Learn Bayes Methods Week: Introduction to BayesFlow

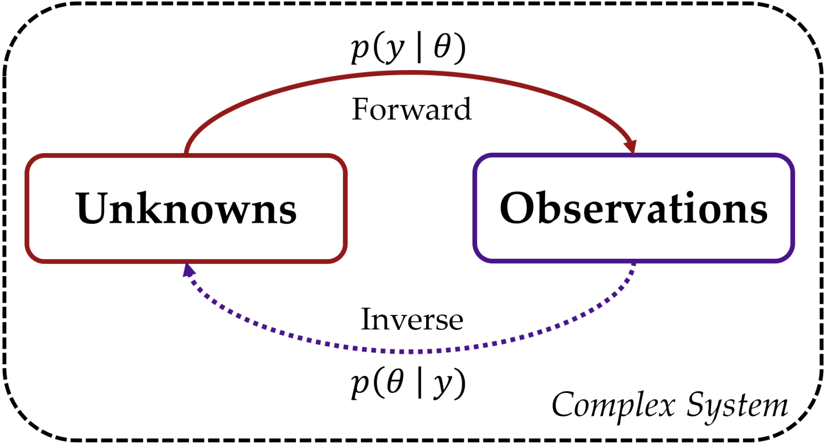
Stefan T. Radev

Center for Modeling, Simulation, & Imaging in Medicine (CeMSIM)
Department of Cognitive Science
Rensselaer Polytechnic Institute

May 12, 2025

Agenda

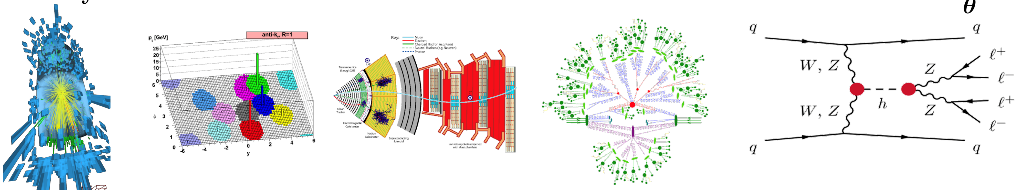
Problem Setting



Example: Particle Physics

forward: simulation

Observables **y** ← Detector interactions ← Shower splittings ← Parton-level momenta ← Theory parameters $\boldsymbol{\theta}$



inverse: inference

Figure adapted from Kyle Cranmer's PhyStat-SBI 2024 talk.

Example: Agent-Based Models for Immersive Environments

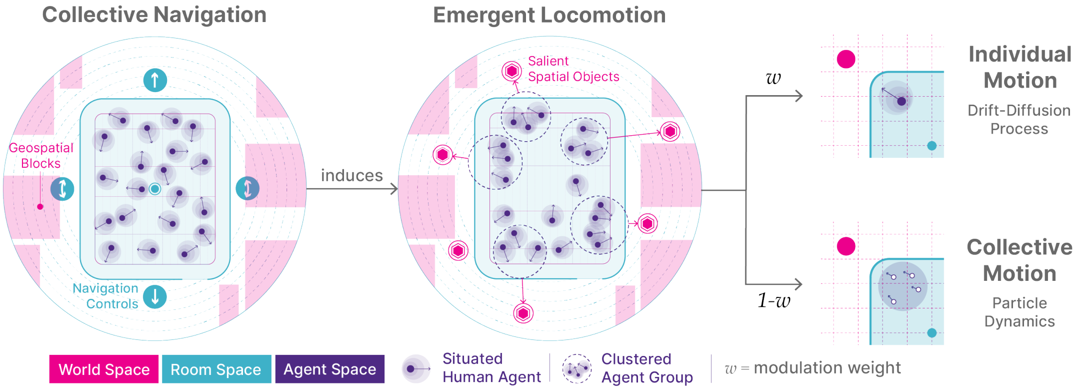
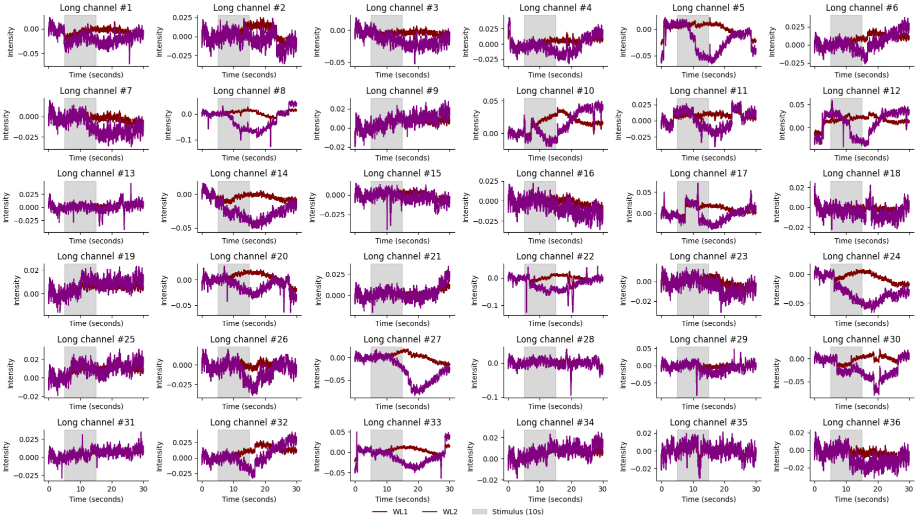


Figure from Huang et al. (*in prep*).

Example: Full-Head fNIRS Simulator



Probabilistic Recipes

- Generative (forward) notation indicates a probabilistic recipe:

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}) \tag{1}$$

$$\mathbf{y} \sim p(\mathbf{y} \mid \boldsymbol{\theta}) \tag{2}$$

  Note: Overloaded semantics of the $\sim$ symbol, means "distributed as" and "sampled from".

- Generative notation mimics probabilistic programs (e.g., Stan):

```
model {
    theta ~ normal(0, 1);
    y ~ normal(theta, 1);
}
```

Simulators vs. Likelihoods

Likelihood-Based (Explicit)

Bayesian model $p(\boldsymbol{\theta}, \mathbf{y})$:

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$$
$$\mathbf{y} \sim p(\mathbf{y} \mid \boldsymbol{\theta})$$

- Prior $p(\boldsymbol{\theta})$ can be sampled and evaluated.
- Data model $p(\mathbf{y} \mid \boldsymbol{\theta})$ can be sampled and evaluated.

