

The frontiers of simulation-based inference (Part I)

PHYSTAT-SBI 2024

May 15, 2024

Gilles Louppe

g.louppe@uliege.be

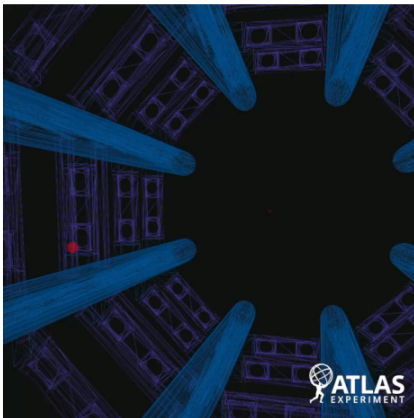
Simulation-based inference

Simulators as generative models

A simulator prescribes a generative model that can be used to simulate data \mathbf{x} .

Collider data

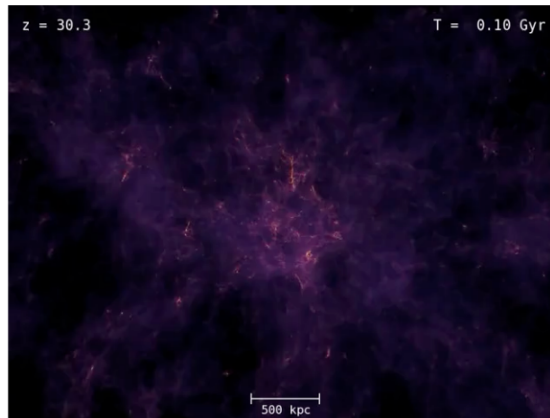
particles $\sim p(\text{particles})$



[C. Cesarotti with ATLAS]

Cosmology data

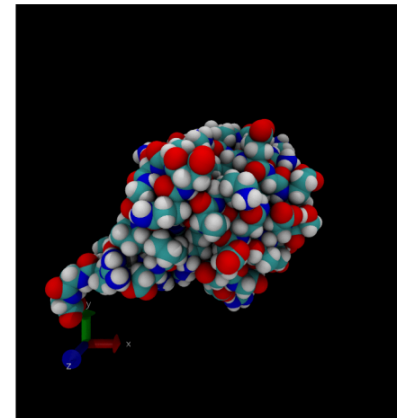
particles $\sim p(\text{particles})$



[Aquarius simulation]

Molecular dynamics

configurations $\sim p(\text{configurations})$



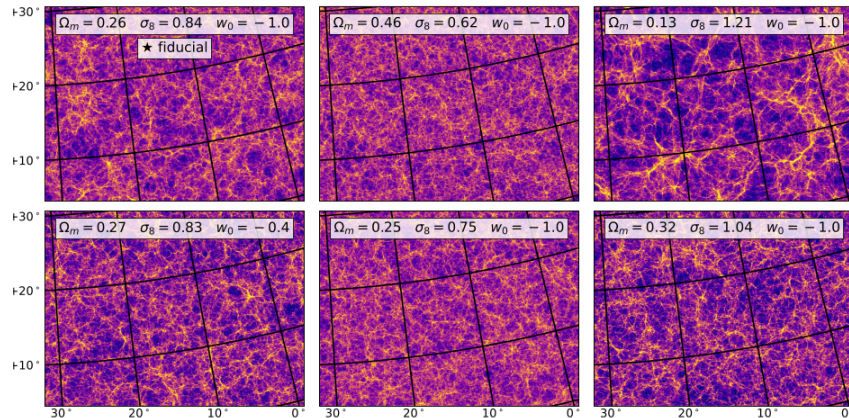
[E. Cances et al]

Conditional simulators

A conditional simulator prescribes a way to sample from the likelihood $p(\mathbf{x}|\theta)$, where θ is a set of conditioning variables or parameters.

