More Expressive Amortized Bayesian Inference via Joint Learning and Self-Consistency?

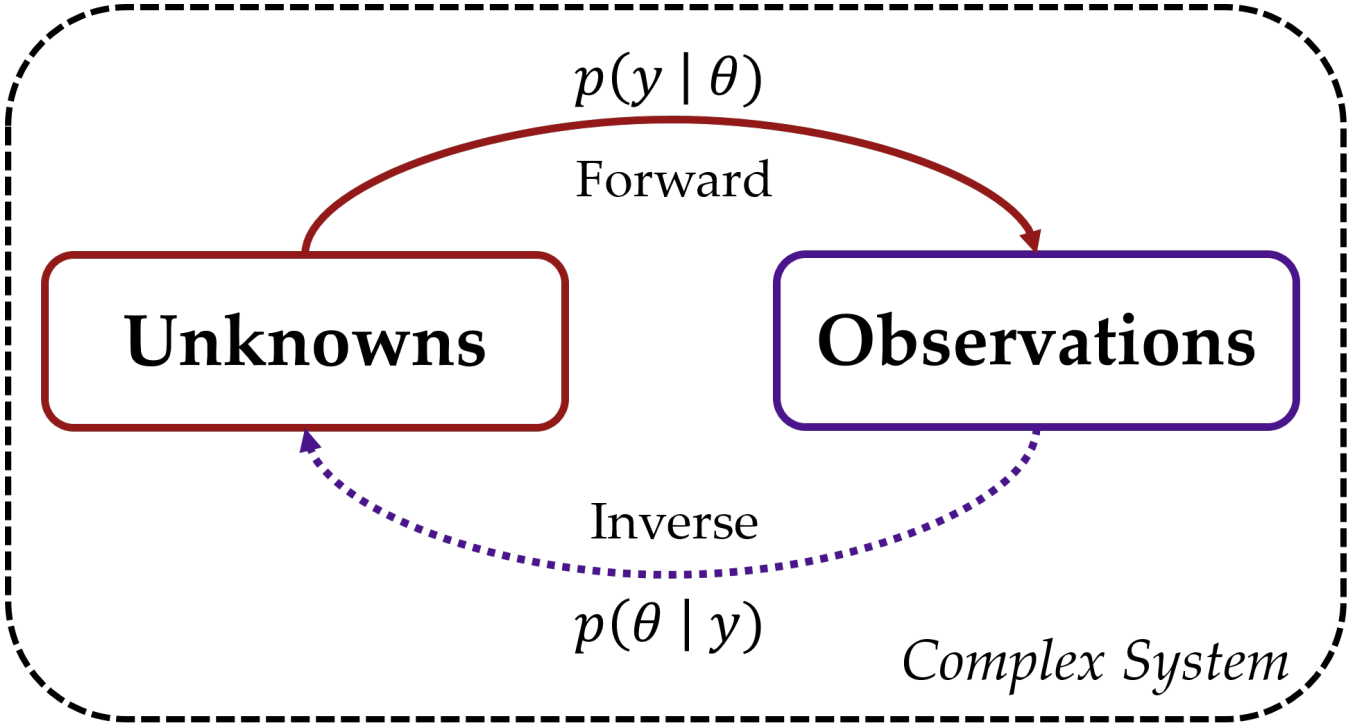
Stefan T. Radev

Cognitive Science Department
Rensselaer Polytechnic Institute

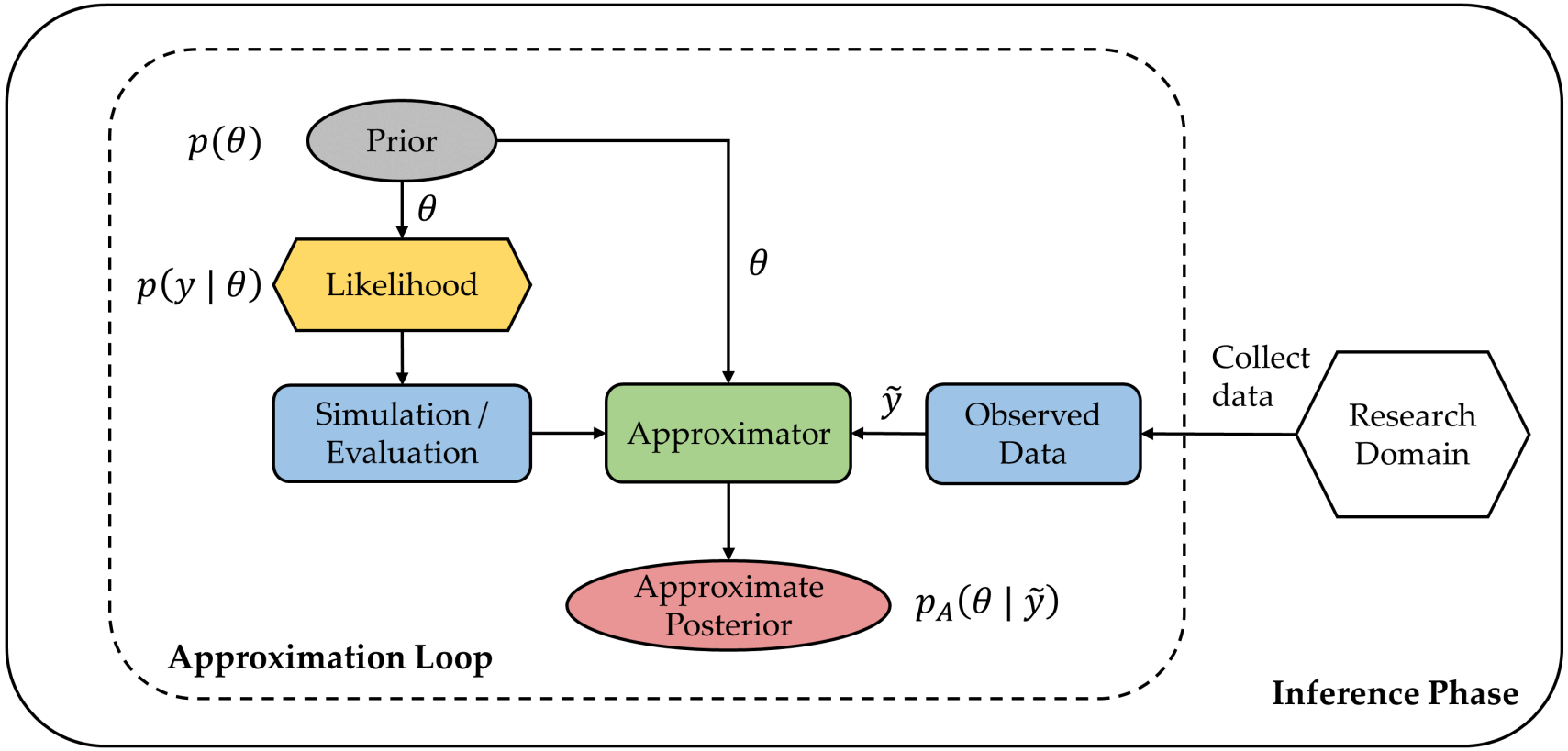
October 26, 2023

Agenda

Problem Setting

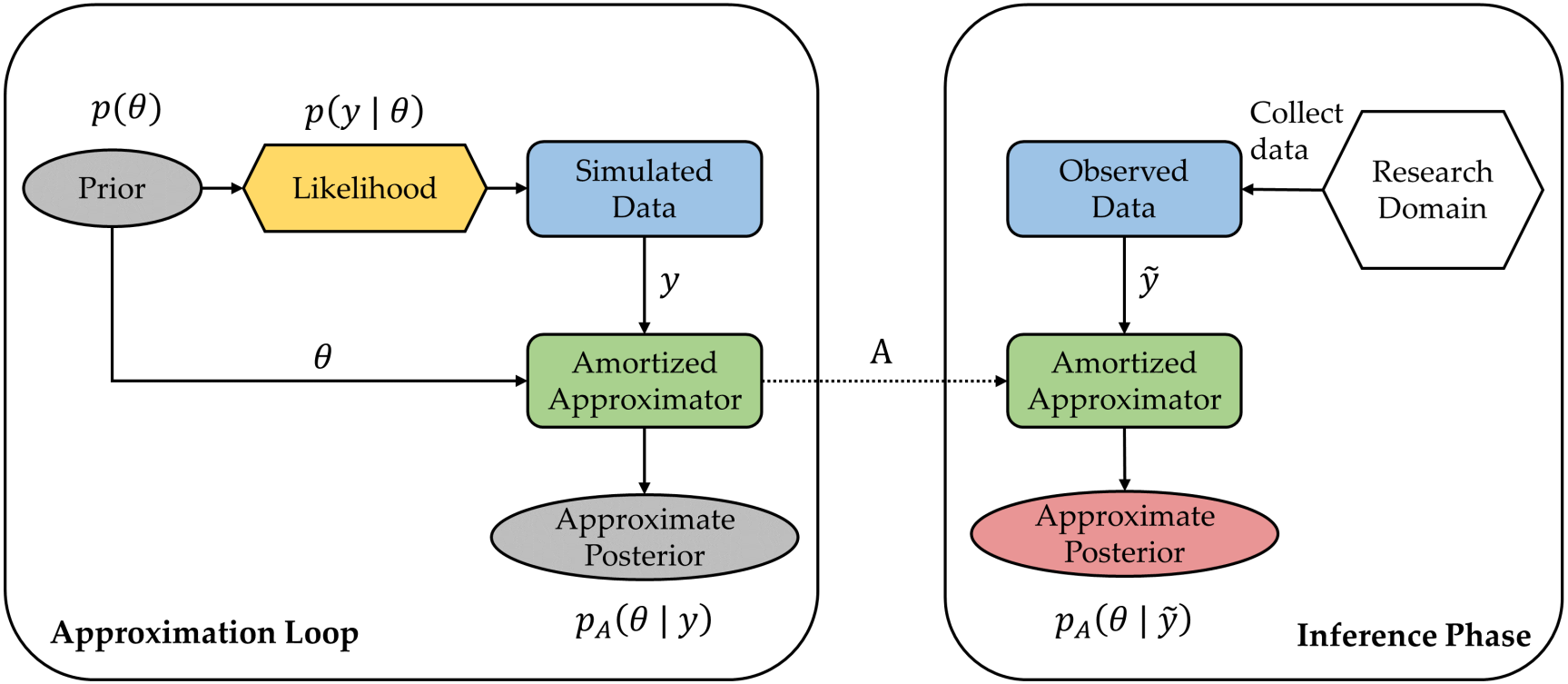


Non-Amortized Bayesian Inference



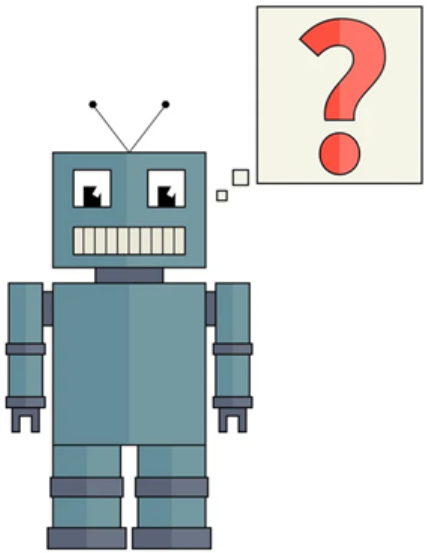
Approximation and inference are coupled. No resource pooling.

