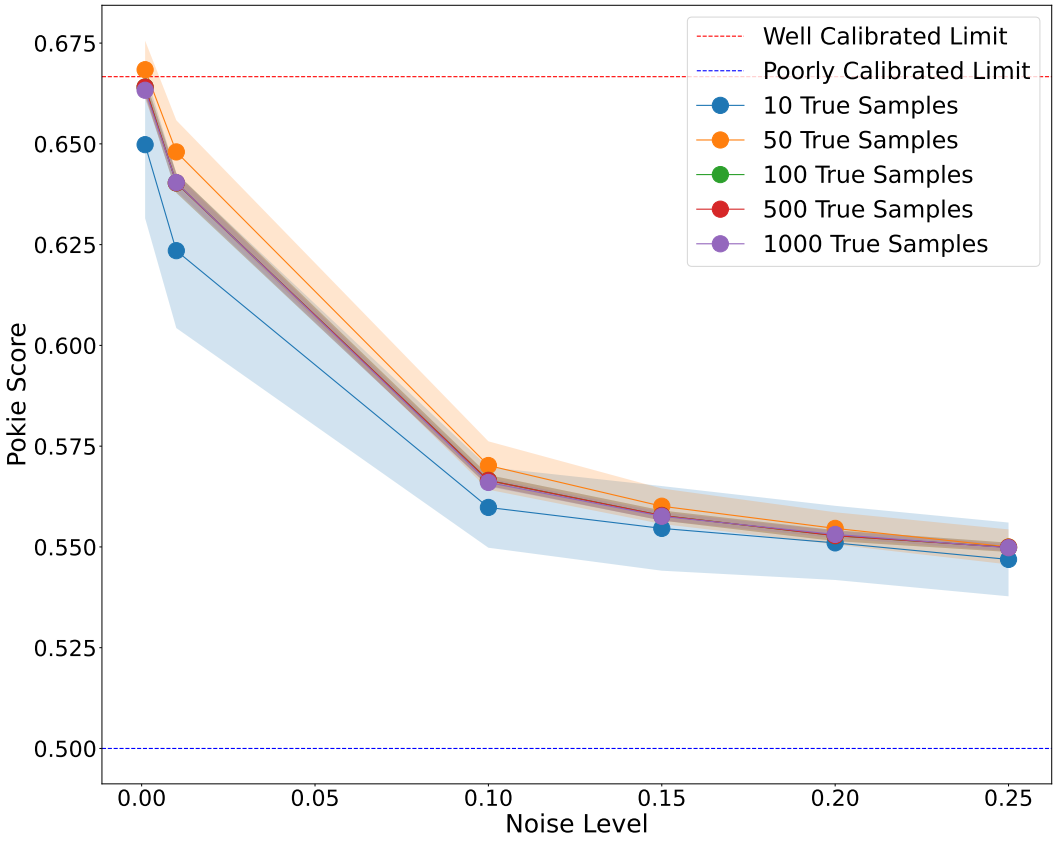

Figure 7: Pokie score vs noise level for varying counts of true posteriors. As the number of true posteriors increases, confidence in the Pokie estimate improves, and the distinction between well- and poorly-calibrated models becomes more pronounced.

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 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 7.

As the number of true posteriors increases, 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, demonstrating robustness to the number of calibration targets.

In addition to sensitivity to data-related factors, we also evaluate the robustness of Pokie to the choice of the distance metric, the distribution used to sample centers, and the shape of the region.

First, we test Pokie's sensitivity to the choice of distance metric used to compute distances between the randomly sampled center, the posterior samples, and the ground-truth parameters. To isolate the effect of the metric, we replicate the linear regression experiment from Section 3.1. We evaluate Pokie using five distance metrics: L2 (Euclidean), L1 (Manhattan), Chebyshev, Cosine, and Minkowski (with $p = 3$). As shown in Table 9, all metrics successfully identify the least biased model (with $\eta = 0.001$) as the best-calibrated, and correctly rank the remaining models in order of increasing

noise. This demonstrates that Pokie is robust to the choice of distance metric, and we recommend that users select the metric most aligned with the structure of their domain or miscalibration pattern.

Table 9: Pokie Scores (68% CI) by Distance Metric and Noise Level

Noise Level	L2	L1	Chebyshev	Cosine	Minkowski
0.001	0.6640 ± 0.0007	0.6636 ± 0.0008	0.6640 ± 0.0007	0.6600 ± 0.0011	0.6639 ± 0.0007
0.010	0.6405 ± 0.0009	0.6417 ± 0.0007	0.6422 ± 0.0007	0.6471 ± 0.0011	0.6404 ± 0.0009
0.100	0.5660 ± 0.0005	0.5764 ± 0.0005	0.5617 ± 0.0006	0.6093 ± 0.0012	0.5625 ± 0.0006
0.150	0.5574 ± 0.0004	0.5681 ± 0.0004	0.5524 ± 0.0005	0.5994 ± 0.0012	0.5539 ± 0.0004
0.200	0.5528 ± 0.0004	0.5624 ± 0.0005	0.5484 ± 0.0005	0.5922 ± 0.0011	0.5498 ± 0.0005
0.250	0.5497 ± 0.0004	0.5580 ± 0.0004	0.5464 ± 0.0004	0.5865 ± 0.0011	0.5474 ± 0.0005