Cosmology data

$$\text{map} \sim p(\text{map} \mid \{\Omega_m, \sigma_8, w_0\})$$



[Kacprzak et al 2022]

$$x \sim p(x; \mathcal{M})$$

Model

or

$$x \sim p(x \mid \theta)$$

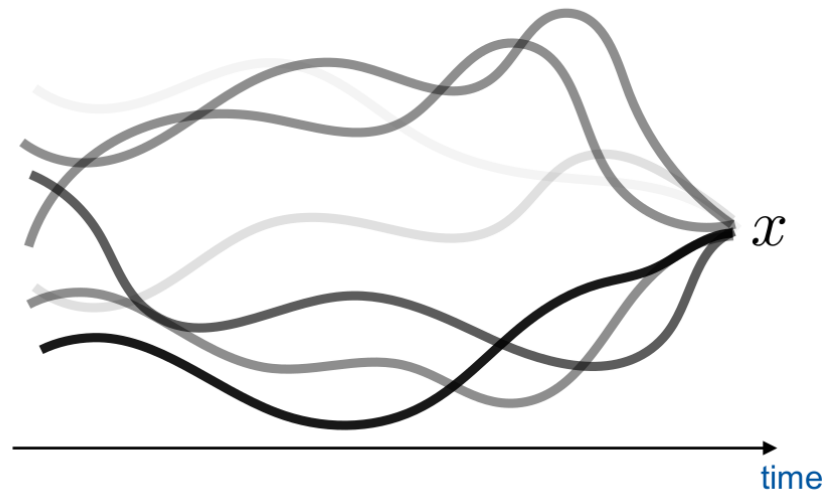
Model parameters

Intractable likelihoods

The (modeled) data generating process may involve additional latent variables \mathbf{z} that are not observed, leading to likelihoods

$$p(\mathbf{x}|\theta) = \int p(\mathbf{x}, \mathbf{z}|\theta) d\mathbf{z}.$$

In this case, evaluating the likelihood becomes intractable.



$$p(\mathbf{z}_p | \theta)$$

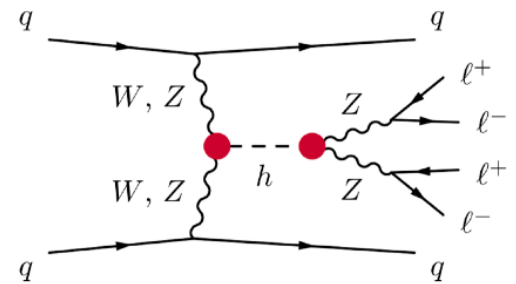
Latent variables

Parameters
of interest

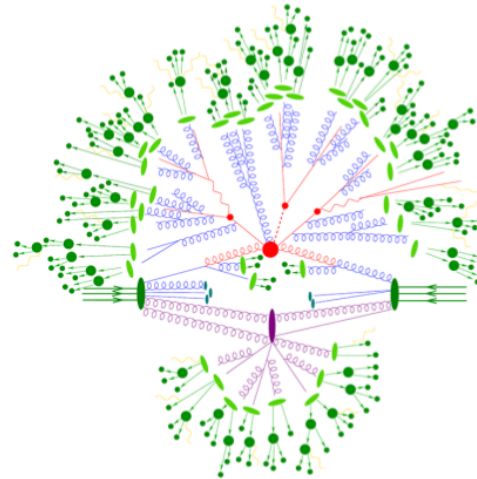
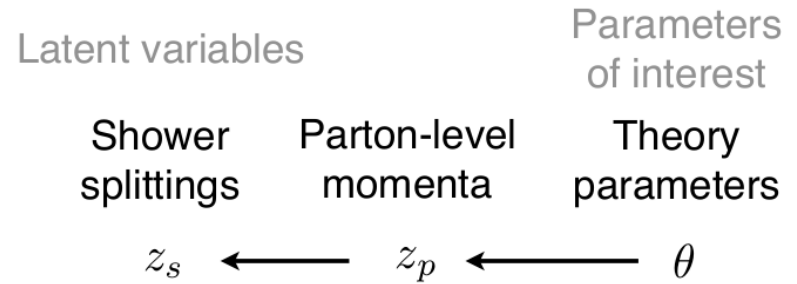
Parton-level
momenta