Amortized Bayesian Inference



Approximation and inference are decoupled. Pooling of resources.

More Expressive?



1. Qualitative: New capabilities / new tasks.
2. Quantitative: Better performance in certain scenarios.

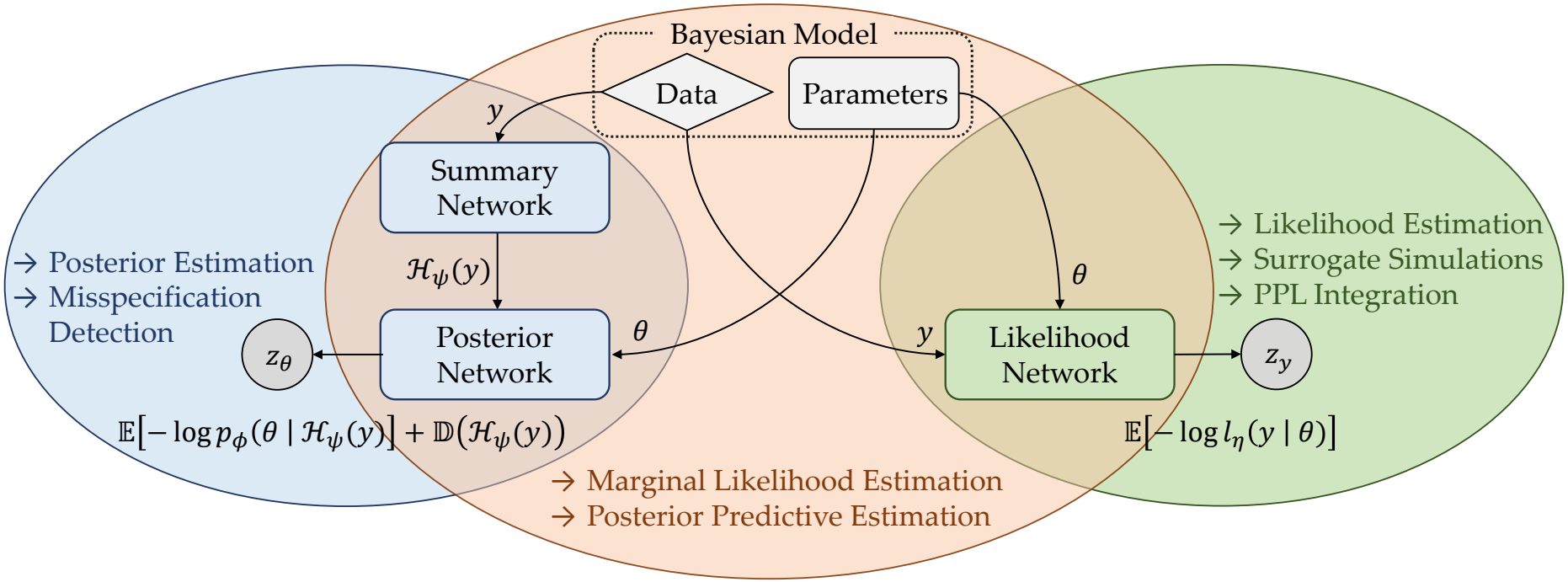
Desiderata

1. End-to-end learning – enables automatic learning of maximally informative summary statistics.
2. Fully amortized inference (i.e., pre-paid computational effort) – enables fully Bayesian inference on millions of data sets.
3. No structural restrictions on priors, likelihoods, or posteriors – enables creativity on the part of modelers.
4. Reliable inference in the open world – option to reject queries for which no competent inference can be provided.
5. Offline efficiency – reasonable approximations for limited simulation budgets (i.e., expensive simulators).

Agenda

Joint Learning: General Overview

- JANA (Radev et al., 2023), inspired by SNPLA (Wiqvist, Frellsen, & Picchini, 2021):



Simulation-Based Training

- Three neural networks amortize various Bayesian tasks:
  1. A summary network $\mathcal{H}_\psi(y)$ compresses complicated data (e.g., entire data sets or multivariate time series) into informative embeddings.
  2. A posterior network $\mathcal{P}_\phi(\theta; \mathcal{H}_\psi(y))$ approximates the true posterior given the outputs of the summary network.
  3. A likelihood network $\mathcal{L}_\eta(y; \theta)$ approximates the true likelihood of the raw data.
- Criterion:

$$\min_{\phi,\psi,\eta} \mathbb{E}_{p(\theta,y)} \big[-(\log p_\phi(\theta \mid \mathcal{H}_\psi(y)) + \log l_\eta(y \mid \theta)) \big] + \lambda \cdot \mathbb{MMD}^2 \big[p(\mathcal{H}_\psi(y)) \,||\, p(z) \big] \tag{1}$$

Why the Open World Matters

- What we want to optimize:

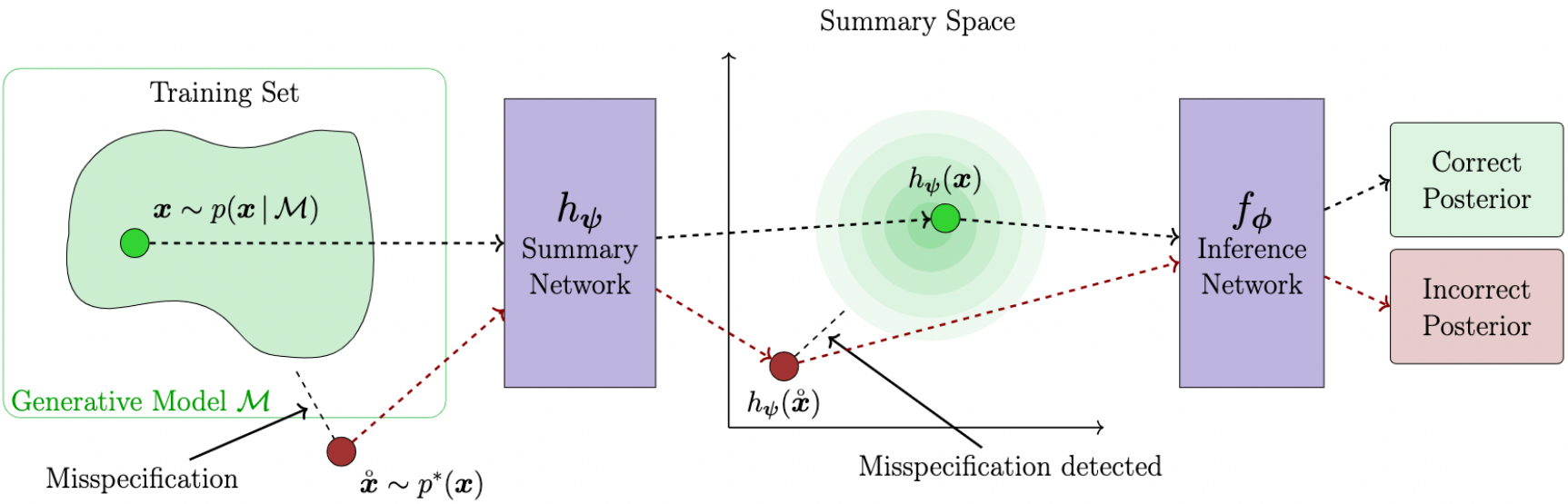
$$
\begin{aligned}
(\phi^*, \psi^*) &= \arg\min_{\phi,\psi} \mathbb{E}_{p^*(y)} \left[\mathbb{KL}(p(\theta \mid y) \,||\, p_\phi(\theta \mid \mathcal{H}_\psi(y)))\right] \\
&= \arg\min_{\phi,\psi} \mathbb{E}_{p^*(y)} \left[\mathbb{E}_{p(\theta \mid y)} \left[\log p(\theta \mid y) - \log p_\phi(\theta \mid \mathcal{H}_\psi(y))\right]\right] \\
&= \arg\min_{\phi,\psi} \mathbb{E}_{p^*(y)} \left[\mathbb{E}_{p(\theta \mid y)} \left[-\log p_\phi(\theta \mid \mathcal{H}_\psi(y))\right]\right] \qquad (2)
\end{aligned}
$$

- What we actually optimize:

$$
\begin{aligned}
(\hat{\phi}, \hat{\psi}) &= \arg\min_{\phi,\psi} \mathbb{E}_{p(y)} \left[\mathbb{E}_{p(\theta \mid y)} \left[-\log p_\phi(\theta \mid \mathcal{H}_\psi(y))\right]\right] \\
&= \arg\min_{\phi,\psi} \mathbb{E}_{p(\theta,y)} \left[-\log p_\phi(\theta \mid \mathcal{H}_\psi(y))\right] \qquad (3)
\end{aligned}
$$

Out-of-Simulation (OOSim) Detection

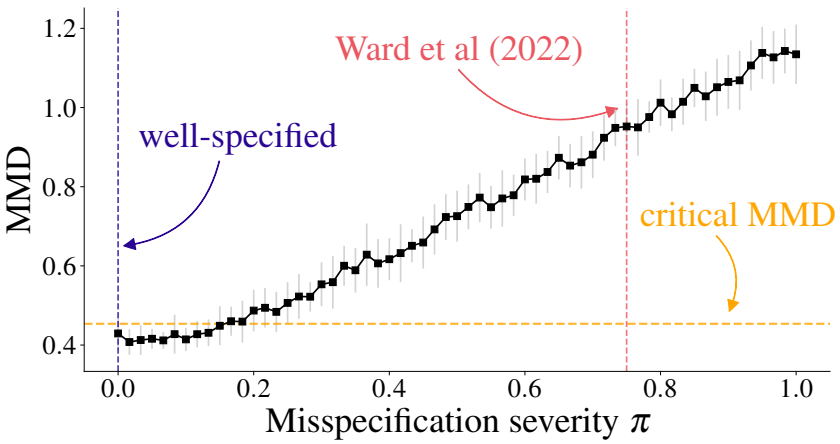
- Re-frame model misspecification as out-of-distribution (OOD) detection (Schmitt et al., 2021).
- Trust inferences only on inlier observations:



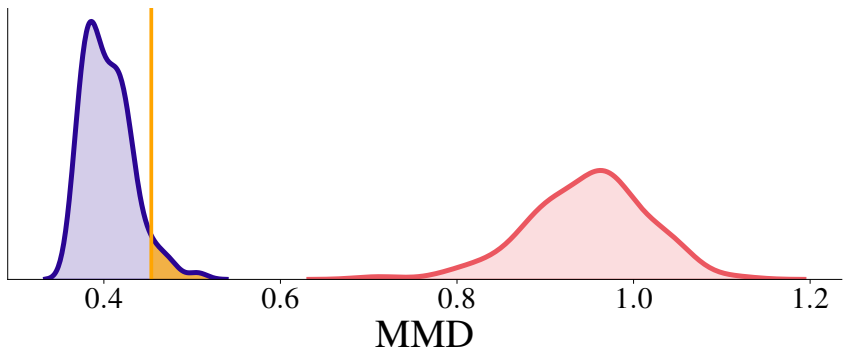
- See Ward, Cannon, Beaumont, Fasiolo, and Schmon (2022) for an MCMC-based correction following detection.

OOSim in Action

- MMD can reliably highlight simulation gaps (Schmitt et al., 2021):



(a) MMD by misspecification severity.



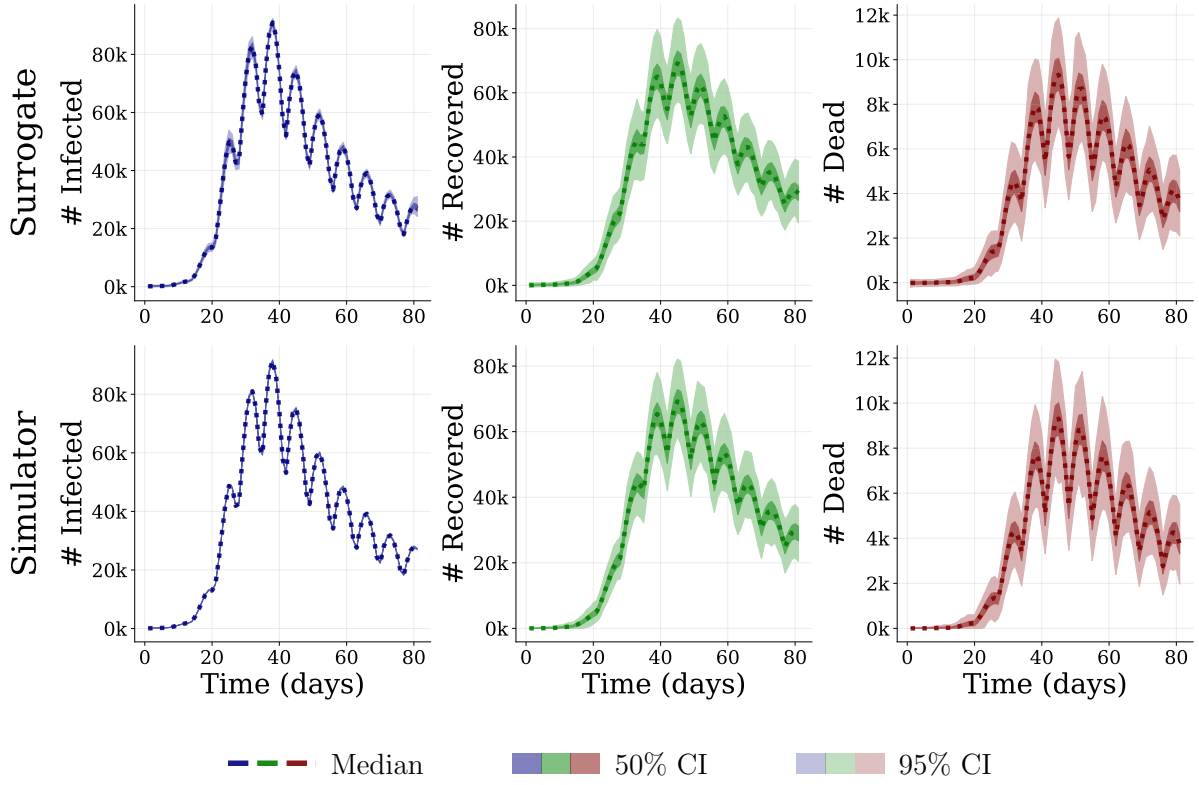
(b) Power for $\pi = 0.75$ is essentially 1.

Training model $\mathscr{M}$ | Observed process $\mathscr{M}^*$ | H_0 rejection area

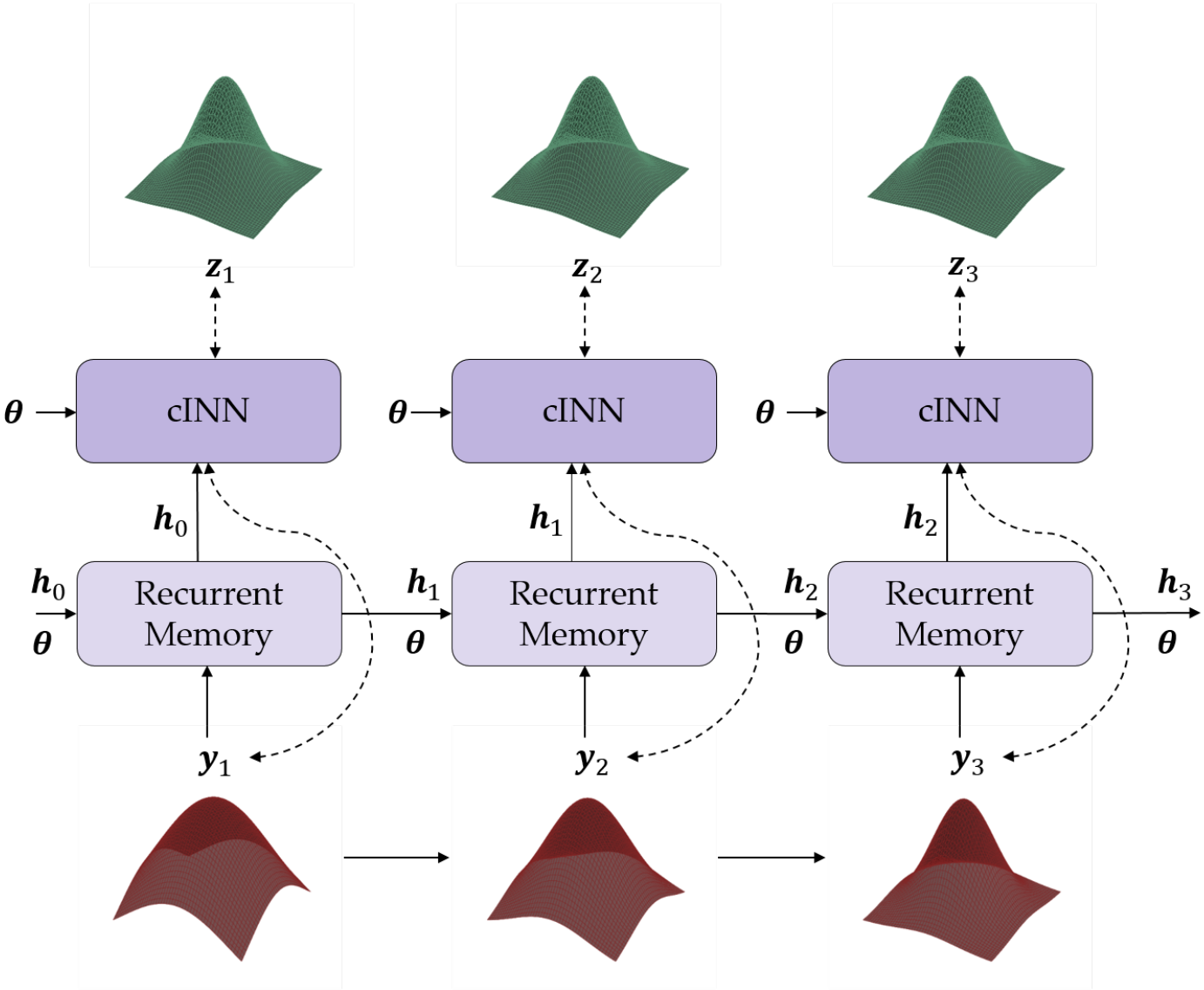
Bonus: We can also use MMD to reject bad surrogate simulations.