Simulation-Based (Implicit)

The same Bayesian model $p(\boldsymbol{\theta}, \mathbf{y})$:

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$$
$$\mathbf{y} = \mathrm{Sim}(\boldsymbol{\theta}, \mathbf{z}), \ \mathbf{z} \sim p(\mathbf{z} \mid \boldsymbol{\theta})$$

- Prior $p(\boldsymbol{\theta})$ can be sampled and **optionally** evaluated.
- Data model $p(\mathbf{y} \mid \boldsymbol{\theta})$ can be sampled **but not** evaluated.

The Hardships of Being Bayesian

- Bayesian inference is notationally straightforward:

$$p(\boldsymbol{\theta} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{3}$$

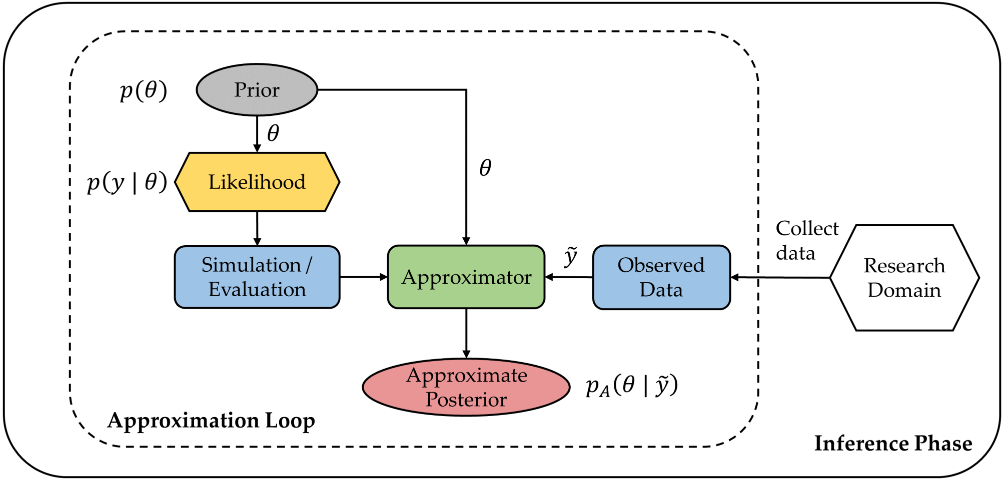
- "Bayesian inference is hard because integration is hard":

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \times \left(\int_{\Theta} p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \right)^{-1} \tag{4}$$

- Simulation-based inference is harder because integration is twice as hard:

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \int_{\mathcal{Z}} p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) \mathrm{d}\mathbf{z} \times p(\boldsymbol{\theta}) \times \left(\int_{\Theta} p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \right)^{-1} \tag{5}$$

Non-Amortized Bayesian Inference

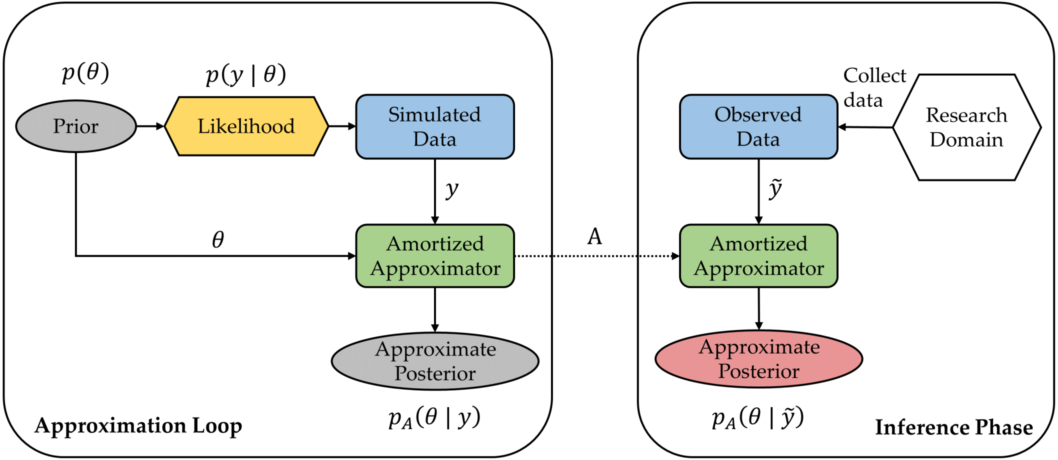


Bayesians Now and Then

Solves complex integrals; finds conjugate models and compares them analytically; proves new theorems in probability theory; computes Bayes factors by hand.

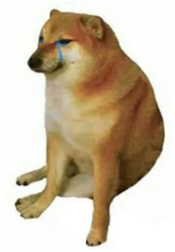
Sampler did not converge. Help.

Amortized Bayesian Inference



Bayesians Now

Bayesians Now



Sampler did not converge. Help.

Also Bayesians Now

Network did not converge. Help.

Amortized Bayesian Inference in a Nutshell

1. Specify a parametric Bayesian model $p(\boldsymbol{\theta}, \mathbf{y})$.
2. Generate a data set of model simulations $\mathcal{D}_{\text{train}} := \{\boldsymbol{\theta}^{(b)}, \mathbf{y}^{(b)}\}_{b=1,\ldots,B}$.
3. Specify a conditional generative network $\boldsymbol{\theta} \sim G(\boldsymbol{\theta}; \mathbf{y})$ with proper inductive biases.
4. Train $G(\boldsymbol{\theta}; \mathbf{y})$ on $\mathcal{D}_{\text{train}}$ until convergence.
5. Diagnose the statistical accuracy of G on a test set $\mathcal{D}_{\text{test}} := \{\boldsymbol{\theta}^{(s)}, \mathbf{y}^{(s)}\}_{s=1,\ldots,S}$.
6. Apply $G(\boldsymbol{\theta}; \mathbf{y})$ to arbitrary many observed data sets $\mathcal{D}_{\text{obs}} := \{\boldsymbol{\theta}^{(l)}, \mathbf{y}^{(l)}\}_{l=1,\ldots,L}$.
7. Detect potential simulation gaps and diagnose sample quality.
8. Perform your favorite posterior predictive (PPC) checks.

An Informal Working Definition

Clarifying Amortized Bayesian Inference (ABI)

The term amortized has been used inconsistently throughout the literature, often denoting different generalization scopes.