Theory
parameters

$$z_p \longleftarrow \theta$$



$$p(\mathbf{z}_s | \theta) = \int p(\mathbf{z}_p | \theta) p(\mathbf{z}_s | \mathbf{z}_p) d\mathbf{z}_p$$



$$p(\mathbf{z}_d|\theta) = \iint p(\mathbf{z}_p|\theta)p(\mathbf{z}_s|\mathbf{z}_p)p(\mathbf{z}_d|\mathbf{z}_s)d\mathbf{z}_pd\mathbf{z}_s$$

Latent variables

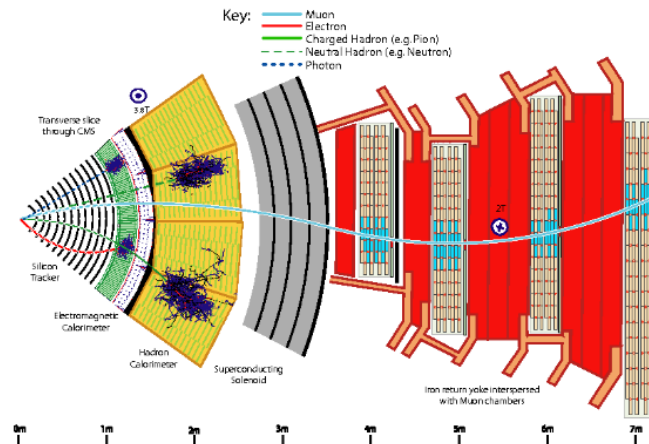
Parameters of interest

Detector interactions

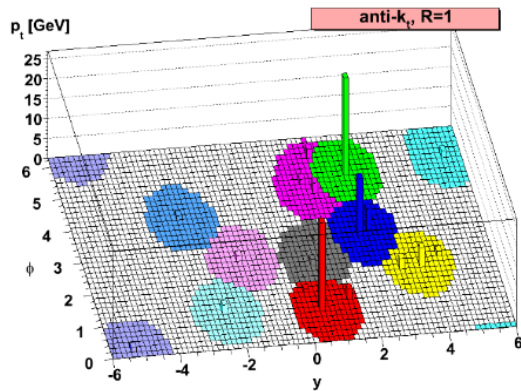
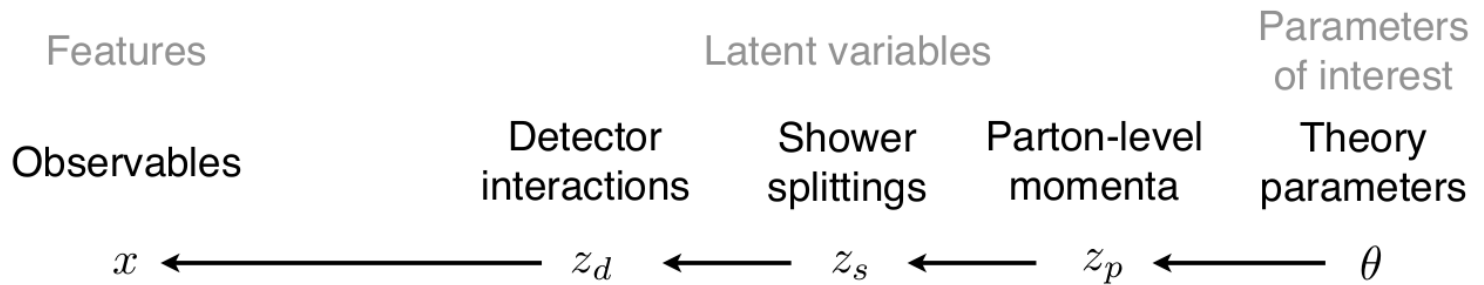
Shower splittings

Parton-level momenta

Theory parameters



$$p(\mathbf{x}|\theta) = \underbrace{\iiint}_{\text{yikes!}} p(\mathbf{z}_p|\theta)p(\mathbf{z}_s|\mathbf{z}_p)p(\mathbf{z}_d|\mathbf{z}_s)p(\mathbf{x}|\mathbf{z}_d)d\mathbf{z}_p d\mathbf{z}_s d\mathbf{z}_d$$

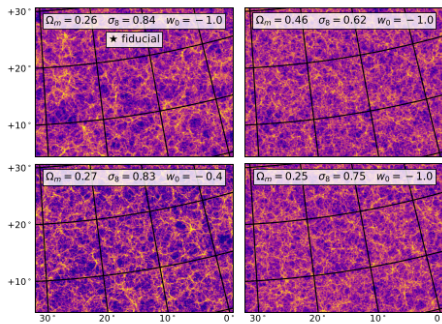


[Image source: M. Cacciari,
G. Salam, G. Soyez 0802.1189]

What can we do with generative models?