Amortized Likelihood Emulation

- Upon convergence, we can generate surrogate simulations for any θ with or without summary statistics.
- Complex COVID-19 model with 34 (identifiable and non-identifiable) parameters (Radev et al., 2021):

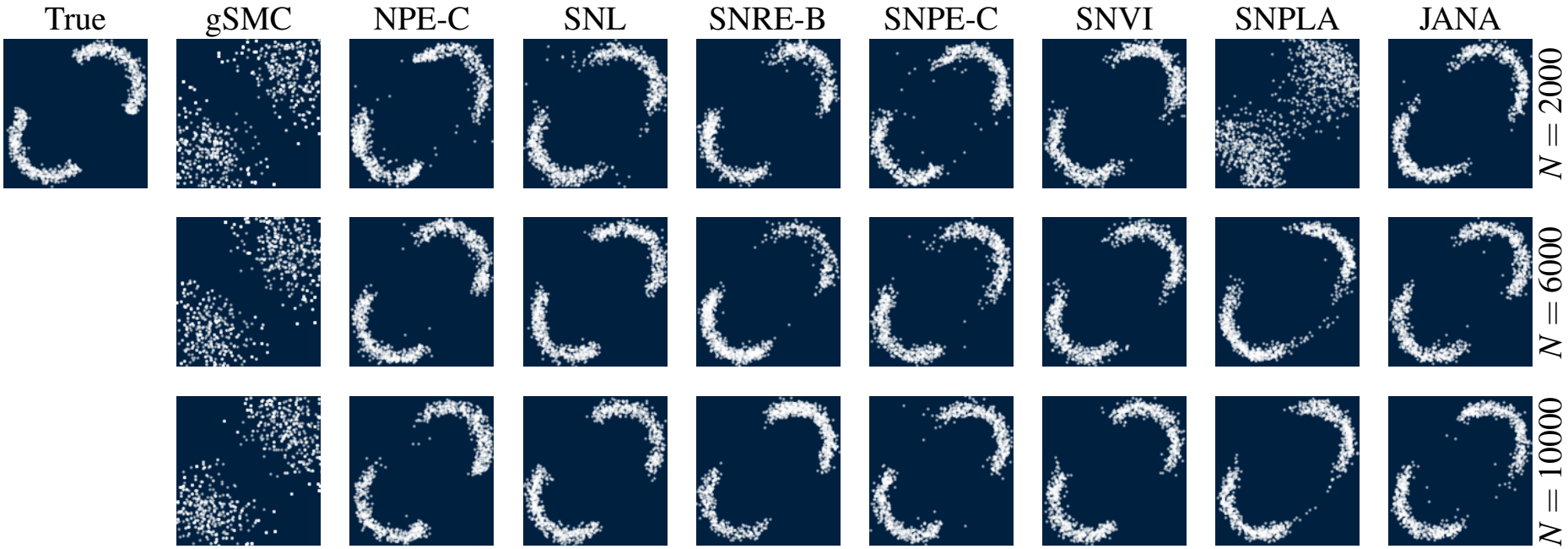


Aside: Recurrent Flows



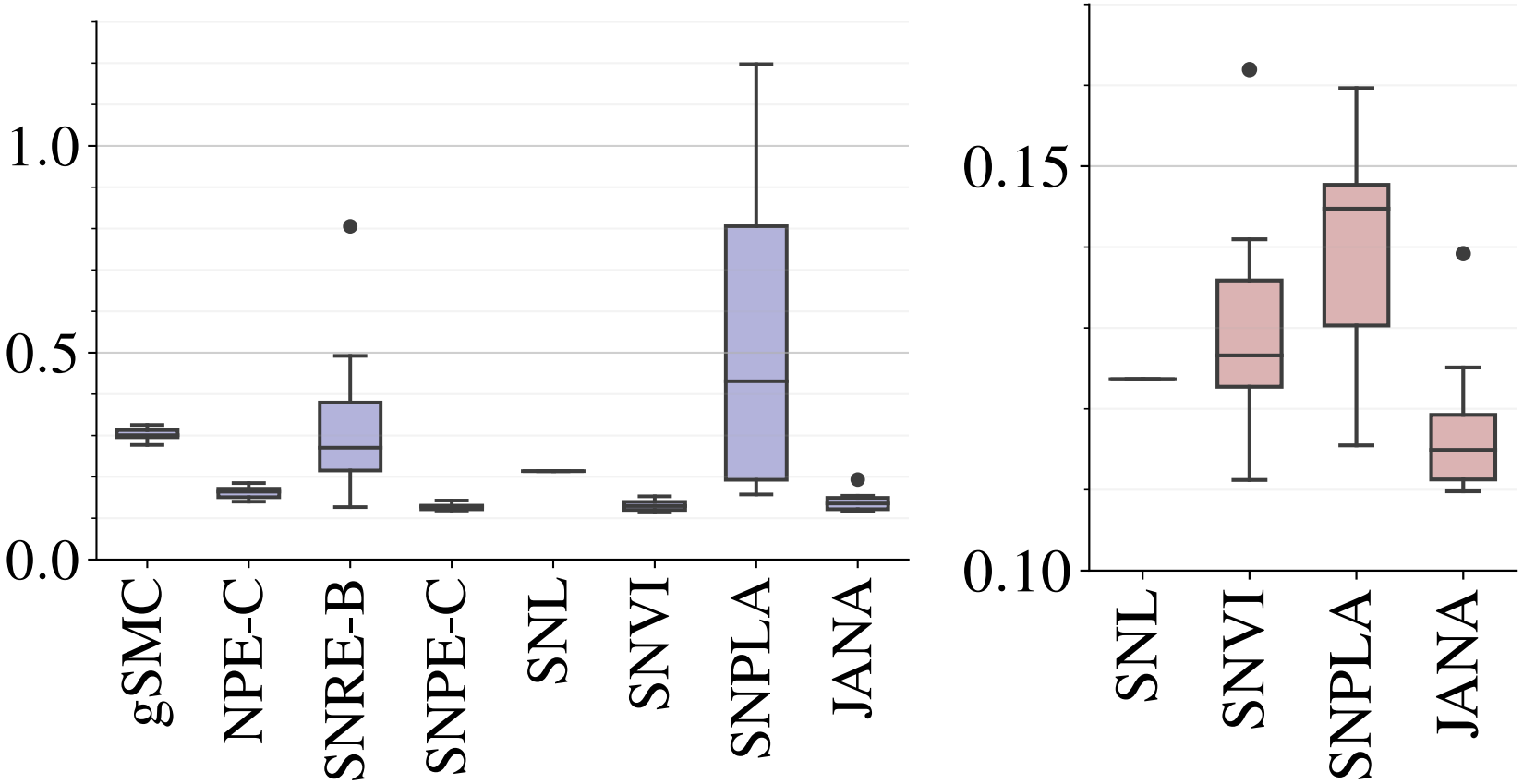
Some Benchmarking: Two Moons (I)

- Amortized posterior inference is on par with non-amortized inference on weird posteriors:



Some Benchmarking: Two Moons (II)

- A more detailed look ($N = 10\,000$):



(a) Posterior MMD

(b) Joint MMD

Small World Validation Methodology

- Simulation-based calibration (Talts et al., 2018, SBC): For all quantiles $q \in (0, 1)$, all uncertainty regions are well calibrated, as long as we have the true model and posterior computation is exact. Formally:

$$q = \int_{\mathcal{Y}} \int_{\Theta} \mathbb{I}\left[\theta^* \in U_q(\theta \mid y)\right] p(\theta^*, y)\, d\theta^* dy \tag{4}$$

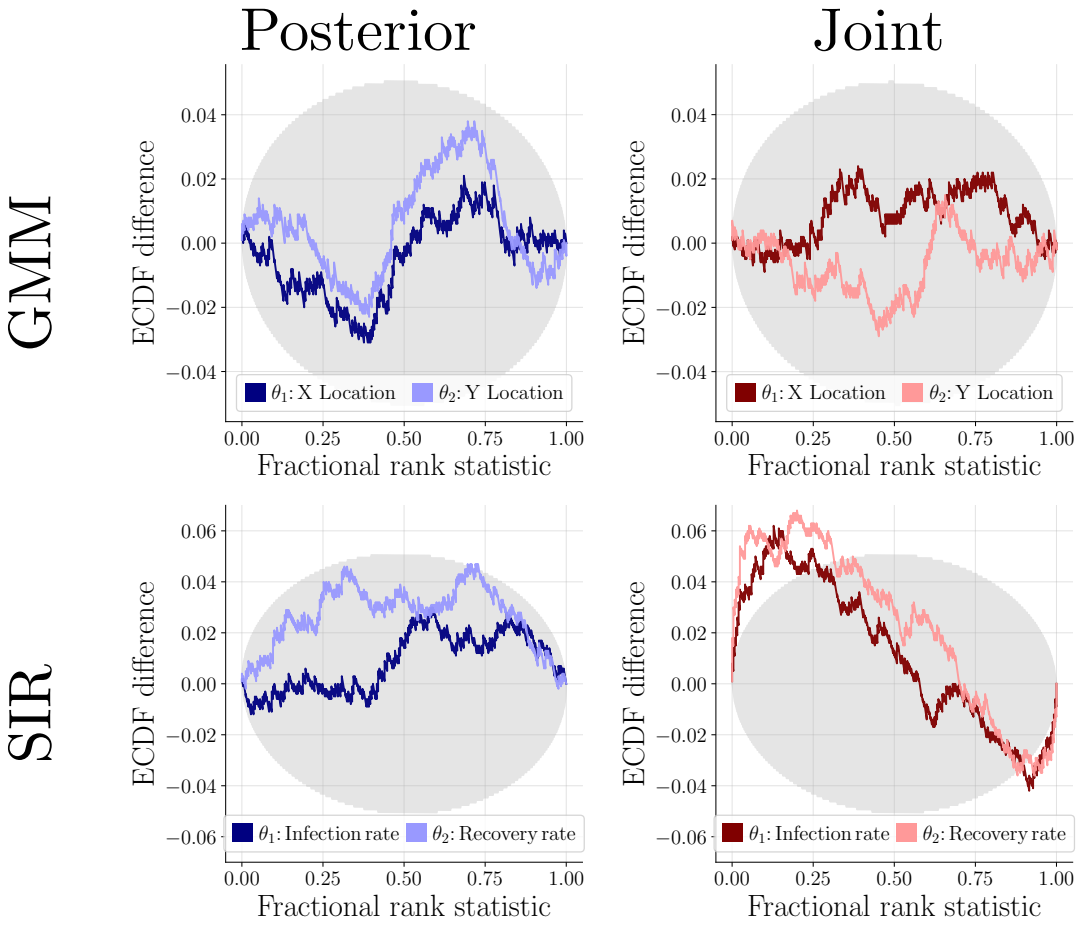
- Idea: Approximate SBC via simulations from the generative model and (fractional) rank statistics of the posterior samples:

$$R(\theta_m^*, \theta_{1:S}) = \sum_{s=1}^{S} \mathbb{I}\left[\theta_m^* > \theta_s\right] \tag{5}$$

- SBC can be performed for free thanks to amortized inference!