Definition

Let $\mathcal{A}$ denote a learner, $\mathbf{Y} \in \mathbb{R}^D$ denote target variables, $\mathbf{X} \in \mathbb{X}$ represent input data, and $\mathbf{C} \in \mathbb{C}$ denote context variables. A learner $\mathbf{Y} \sim \mathcal{A}(\mathbf{X}, \mathbf{C})$ is an *amortized Bayesian approximator* of a target quantity $\mathbf{Y}$ with respect to a joint distribution $p(\mathbf{X}, \mathbf{Y}, \mathbf{C})$ if it can directly approximate $p(\mathbf{Y} \mid \mathbf{X}, \mathbf{C})$ for any $(\mathbf{X}, \mathbf{C}) \sim p(\mathbf{X}, \mathbf{C})$ without further training or additional approximation algorithms.

Inference as Optimization

- A straightforward objective for amortizing inference (forward KL):

$$q^* = \arg\min_q \mathbb{E}_{p(\mathbf{y})} \left[\mathbb{KL}(p(\boldsymbol{\theta} \mid \mathbf{y}) \,||\, q(\boldsymbol{\theta} \mid \mathbf{y}))\right] \quad (6)$$

$$= \arg\min_q \mathbb{E}_{p(\mathbf{y})} \left[\mathbb{E}_{p(\boldsymbol{\theta} \mid \mathbf{y})} \left[\log p(\boldsymbol{\theta} \mid \mathbf{y}) - \log q(\boldsymbol{\theta} \mid \mathbf{y})\right]\right] \quad (7)$$

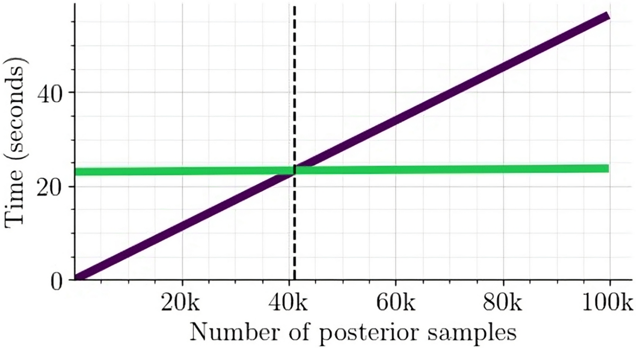
$$= \arg\min_q \mathbb{E}_{p(\boldsymbol{\theta}, \mathbf{y})} \left[- \log q(\boldsymbol{\theta} \mid \mathbf{y})\right] \quad (8)$$

- In practice, we minimize the empirical mean of Eq.8:

$$\hat{q} = \arg\min_q \frac{1}{B} \sum_{b=1}^{B} - \log q(\boldsymbol{\theta}^{(b)} \mid \mathbf{y}^{(b)}) \quad (9)$$

- Can be generalized to any (proper) scoring rule (e.g., Pacchiardi, Khoo, & Dutta, 2024).

Breakeven Points



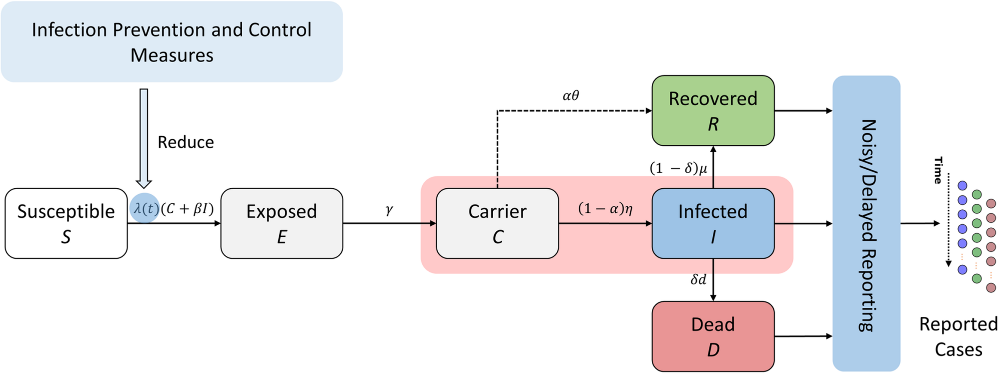
MCMC:

- no upfront training
- 5.66 seconds for 10k posterior draws

Amortized inference:

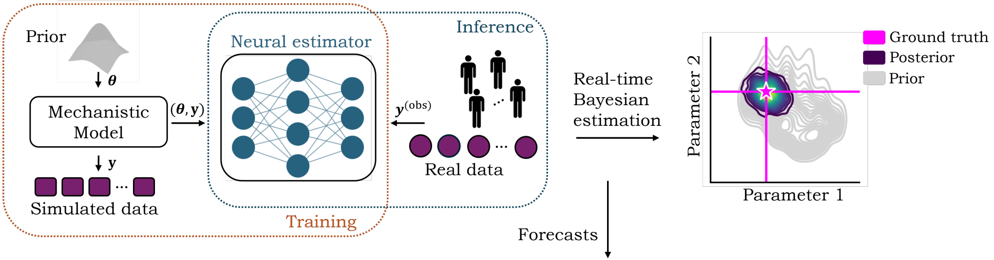
- 23 seconds upfront training
- 0.07 seconds for 10k posterior draws

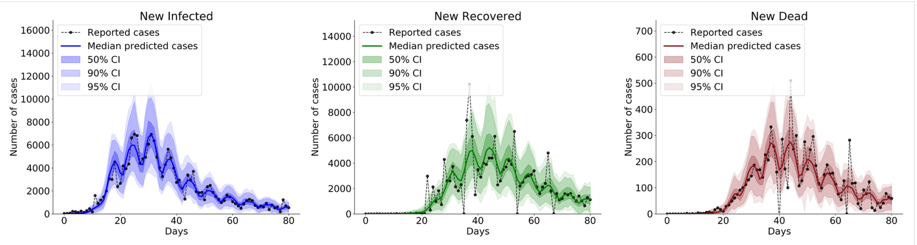
Example: Epidemiological Modeling



Simplified compartmental model from Radev et al. (2021).