We then test Pokie 's sensitivity to the choice of distribution used to sample centers from the parameter space. To isolate this factor, we again replicate the linear regression setup from Section 3.1, using the same biased posterior construction. We compare three representative center distributions: the uniform distribution $\mathcal{U}[0, 1]$, the standard normal distribution $\mathcal{N}(0, 1)$, and the asymmetric Beta distribution $\mathcal{B}(2, 5)$.

As shown in Table 10, Pokie scores remain stable across all three distributions, particularly in low-noise regimes where all methods report nearly identical values. Under higher noise, the uniform distribution achieves slightly higher scores, though absolute differences remain modest. These results indicate that Pokie is robust to the choice of center distribution. We leave it to the user to define a distance metric that is informative for their use case.

Table 10: Pokie Scores (68% CI) by Center Distribution and Noise Level

Noise Level	$\mathcal{U}[0, 1]$	$\mathcal{N}(0, 1)$	$\mathcal{B}(2, 5)$
0.001	0.6637 ± 0.0008	0.6636 ± 0.0007	0.6645 ± 0.0006
0.010	0.6402 ± 0.0010	0.6396 ± 0.0009	0.6445 ± 0.0007
0.100	0.5660 ± 0.0005	0.5532 ± 0.0005	0.5550 ± 0.0005
0.150	0.5574 ± 0.0004	0.5412 ± 0.0004	0.5389 ± 0.0005
0.200	0.5528 ± 0.0004	0.5348 ± 0.0004	0.5294 ± 0.0005
0.250	0.5498 ± 0.0004	0.5305 ± 0.0005	0.5229 ± 0.0004

Lastly we test the sensitivity of Pokie on the choice of region. We test Pokie when using ellipses of different scales and demonstrate our result in Table 11, in which we repeat the experiment from Section 3.1. We note that the scaling of more complex shapes remains a limiting factor in high dimensional spaces.

Table 11: Pokie Scores (68% CI) by Ellipse Stretch Configuration and Noise Level

Noise Level	**(1.0, 2.0)**	**(0.5, 1.5)**	**(1.5, 1.0)**	**(2.0, 0.5)**	**(0.8, 1.2)**
0.001	0.6653 ± 0.0009	0.6681 ± 0.0009	0.6627 ± 0.0007	0.6618 ± 0.0009	0.6669 ± 0.0008
0.010	0.6504 ± 0.0008	0.6554 ± 0.0009	0.6358 ± 0.0009	0.6317 ± 0.0010	0.6494 ± 0.0008
0.100	0.5795 ± 0.0004	0.5805 ± 0.0007	0.5508 ± 0.0006	0.5347 ± 0.0006	0.5773 ± 0.0006
0.150	0.5647 ± 0.0003	0.5619 ± 0.0007	0.5417 ± 0.0005	0.5270 ± 0.0004	0.5657 ± 0.0005
0.200	0.5555 ± 0.0003	0.5504 ± 0.0005	0.5372 ± 0.0005	0.5247 ± 0.0004	0.5590 ± 0.0004
0.250	0.5492 ± 0.0003	0.5430 ± 0.0005	0.5345 ± 0.0004	0.5239 ± 0.0005	0.5543 ± 0.0004

Overall, Pokie remains robust given the sensitivity tests. 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. Pokie also remains flexible when selecting the center distribution, distance metric, or region shape. These properties make Pokie well-suited for practical use in SBI tasks under practical constraints.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, through the Methods section, Proof in the Appendix, and the experiments, we demonstrate that the claims made in the abstract and introduction accurately reflect the paper's contribution and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss our assumptions, the limitations of our metric, and perform a sensitive analysis to communicate the limits of Pokie.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Please see the Proofs in Appendix$D.1$.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide our code, the details of our experiments, and our model hyperparameters.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See link to repository at the end of the abstract.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and test details are provided in the appendix for all relevant experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experiments show the standard deviation of our Pokie score. Furthermore, the Tarp results also showcase their error as well.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details about compute resources are provided in the relevant experiments, and we note that unless otherwise specified, we use the MacBook Air 8 GB Ram to run experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We adhere to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper serves no direct societal impacts. It simply provides a way to assess the quality of the posterior distribution.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All works are credited, correctly cited, and terms of use are properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Source Code for Pokie is provided, and details of how to use are provided in the ReadMe, Notebooks, and source code itself.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: This paper and the developed metric in this research do not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.