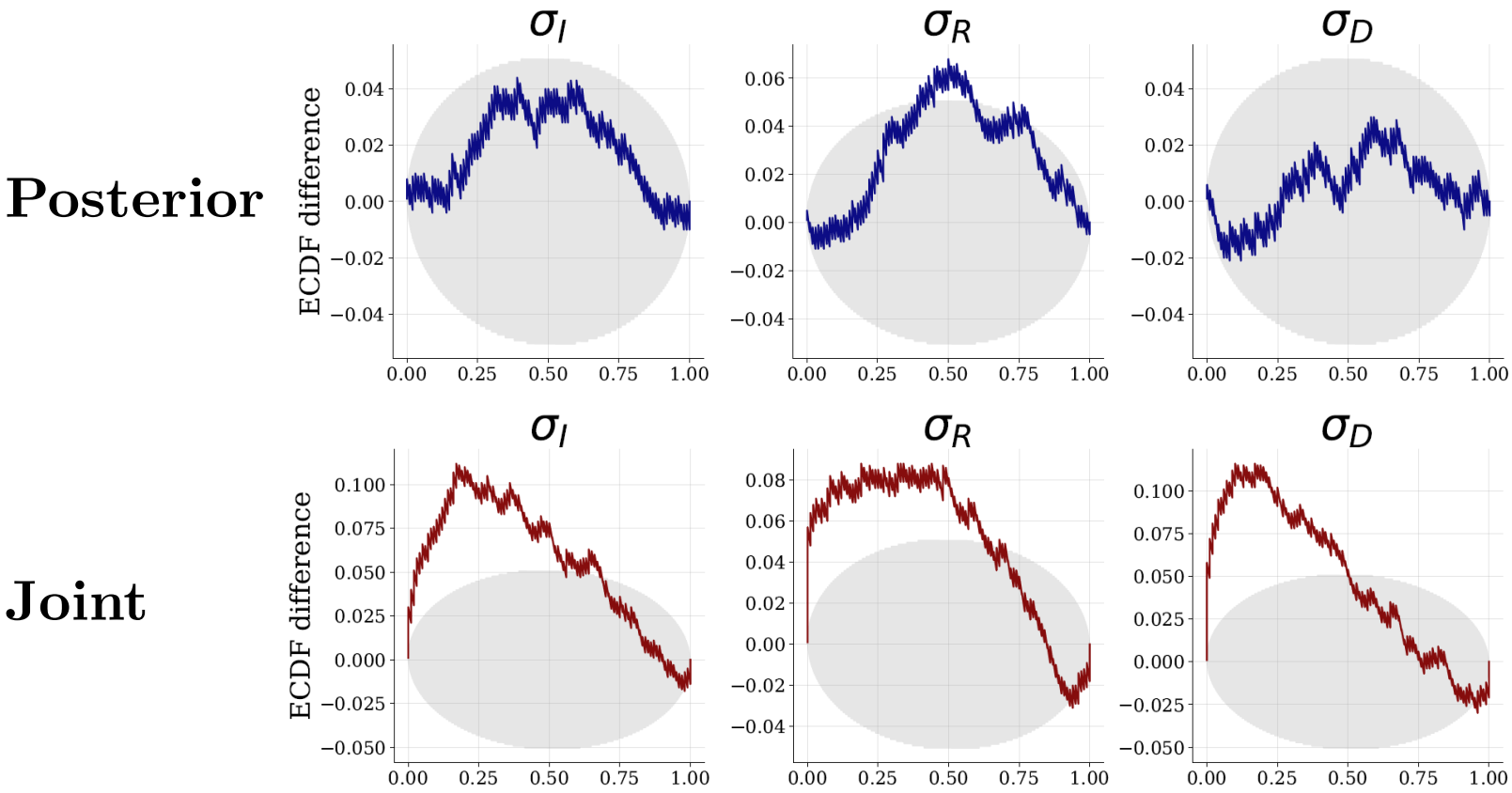
Some Benchmarking: Joint SBC (I)

- Joint simulation-based calibration can detect subtle deficiencies on toy benchmark examples (Lueckmann et al., 2021):



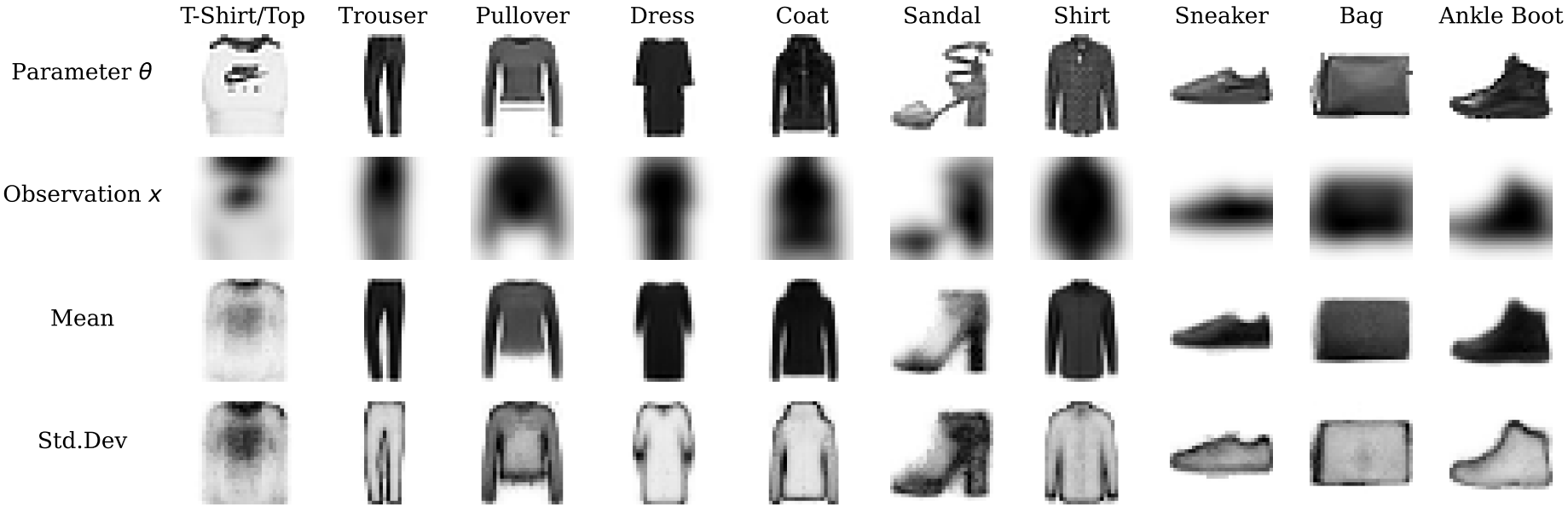
Some Benchmarking: Joint SBC (II)

- But also on representative examples, e.g., aleatoric noise parameters (Radev et al., 2021):



Some Benchmarks: Higher Dimensions

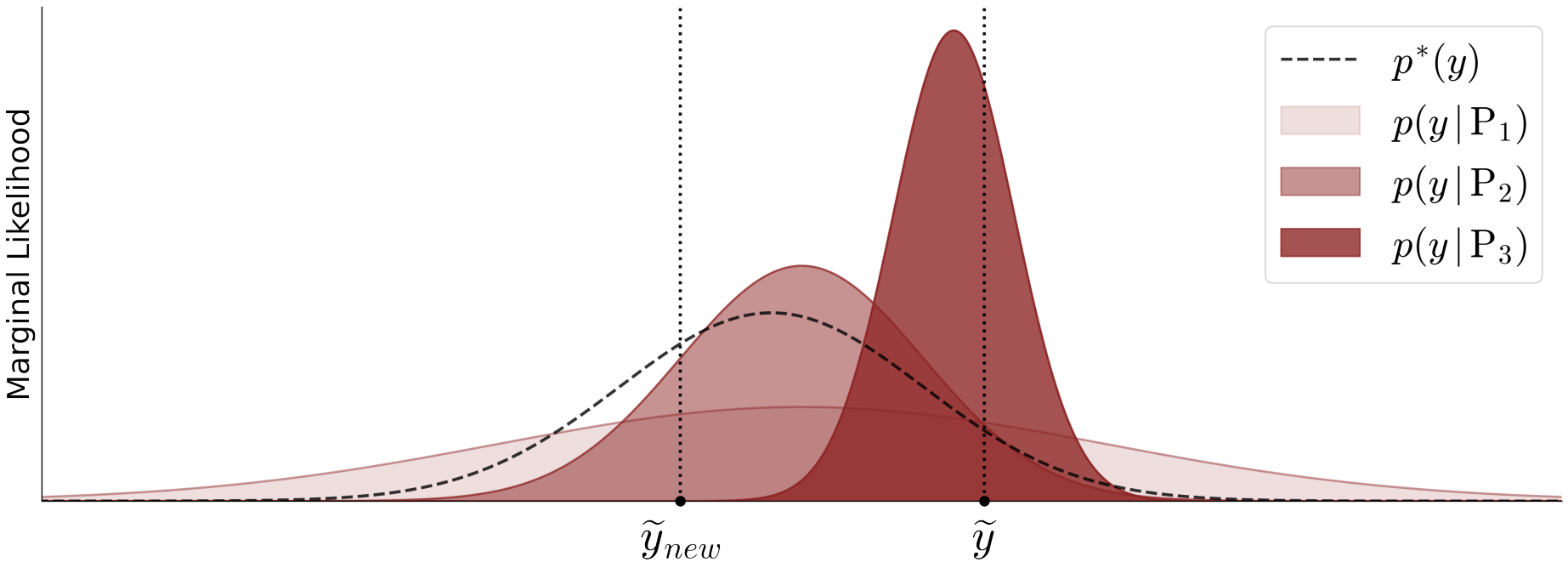
- Bayesian denoising of Fashion MNIST (inspired by Ramesh et al., 2022):



Marginal Likelihoods and Occam's Razor

- Canonical quantity for prior predictive model comparison:

$$p(y) = \int_{\Theta} p(y \mid \theta)\, p(\theta)\, d\theta \tag{6}$$



Approximating Marginal Likelihoods

- Bayes' rule as a probabilistic change of variable (Gelfand & Dey, 1994):

$$p(y) = \frac{p(\theta)}{p(\theta \mid y)} p(y \mid \theta) \tag{7}$$

- Approximate log marginal lieklihood (LML) through all three networks:

$$\log \hat{p}(y) = \log p(\theta) + \log l_\eta(y \mid \theta) - \log p_\phi(\theta \mid \mathcal{H}_\psi(y)) \tag{8}$$

- Variation in the RHS for different θ values is a measure of (epistemic) approximation error. Will use later as a self-consistency loss.