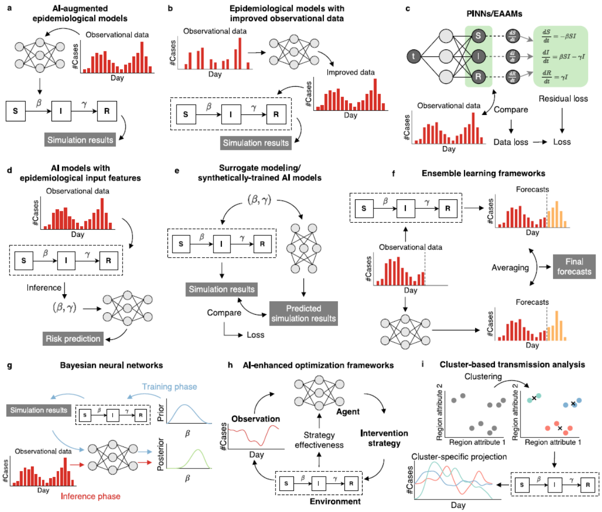
Example: Epidemiological Modeling (OutbreakFlow)





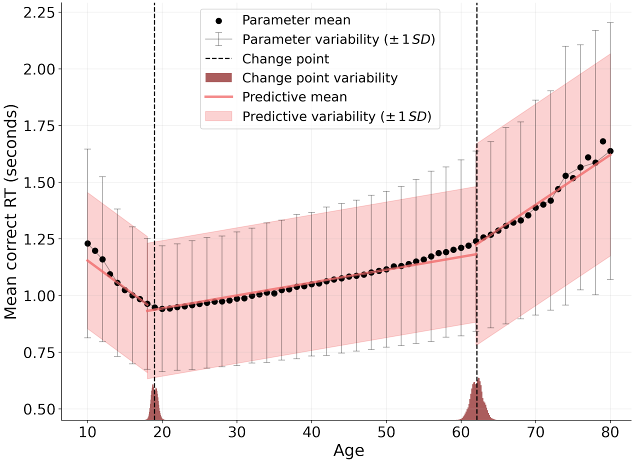
It’s Just the Beginning

- There is a whole zoo of hybrid methods and hybrid models (Figure from Ye et al., 2025).
- The line between statistical models and statistical methods becomes increasingly blurry (Bürkner, Scholz, & Radev, 2023).
- Amortized Bayesian inference can be viewed as part of the new simulation intelligence wave (Lavin et al., 2021).



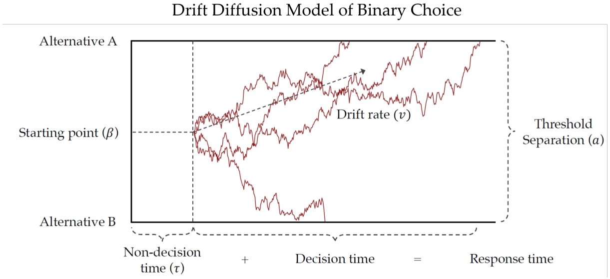
Example: Decision Making (I)

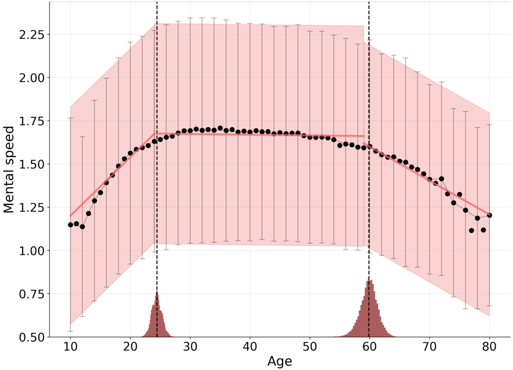
- **Assumption**: Mental speed in simple cognitive tasks declines over time.



Example: Decision Making (II)

- Challenge: Bayesian model-based analysis of 1.2 million participants (von Krause, Radev, & Voss, 2022):





Example: Decision Making (III)

- How the media read the paper...

Brains do not slow down until after age of 60, study finds

Findings go against the assumption that mental processing speed declines from a peak at age 20

NEWSLETTERS
Sign up to read our regular email newsletters

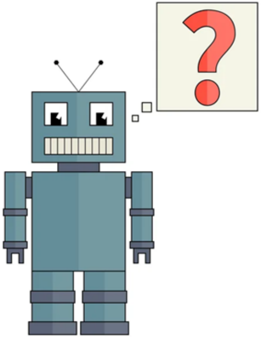
NewScientist

News Podcasts Video Technology Space Physics Health More Shop Courses Events

Your brain doesn't slow down until your 60s – later than we thought

Although people take longer to make decisions from age 20 onwards, this may not be due to a decline in the speed of information processing, a large study has found

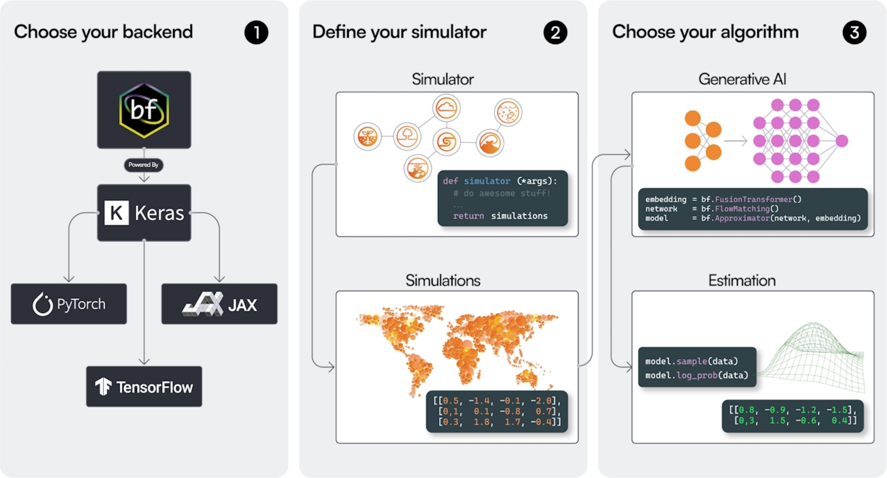
What Do We Stand to Gain?