Produce samples and make predictions

$$\mathbf{x} \sim p(\mathbf{x}|\theta)$$

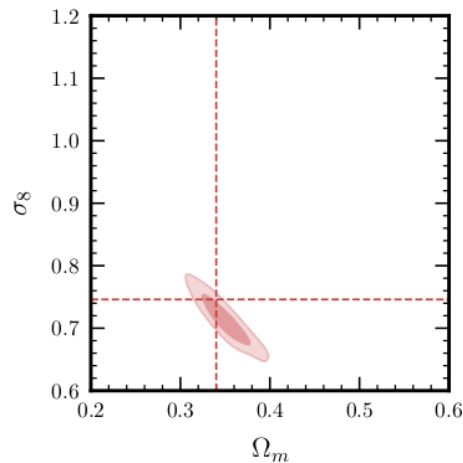


[Kacprzak et al 2022]

Evaluate densities

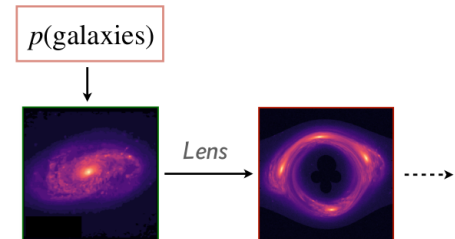
$$p(\mathbf{x}|\theta)$$

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

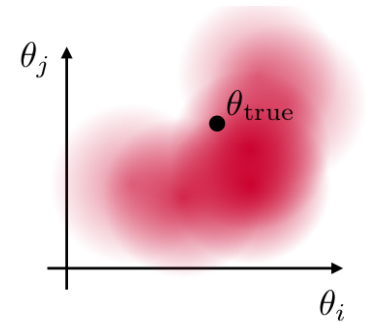
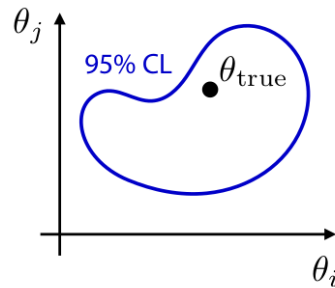
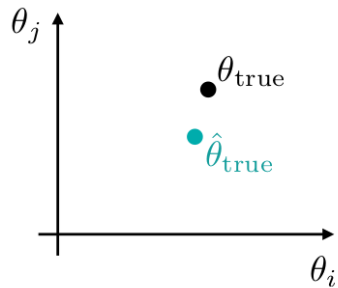


Encode complex priors

$$p(\mathbf{x})$$

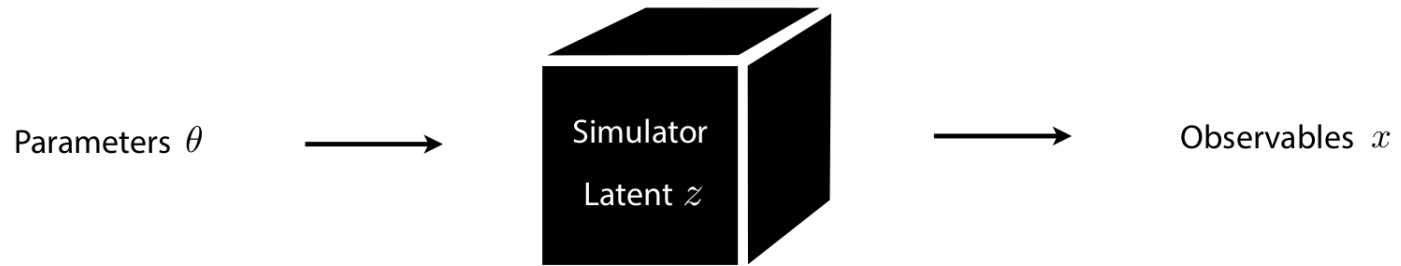


Inference



- Frequentist inference: find $\hat{\theta}$ that maximizes the likelihood $p(\mathbf{x}|\theta)$ or build a confidence interval thereof.
- Bayesian inference: compute the posterior distribution $p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$.