Approximating Posterior Predictive Performance

- Posterior predictive quantities (fixed y):

$$p(y_{\text{new}} \mid y) = \int_{\Theta} p(y_{\text{new}} \mid \theta, y)\, p(\theta \mid y) d\theta \tag{9}$$

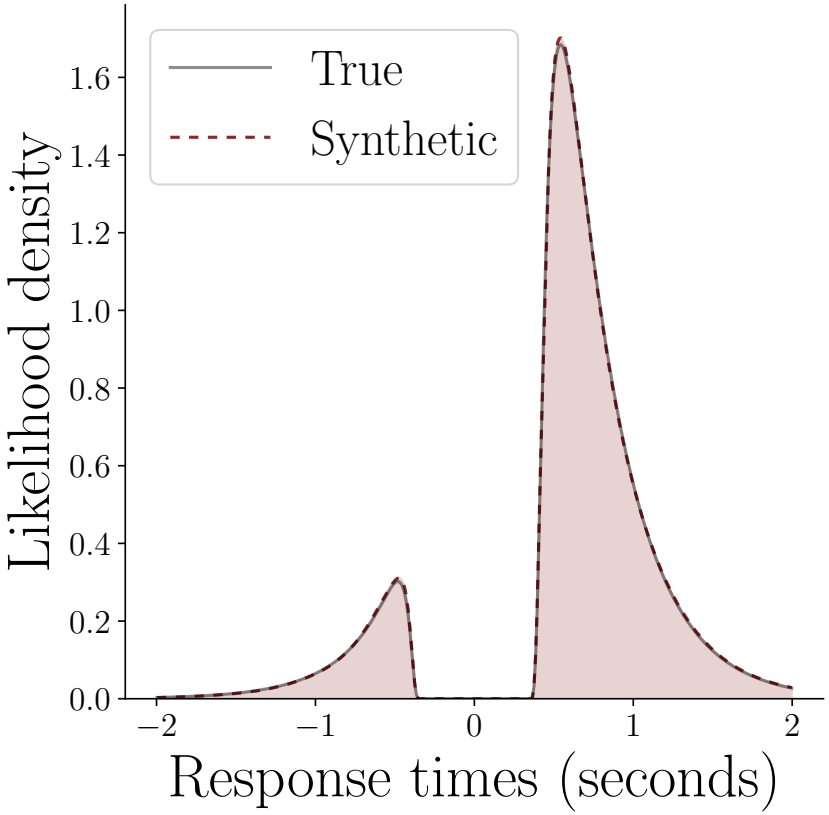
$$\text{ELPD} = \sum_{m=1}^{M} \log p(y_{\text{new}}^{(m)} \mid y) \tag{10}$$

- Monte Carlo approximation via JANA (fixed y):

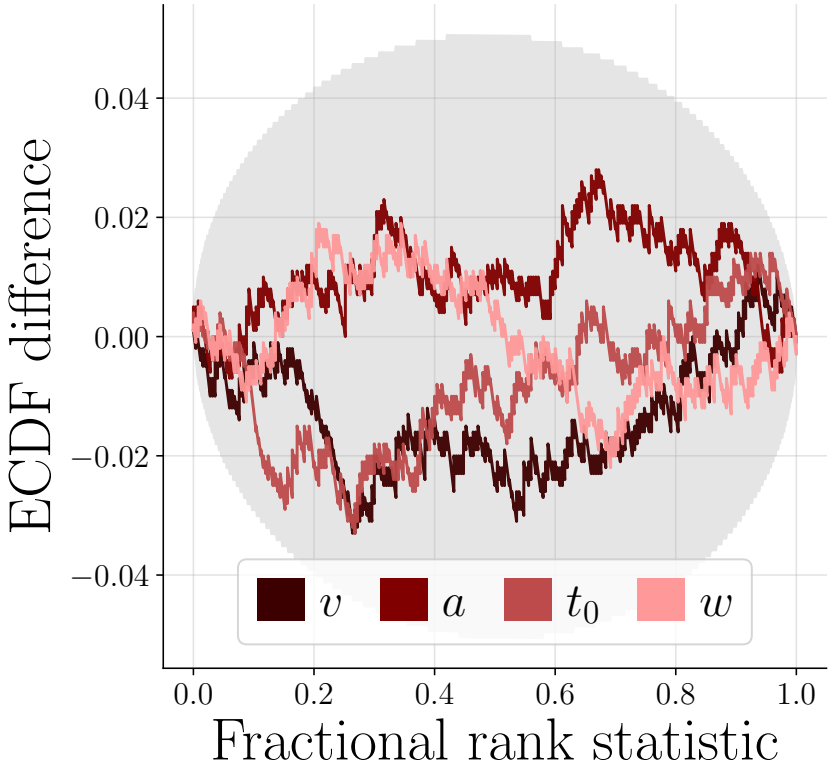
$$\theta^{(s)} \sim p_{\phi}(\theta \mid \mathcal{H}_{\psi}(y)) \quad \text{for} \quad s = 1, \ldots, S \tag{11}$$

$$\text{ELPD} \approx \sum_{m=1}^{M} \log \frac{1}{S} \sum_{s=1}^{S} l_{\eta}(y_{\text{new}}^{(m)} \mid \theta^{(s)}, y) \tag{12}$$

Results: Diffusion Model of Decision Making (I)

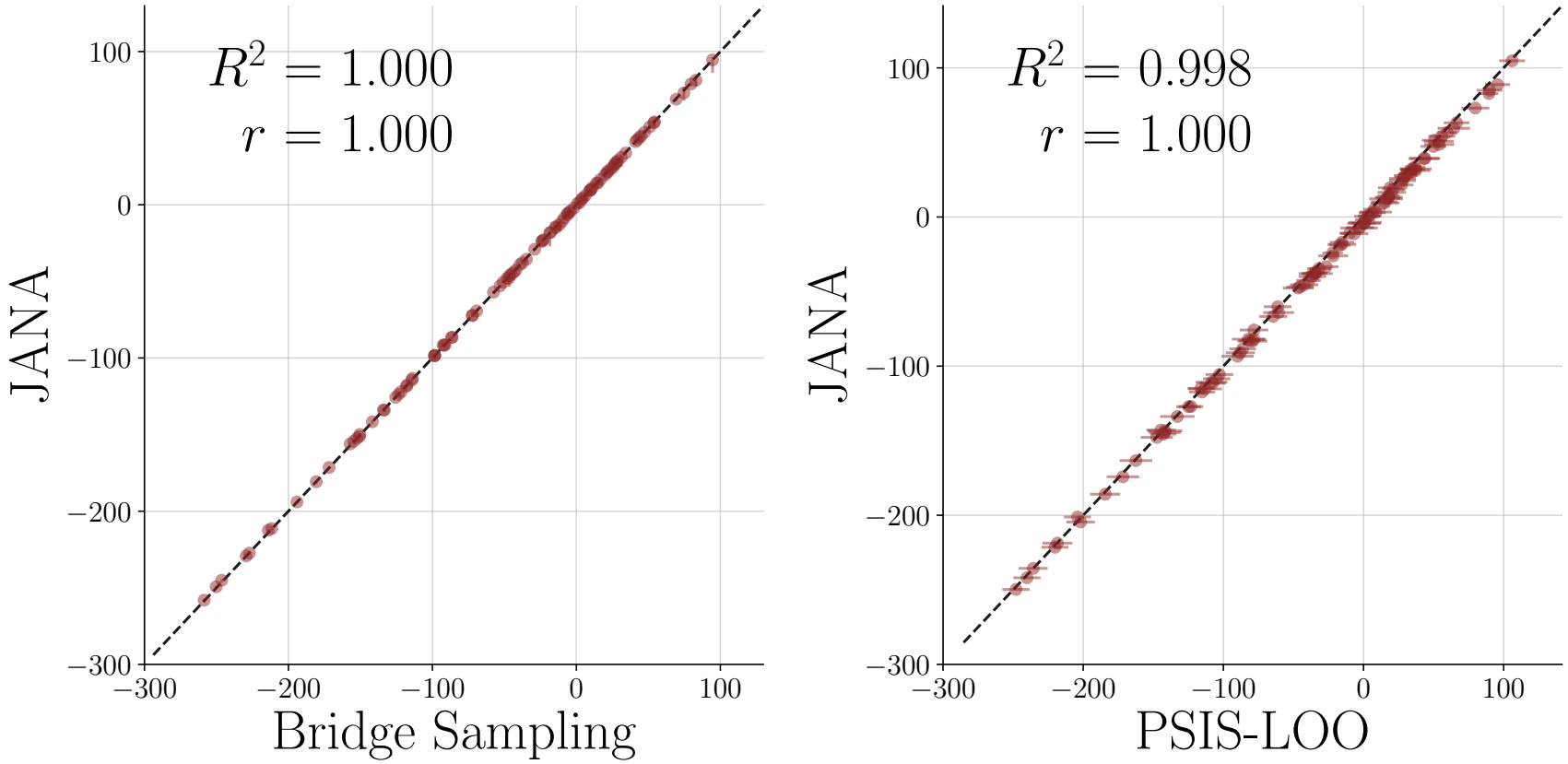


(a) Likelihood emulation



(b) Joint calibration

Results: Diffusion Model of Decision Making (II)



(a) Prior predictive (LML)