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

Agenda

Overview of BayesFlow

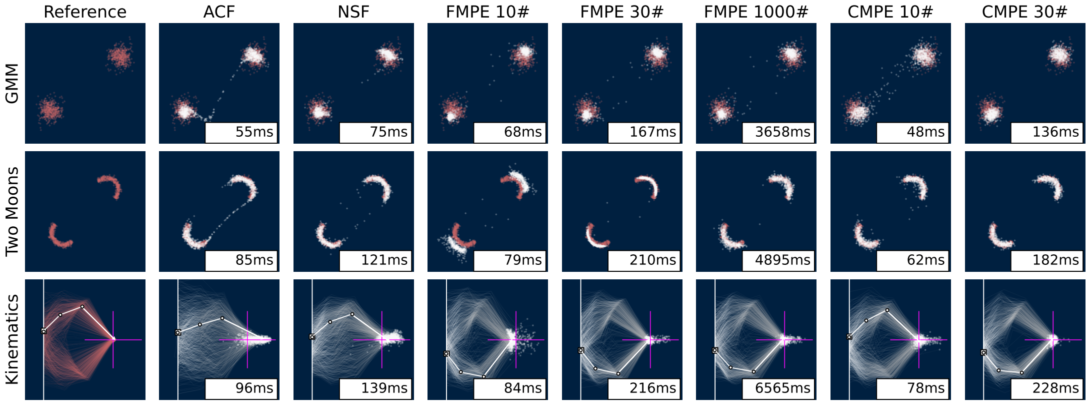


Important Links

- GitHub page: https://github.com/bayesflow-org/bayesflow
- Awesome amortized list: https://github.com/bayesflow-org/awesome-amortized-inference
- Project page: https://bayesflow.org/
- Forums: discuss.bayesflow.org
- BlueSky: https://bsky.app/profile/bayesflow.org
- Zulip: https://bayesflow.zulipchat.com/

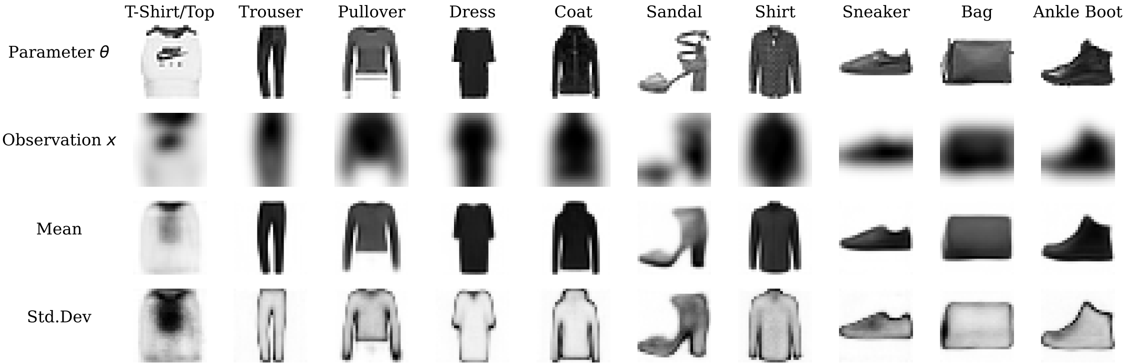
Agenda

Notes on Benchmarking (I)



A subset of the benchmark problems from Lueckmann et al. (2021); Kruse et al. (2021).

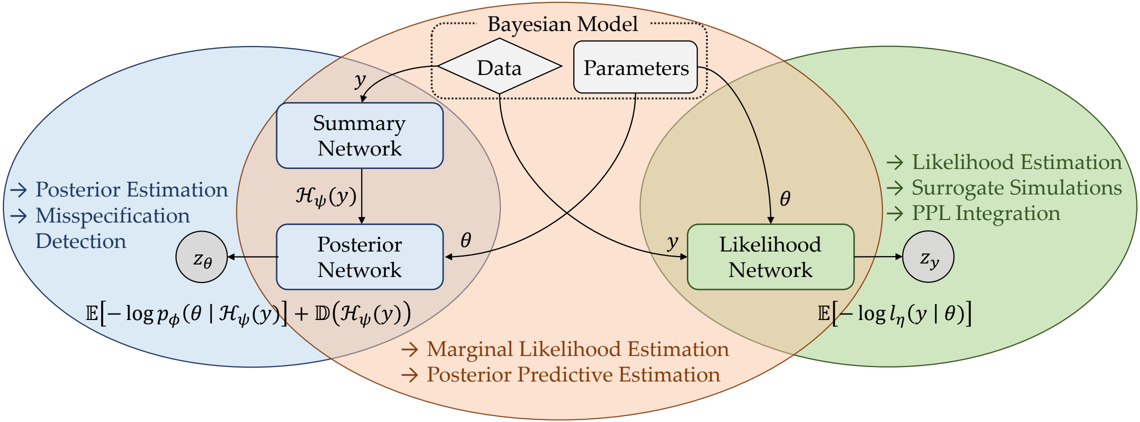
Notes on Benchmarking (II)



Consistency models can easily sample from high-dimensional posteriors Schmitt, Pratz, et al. (2024).