Statistical inference becomes challenging when the likelihood $p(\mathbf{x}|\theta)$ is implicit or intractable. **Simulation-based inference algorithms are needed.**

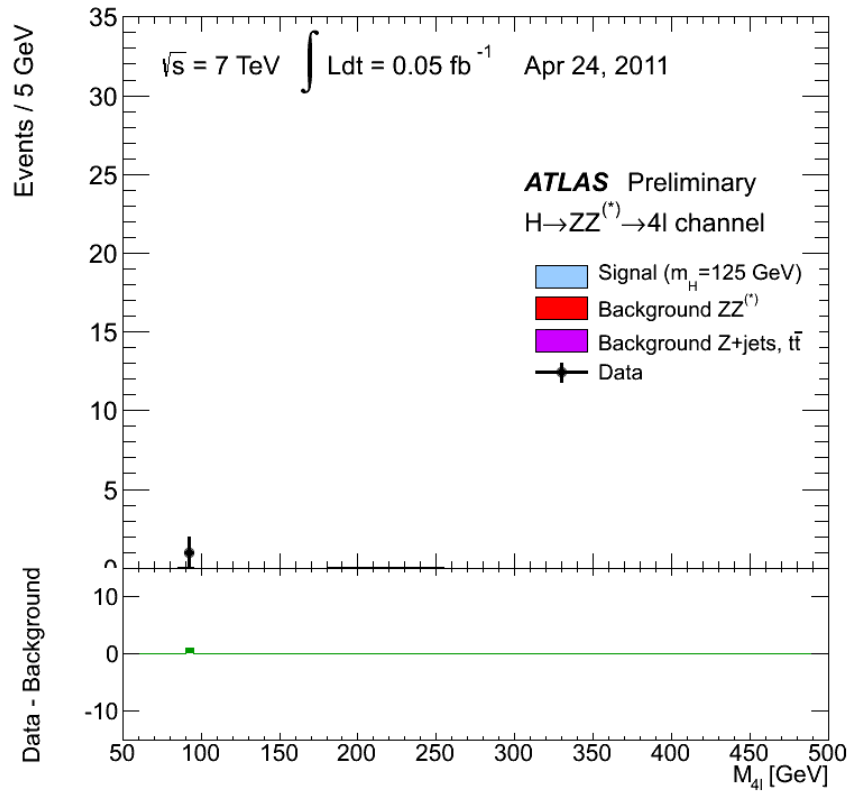


- Prediction:
- Well-motivated mechanistic, causal model
 - Simulator can generate samples $x \sim p(x|\theta)$