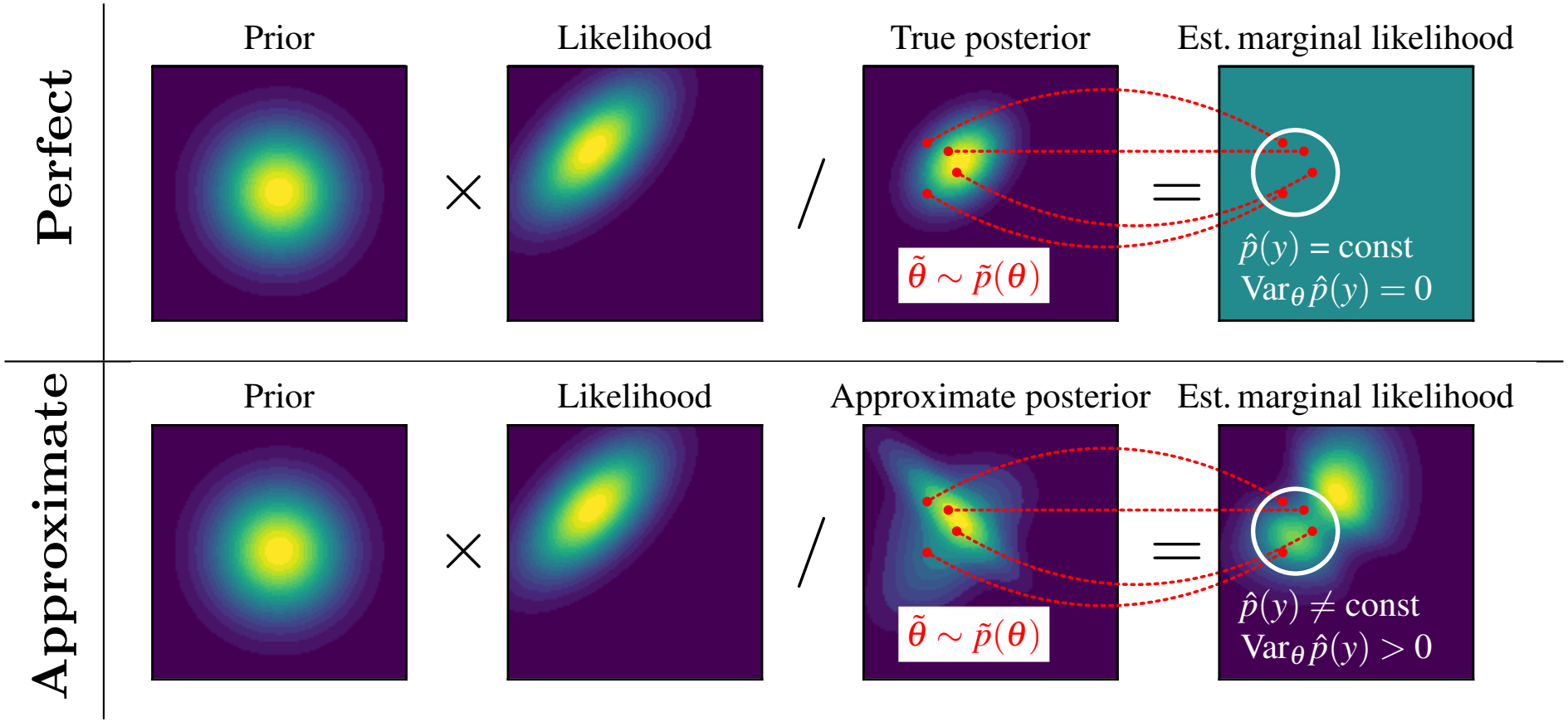
(b) Post. predictive (ELPD)

Agenda

Bayes' Rule and Symmetry

- Trivial, but typically non-actionable observation:

$$p(y) = p(\theta)\, p(y \mid \theta) / p(\theta \mid y) \tag{13}$$



Self-Consistency Loss

- Formulate unwanted variability as a loss (Schmitt et al., 2023):

$$\mathcal{L}_{\mathrm{SC}}(\phi, y) := \mathrm{Var}_{\tilde{\theta} \sim \tilde{p}(\theta)} \Big(\log p(\tilde{\theta}) + \log p(y \mid \tilde{\theta}) - \log p_\phi(\tilde{\theta} \mid y) \Big), \quad (14)$$

- Total loss (e.g., NPE):

$$\mathcal{L}_{\text{NPE-SC}}(\phi) := \mathbb{E}_{p(y)} \Big[\mathbb{E}_{p(\theta \mid y)} \big[-\log p_\phi(\theta \mid y) \big] + \lambda \, \mathcal{L}_{\mathrm{SC}}(\phi, y) \Big] \quad (15)$$

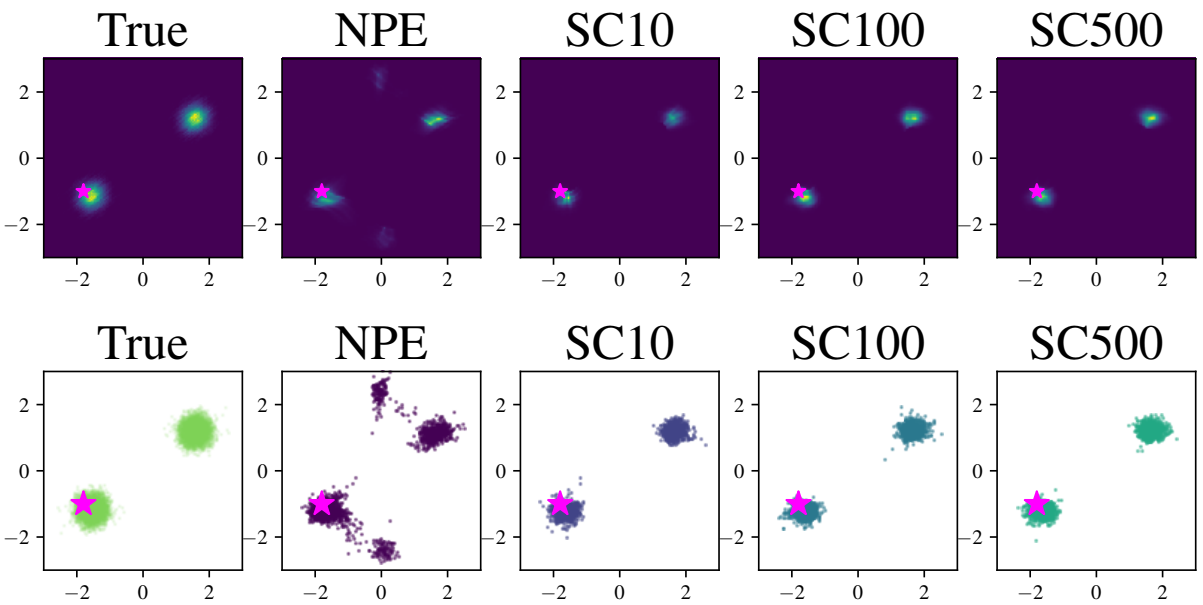
- Care needs to be taken when choosing $\tilde{p}(\theta)$!
- Increases computational cost by $\mathcal{O}(L)$, where L is the number of Monte Carlo samples from $\tilde{p}(\theta)$.

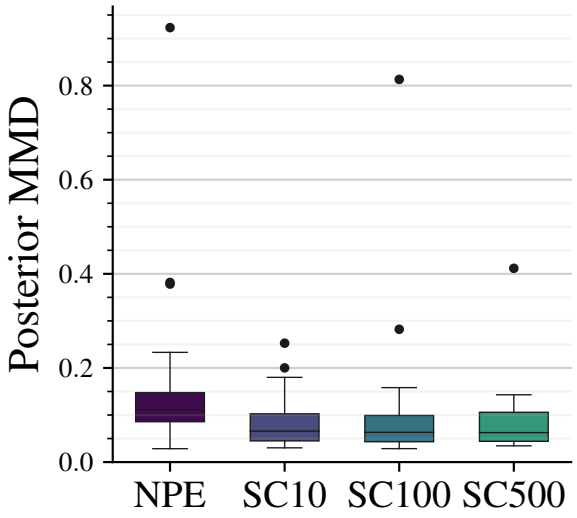
Results: Analytic Likelihood, Toy Model

- Toy model with a bimodal posterior:

$$\theta \sim \mathcal{N}(\theta \mid 0, \mathbf{I}), \qquad y \sim 0.5\,\mathcal{N}(y \mid \theta, \mathbf{I}/2) + 0.5\,\mathcal{N}(y \mid -\theta, \mathbf{I}/2)$$

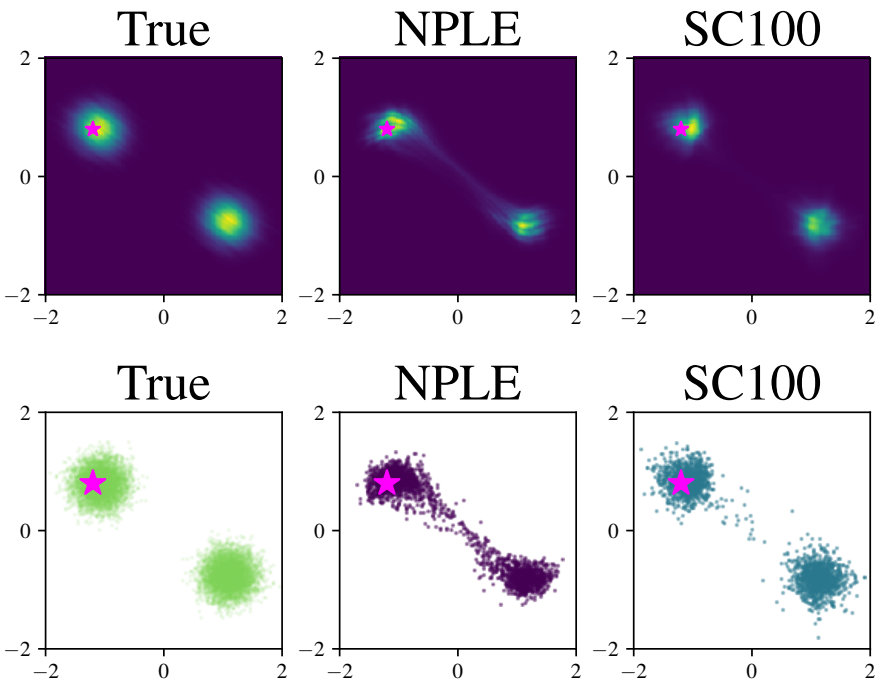
- Drastic improvements for small simulation budgets ($N = 1024$):