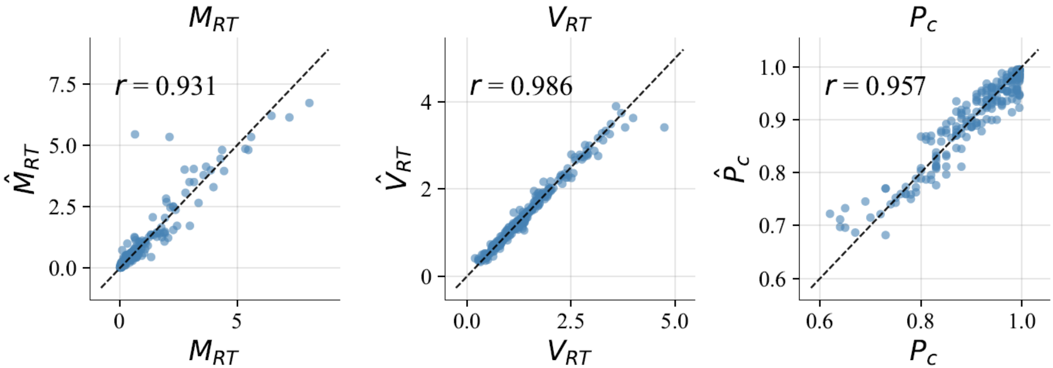
Joint Learning: General Overview

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

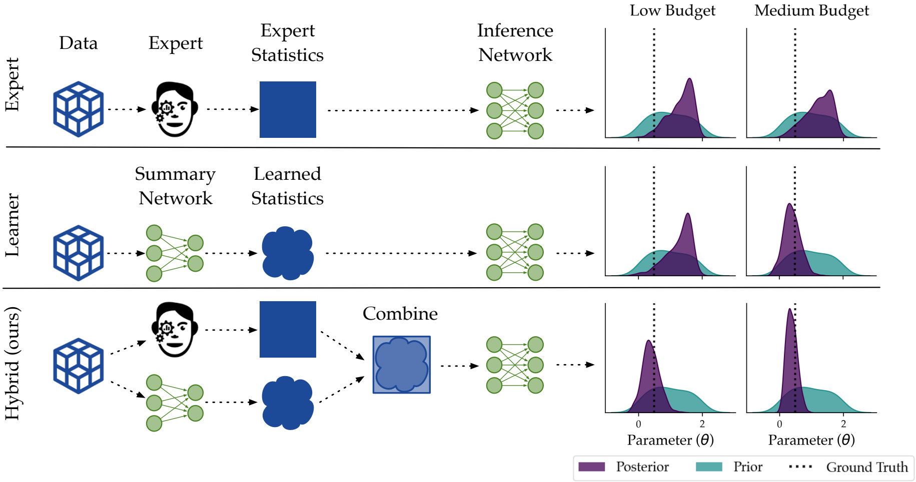


Summary Networks Approximate Sufficient Summaries

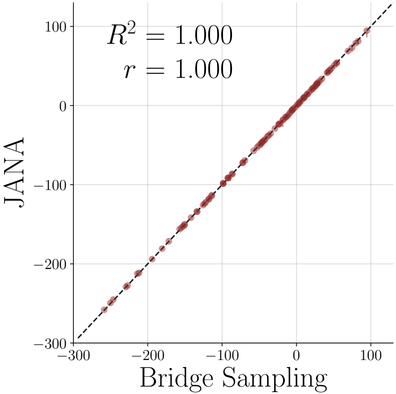
- Random forest regression of summary space variables on known sufficient summaries (Wu, Radev, & Tuerlinckx, 2024):



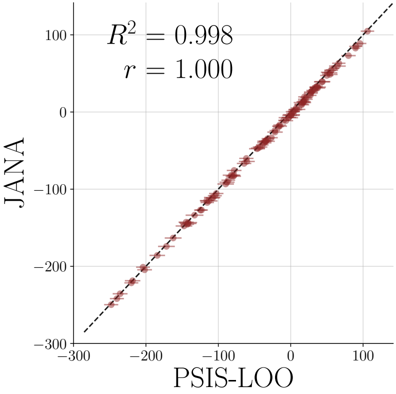
Expert vs. End-To-End Statistics



Joint Learning Can Approximate Hard Bayesian Problems



(a) Prior predictive (LML)



(b) Post. predictive (ELPD)

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:

$$q = \int_{\mathcal{Y}} \int_{\Theta} \mathbb{I}\left[\theta^* \in U_q(\theta \mid \mathbf{y})\right] p(\theta^*, \mathbf{y})\, d\theta^* d\mathbf{y} \tag{10}$$

- Idea: Approximate SBC via many simulations from $p(\theta^*, \mathbf{y})$ (or a surrogate) and (fractional) rank statistics of the posterior samples $\{\theta^{(s)}\}$:

$$R(\theta^*_m, \theta_{1:S}) = \sum_{s=1}^{S} \mathbb{I}\left[\theta^*_m > \theta_s\right] \tag{11}$$

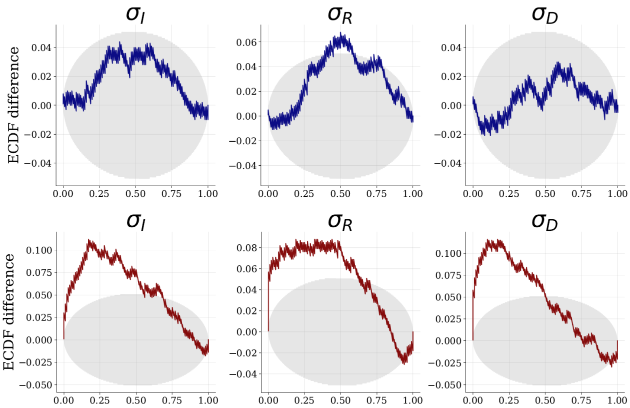
- SBC can be performed for free thanks to amortized inference!

Simulation-Based Calibration in Action

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

Posterior calibration

Joint calibration



One Analysis Must Not Rule Them All

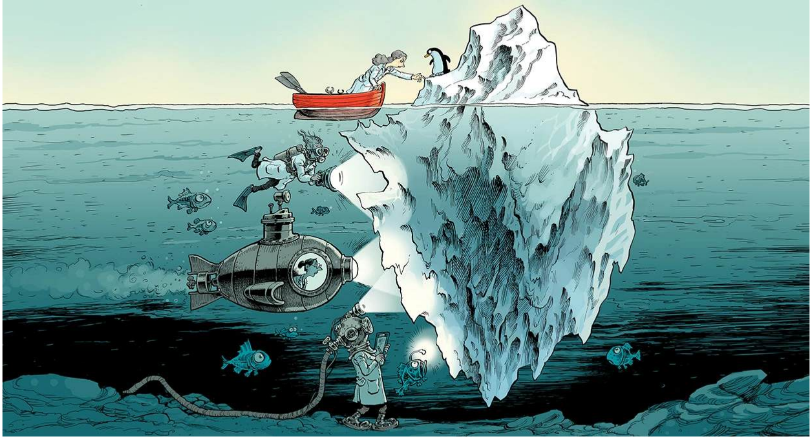


Illustration by David Parkins, reprinted in Wagenmakers et al. (2022).

Meta-Amortized Inference (I)

- In regular amortized inference, we typically minimize a strictly proper scoring loss $\mathcal{S}$ in expectation over the Bayesian model $p(\boldsymbol{\theta}, \mathbf{y})$:

$$\min_q \left\{ \mathbb{E}_{p(\boldsymbol{\theta},\mathbf{y})} \left[\mathcal{S}(q(\cdot \mid \mathbf{y}), \boldsymbol{\theta}) \right] \approx \frac{1}{B} \sum_{b=1}^{B} \mathcal{S}(q(\cdot \mid \mathbf{y}^{(b)}), \boldsymbol{\theta}^{(b)}) \right\}. \tag{12}$$

- If we want to extend the amortization scope, we can generalize over any number of context variables $\mathbf{C}$ by adding additional conditions:

$$\min_q \mathbb{E}_{p(\boldsymbol{\theta},\mathbf{y},\mathbf{C})} \left[\mathcal{S}(q(\cdot \mid \mathbf{y}, \mathbf{C}), \boldsymbol{\theta}) \right] \tag{13}$$

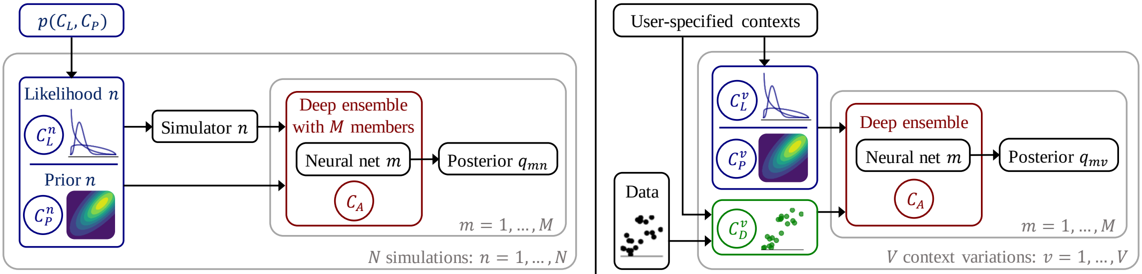
Meta-Amortized Inference (II)

Stage 1: Training

Extending the amortization scope

Stage 2: Inference

Amortized sensitivity analysis

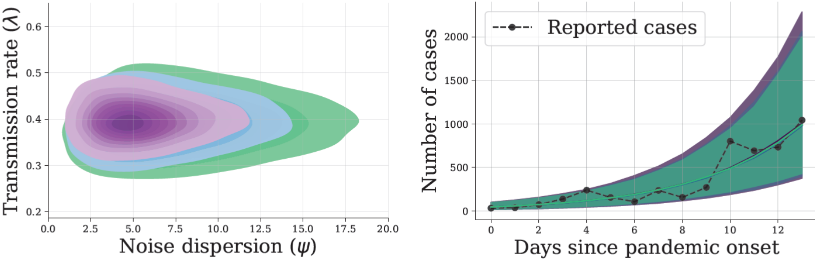


→ learn $N \cdot M$ posteriors during training

→ efficiently evaluate $V \cdot M$ posteriors during inference

Expanding the amortization scope via context awareness (Elsemüller et al., 2024).

Easy Prior Sensitivity Analysis

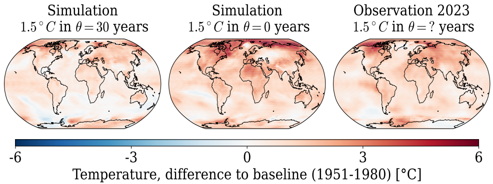


(a) Bivariate posteriors. (b) Posterior predictives.

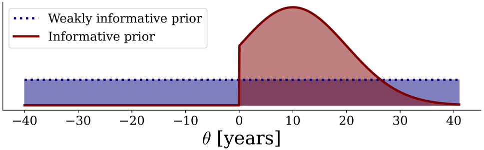
$\gamma = 0.5$ $\gamma = 1.0$ $\gamma = 2.0$

Re-analyzing over a tilted $p(\boldsymbol{\theta})^{\gamma}$ prior (Elsemüller et al., 2024).

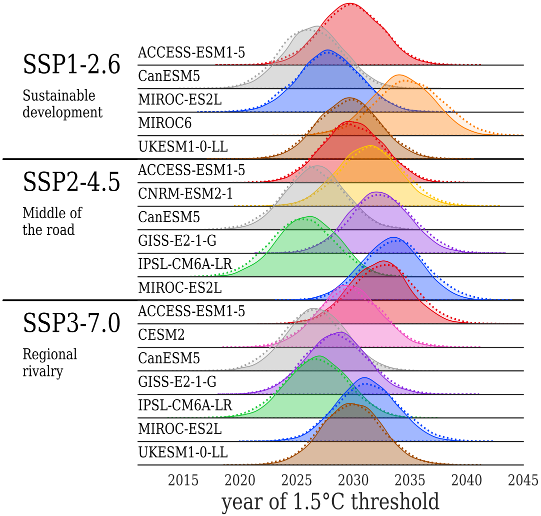
Sensitivity-Aware Inference of Critical Thresholds



Simulation snapshots (**left**, **center**) and the empirical observation (**right**).

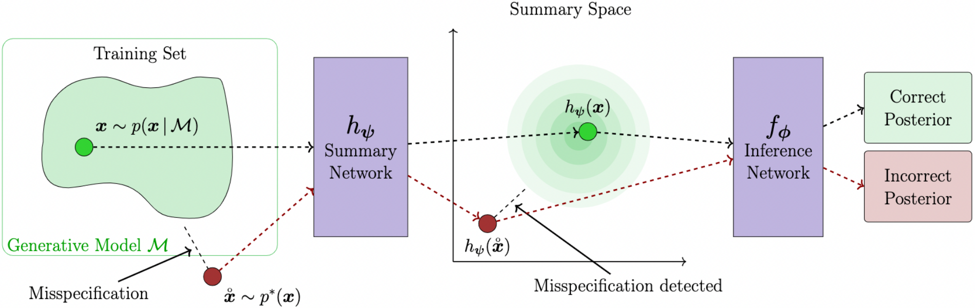


Qualitatively different prior distributions.



Out-of-Simulation (OOSim) Detection

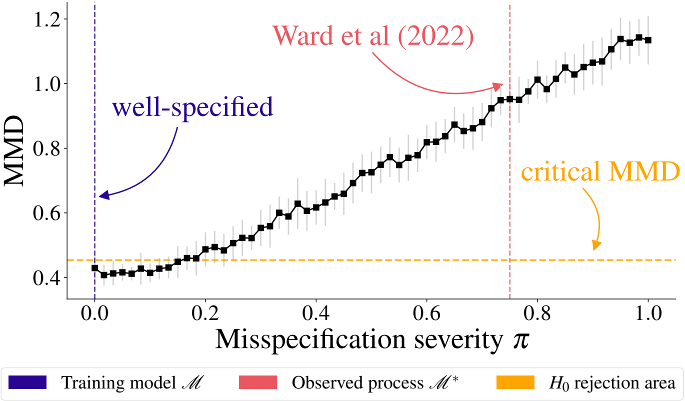
- Re-frame model misspecification as out-of-distribution (OOD) detection (Schmitt et al., 2021).
- Trust inferences only on inlier observations:



- See Siahkoohi, Rizzuti, Orozco, and Herrmann (2023) for a physics-based latent space correction.