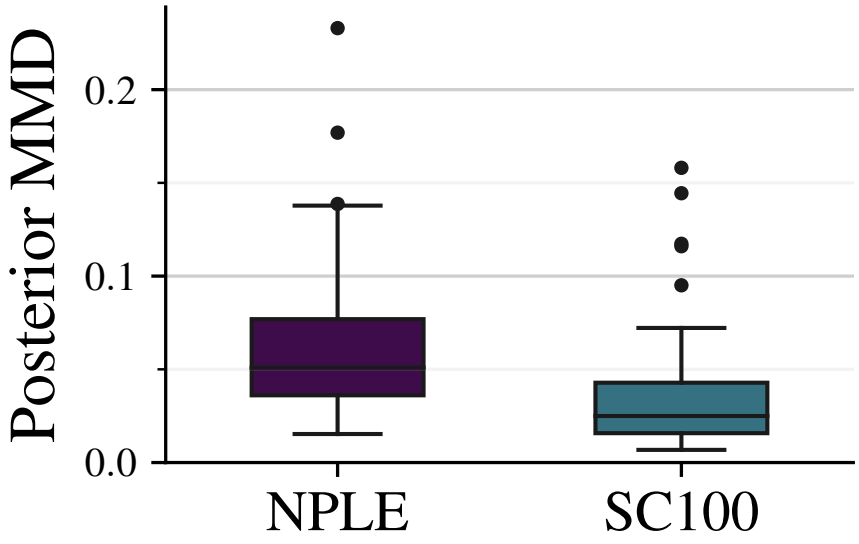




Results: Approximate Likelihood, Toy Model

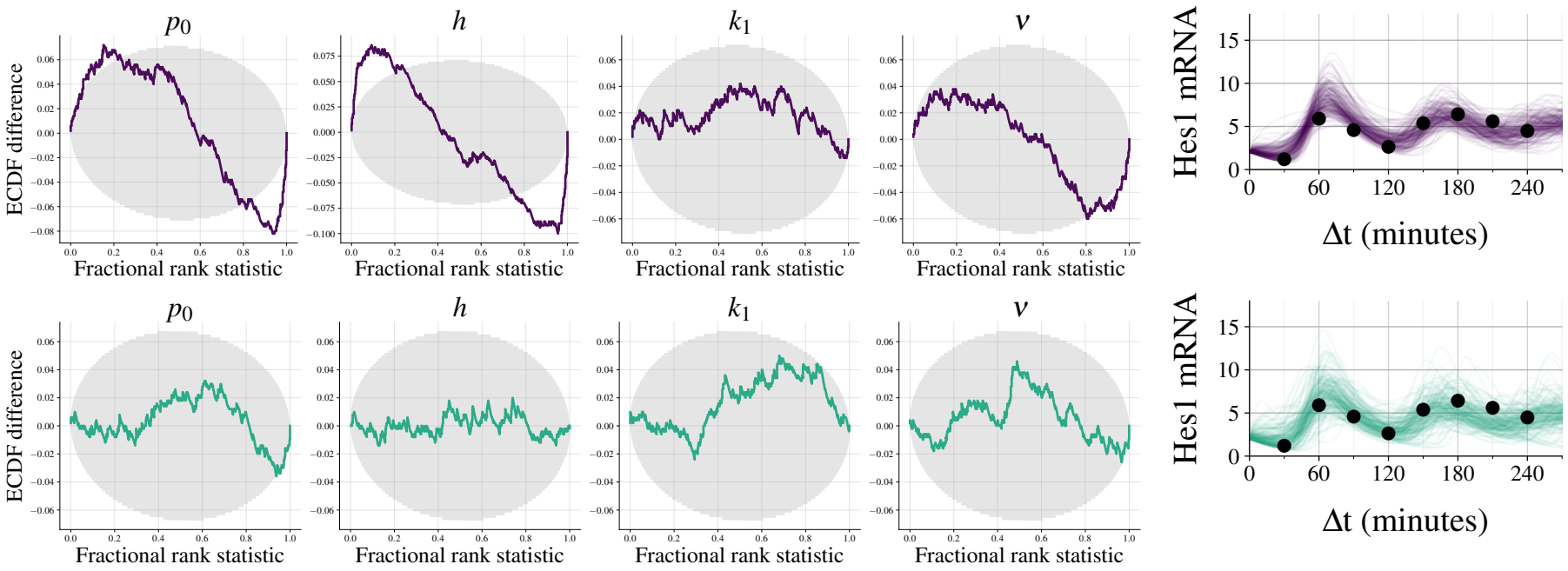
- Pretend that the likelihood were intractable and re-do experiment.
- Still notable improvements over straightforward NPLE:





Results: Approximate Likelihood, Representative Model

- An ODE model of Hes1 mRNA concentration after serum injection (Silk et al., 2011).
- No notable improvements in terms of posterior quality. However, better calibration when using the SC loss:



Agenda

Selected Prior Work

- Amortized posterior estimation (Radev et al., 2020; Gonçalves et al., 2020; Avecilla et al., 2022; Geffner et al., 2022; Sharrock et al., 2022)
- Neural likelihood estimation (Papamakarios et al., 2019; Lueckmann et al., 2019; Hermans et al., 2020; Munk et al., 2022).
- Learning end-to-end summary statistics (Chan et al., 2018; Chen et al., 2020; Radev et al., 2020)
- Importance sampling for inference correction and evidence estimation (Glöckler et al., 2022; Dax et al., 2023)
- Sequential joint learning (Wiqvist et al., 2021; Glöckler et al., 2022)
- Connections between generative and bottleneck models (Köthe, 2023)

Thank You!

Acknowledgements

- The `BayesFlow` team (https://bayesflow.org/), and specifically for the topics of this talk:
  - **Marvin Schmitt** (https://www.marvinschmitt.com/)
  - **Valentin Pratz** (https://valentinpratz.de/)
  - **Paul Bürkner** (https://paul-buerkner.github.io/)
  - **Ullrich Köthe** (https://hci.iwr.uni-heidelberg.de/vislearn/people/ullrich-koethe/)
  - **Daniel Habermann** (https://daniel-habermann.de/)
  - **Umberto Picchini** (https://umbertopicchini.github.io/)
- Much of this work was done during my time at the STRUCTURES Cluster of Excellence, Heidelberg University (https://www.thphys.uni-heidelberg.de/~structures/)

Agenda

Discussion

- How to best choose/schedule the self-consistency proposal $\tilde{p}(\theta)$ such that pre-asymptotic behavior is optimal?
- Can self-consistency losses be beneficial beyond small simulation budgets?
- Error analysis for marginal neural estimators. Can we port known results (e.g., Gelfand & Dey, 1994)?
- More expressive calibration diagnostics of the joint model (Modrák et al., 2022)? For instance, prior predictive and posterior predictive?

References I

Avecilla, G., Chuong, J. N., Li, F., Sherlock, G., Gresham, D., & Ram, Y. (2022). Neural networks enable efficient and accurate simulation-based inference of evolutionary parameters from adaptation dynamics. *PLoS Biology*, *20*(5), e3001633.

Chan, J., Perrone, V., Spence, J., Jenkins, P., Mathieson, S., & Song, Y. (2018). A likelihood-free inference framework for population genetic data using exchangeable neural networks. *Advances in neural information processing systems*, *31*.

Chen, Y., Zhang, D., Gutmann, M., Courville, A., & Zhu, Z. (2020). Neural approximate sufficient statistics for implicit models. *arXiv preprint arXiv:2010.10079*.

Dax, M., Green, S. R., Gair, J., Pürrer, M., Wildberger, J., Macke, J. H., ... Schölkopf, B. (2023). Neural importance sampling for rapid and reliable gravitational-wave inference. *Physical Review Letters*, *130*(17), 171403.

Geffner, T., Papamakarios, G., & Mnih, A. (2022). *Compositional score modeling for simulation-based inference.*

References II

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(3), 501–514.

Glöckler, M., Deistler, M., & Macke, J. H. (2022). Variational methods for simulation-based inference. *arXiv preprint arXiv:2203.04176*.

Gonçalves, P. J., Lueckmann, J.-M., Deistler, M., et al. (2020). Training deep neural density estimators to identify mechanistic models of neural dynamics. *Elife*.

Hermans, J., Begy, V., & Louppe, G. (2020). Likelihood-free mcmc with amortized approximate ratio estimators. In *International conference on machine learning* (pp. 4239–4248).

Köthe, U. (2023). A review of change of variable formulas for generative modeling. *arXiv preprint arXiv:2308.02652*.

Lueckmann, J.-M., Bassetto, G., Karaletsos, T., & Macke, J. H. (2019). Likelihood-free inference with emulator networks. In *Symposium on advances in approximate bayesian inference*.

References III

Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., & Macke, J. (2021). Benchmarking simulation-based inference. In *International conference on artificial intelligence and statistics* (pp. 343–351).

Modrák, M., Moon, A. H., Kim, S., Bürkner, P., Huurre, N., Faltejsková, K., ... Vehtari, A. (2022). Simulation-based calibration checking for bayesian computation: The choice of test quantities shapes sensitivity. *arXiv preprint arXiv:2211.02383*.

Munk, A., Zwartsenberg, B., Ścibior, A., Baydin, A. G. G., Stewart, A., Fernlund, G., ... Wood, F. (2022). Probabilistic surrogate networks for simulators with unbounded randomness. In *Uncertainty in artificial intelligence* (pp. 1423–1433).

Papamakarios, G., Sterratt, D., & Murray, I. (2019). Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows..

References IV

Radev, S. T., Graw, F., Chen, S., Mutters, N. T., Eichel, V. M., Bärnighausen, T., & Köthe, U. (2021). Outbreakflow: Model-based bayesian inference of disease outbreak dynamics with invertible neural networks and its application to the covid-19 pandemics in germany. *PLoS computational biology*, *17*(10), e1009472.

Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., & Köthe, U. (2020). Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE transactions on neural networks and learning systems*, *33*(4), 1452–1466.

Radev, S. T., Schmitt, M., Pratz, V., Picchini, U., Köthe, U., & Bürkner, P.-C. (2023). Jana: Jointly amortized neural approximation of complex bayesian models. *arXiv preprint arXiv:2302.09125*.

Ramesh, P., Lueckmann, J.-M., Boelts, J., Tejero-Cantero, Á., Greenberg, D. S., Gonçalves, P. J., & Macke, J. H. (2022). Gatsbi: Generative adversarial training for simulation-based inference. *arXiv preprint arXiv:2203.06481*.

References V

Schmitt, M., Bürkner, P.-C., Köthe, U., & Radev, S. T. (2021). Detecting model misspecification in amortized bayesian inference with neural networks. *arXiv preprint arXiv:2112.08866*.

Schmitt, M., Habermann, D., Bürkner, P.-C., Köthe, U., & Radev, S. T. (2023). Leveraging self-consistency for data-efficient amortized bayesian inference. *arXiv preprint arXiv:2310.04395*.

Sharrock, L., Simons, J., Liu, S., & Beaumont, M. (2022). *Sequential neural score estimation: Likelihood-free inference with conditional score based diffusion models.*

Silk, D., Kirk, P. D., Barnes, C. P., Toni, T., Rose, A., Moon, S., ... Stumpf, M. P. (2011). Designing attractive models via automated identification of chaotic and oscillatory dynamical regimes. *Nature Communications*, *2*(1). doi: 10.1038/ncomms1496

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2018). Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*.

References VI

Ward, D., Cannon, P., Beaumont, M., Fasiolo, M., & Schmon, S. (2022). Robust neural posterior estimation and statistical model criticism. *Advances in Neural Information Processing Systems*, *35*, 33845–33859.

Wiqvist, S., Frellsen, J., & Picchini, U. (2021). Sequential neural posterior and likelihood approximation. *arXiv preprint arXiv:2102.06522*.