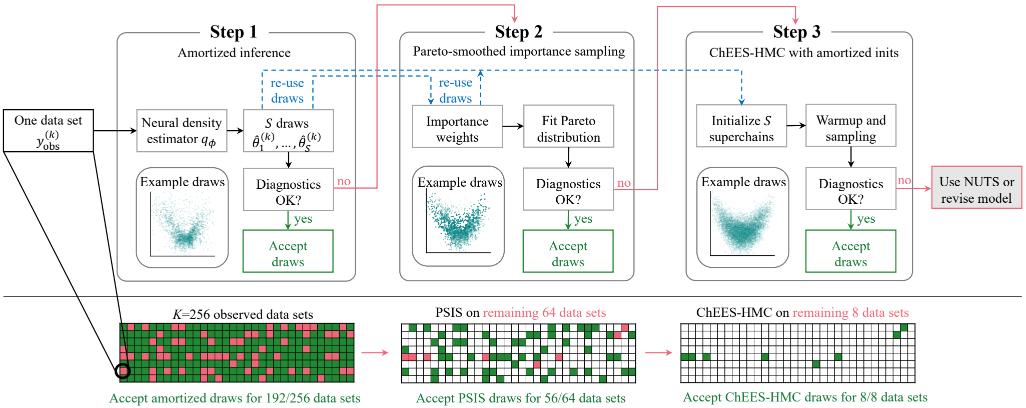
OOSim in Action

- MMD can reliably highlight simulation gaps (Schmitt et al., 2021; Schmitt, Bürkner, et al., 2024):

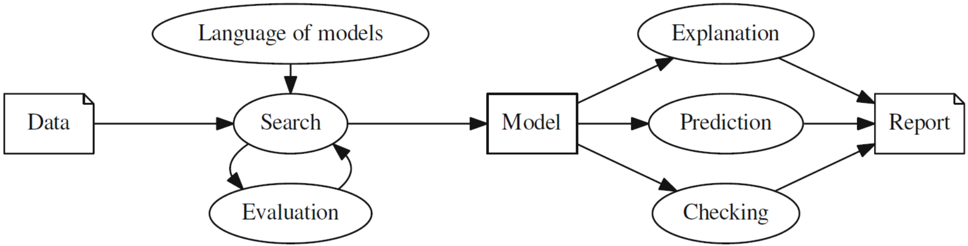


From Bayesian Analysis to Bayesian Workflows

- Towards an automated amortized Bayesian workflow (Schmitt, Li, et al., 2024):



Memo: The Automatic Statistician



A simplified flow diagram outlining the operations of an envisaged report-writing automatic statistician (Steinruecken et al., 2019). As of 2025, the project seems to have died.

Agenda

Discussion

- Low simulation budgets coupled with very high-dimensional data remain challenging. Multi-fidelity simulation-based inference as a potential remedy (Krouglova, Johnson, Confavreux, Deistler, & Gonçalves, 2025)?
- Prior specification requires substantive domain knowledge and remains a challenge.
- Many hyperparameters to tune when things to wrong. Who is to blame: the approximator or the simulation model?
- Unclear how to best augment simulation-based training with analytic information (e.g., likelihood, physical constraints).
- Generalized Bayes vs. Bayes: How much model misspecification is tolerable and when do we want to deviate from $p(\boldsymbol{\theta} \mid \mathbf{y})$?

Thank You!

References I

Bürkner, P.-C., Scholz, M., & Radev, S. T. (2023). Some models are useful, but how do we know which ones? towards a unified bayesian model taxonomy. *Statistic Surveys*, *17*, 216–310.

Elsemüller, L., Olischläger, H., Schmitt, M., Bürkner, P.-C., Koethe, U., & Radev, S. T. (2024). Sensitivity-aware amortized bayesian inference. *Transactions on Machine Learning Research*.

Krouglova, A. N., Johnson, H. R., Confavreux, B., Deistler, M., & Gonçalves, P. J. (2025). Multifidelity simulation-based inference for computationally expensive simulators. *arXiv preprint arXiv:2502.08416*.

Kruse, J., Ardizzone, L., Rother, C., & Köthe, U. (2021). Benchmarking invertible architectures on inverse problems. *arXiv preprint arXiv:2101.10763*.

Lavin, A., Krakauer, D., Zenil, H., Gottschlich, J., Mattson, T., Brehmer, J., ... others (2021). Simulation intelligence: Towards a new generation of scientific methods. *arXiv preprint arXiv:2112.03235*.

References II

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).

Pacchiardi, L., Khoo, S., & Dutta, R. (2024). Generalized bayesian likelihood-free inference. *Electronic Journal of Statistics*, *18*(2), 3628–3686.

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., 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*.

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*.

References III

Schmitt, M., Bürkner, P.-C., Köthe, U., & Radev, S. T. (2024). Detecting model misspecification in amortized bayesian inference with neural networks: An extended investigation. *arXiv preprint arXiv:2406.03154*.

Schmitt, M., Li, C., Vehtari, A., Acerbi, L., Bürkner, P.-C., & Radev, S. T. (2024). Amortized bayesian workflow. *arXiv preprint arXiv:2409.04332*.

Schmitt, M., Pratz, V., Köthe, U., Bürkner, P.-C., & Radev, S. (2024). Consistency models for scalable and fast simulation-based inference. *Advances in Neural Information Processing Systems*, *37*, 126908–126945.

Siahkoohi, A., Rizzuti, G., Orozco, R., & Herrmann, F. J. (2023). Reliable amortized variational inference with physics-based latent distribution correction. *Geophysics*, *88*(3), R297–R322.

Steinruecken, C., Smith, E., Janz, D., Lloyd, J., & Ghahramani, Z. (2019). The automatic statistician. *Automated machine learning: Methods, systems, challenges*, 161–173.

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 IV

von Krause, M., Radev, S. T., & Voss, A. (2022). Mental speed is high until age 60 as revealed by analysis of over a million participants. *Nature Human Behaviour*, *6*(5), 700–708.

Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, *605*(7910), 423–425.

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

Wu, Y., Radev, S., & Tuerlinckx, F. (2024). Testing and improving the robustness of amortized bayesian inference for cognitive models. *arXiv preprint arXiv:2412.20586*.

Ye, Y., Pandey, A., Bawden, C., Sumsuzzman, D. M., Rajput, R., Shoukat, A., ... Galvani, A. P. (2025). Integrating artificial intelligence with mechanistic epidemiological modeling: a scoping review of opportunities and challenges. *Nature Communications*, *16*(1), 1–18.