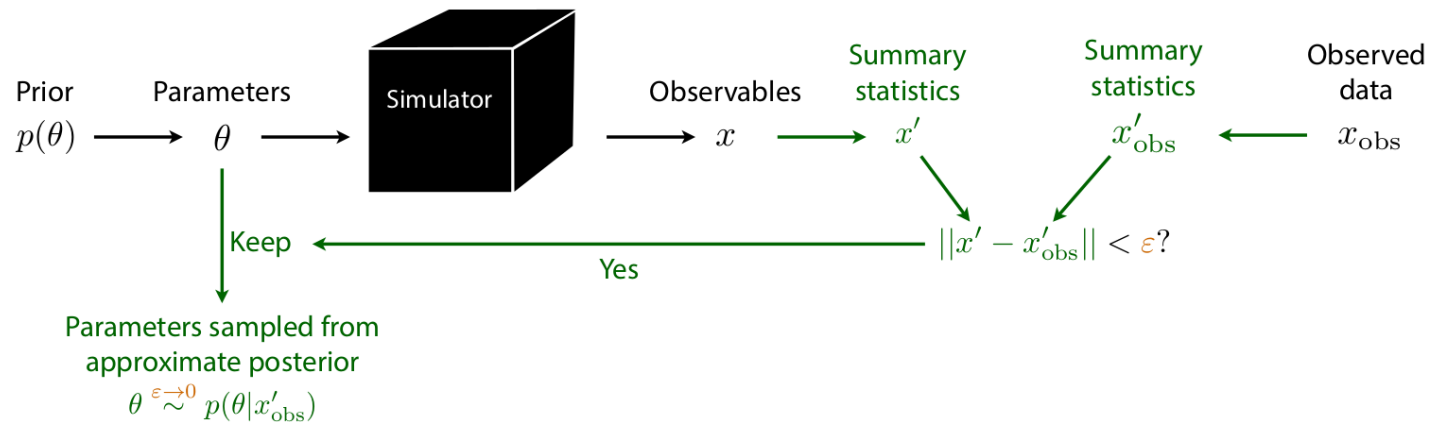
- Inference:
- Interactions between low-level components lead to challenging inverse problems
 - Likelihood $p(x|\theta) = \int dz p(x, z|\theta)$ is intractable

The frontiers

pre-2019



(Frequentist) Approximate the likelihood $p(\mathbf{x}|\theta)$ as
 $p(\mathbf{x}|\theta) \approx \hat{p}(\mathbf{x}|\theta) = p(s(\mathbf{x})|\theta)$ for some (well-chosen) summary statistic $s(\cdot)$.





(Bayesian) Approximate the posterior $p(\theta|\mathbf{x})$ using Approximate Bayesian Computation.

Issues:

- How to choose $\mathbf{x}' = \mathbf{s}(\mathbf{x})$? ϵ ? $\|\cdot\|$?
- No tractable posterior.
- Need to run new simulations for new data or new prior.



The frontier of simulation-based inference

Kyle Cranmer^{a,b,1} , Johann Brehmer^{a,b} , and Gilles Louppe^c

^aCenter for Cosmology and Particle Physics, New York University, New York, NY 10003; ^bCenter for Data Science, New York University, New York, NY 10011; and ^cMontefiore Institute, University of Liège, B-4000 Liège, Belgium

Edited by Jitendra Malik, University of California, Berkeley, CA, and approved April 10, 2020 (received for review November 4, 2019)

Many domains of science have developed complex simulations to describe phenomena of interest. While these simulations provide high-fidelity models, they are poorly suited for inference and lead to challenging inverse problems. We review the rapidly developing field of simulation-based inference and identify the forces giving additional momentum to the field. Finally, we describe how the frontier is expanding so that a broad audience can appreciate the profound influence these developments may have on science.

statistical inference | implicit models | likelihood-free inference | approximate Bayesian computation | neural density estimation

Mechanistic models can be used to predict how systems will behave in a variety of circumstances. These run the gamut of distance scales, with notable examples including particle physics, molecular dynamics, protein folding, population genetics, neuroscience, epidemiology, economics, ecology, climate science, astrophysics, and cosmology. The expressiveness of programming languages facilitates the development of complex, high-fidelity simulations and the power of modern computing provides the ability to generate synthetic data from them. Unfortunately, these simulators are poorly suited for statistical inference. The source of the challenge is that the probability density (or likelihood) for a given observation—an essential ingredient for both frequentist and Bayesian inference methods—is typically intractable. Such models are often referred to as implicit models and contrasted against prescribed models where the likelihood for an observation can be explicitly calculated (1). The problem setting of statistical inference under intractable likelihoods has been dubbed likelihood-free inference—although it is a bit of a misnomer as typically one attempts to estimate the intractable likelihood, so we feel the term simulation-based inference is more apt.

The intractability of the likelihood is an obstruction for scientific progress as statistical inference is a key component of the scientific method. In areas where this obstruction has appeared, scientists have developed various ad hoc or field-specific meth-

the simulator—is being recognized as a key idea to improve the sample efficiency of various inference methods. A third direction of research has stopped treating the simulator as a black box and focused on integrations that allow the inference engine to tap into the internal details of the simulator directly.

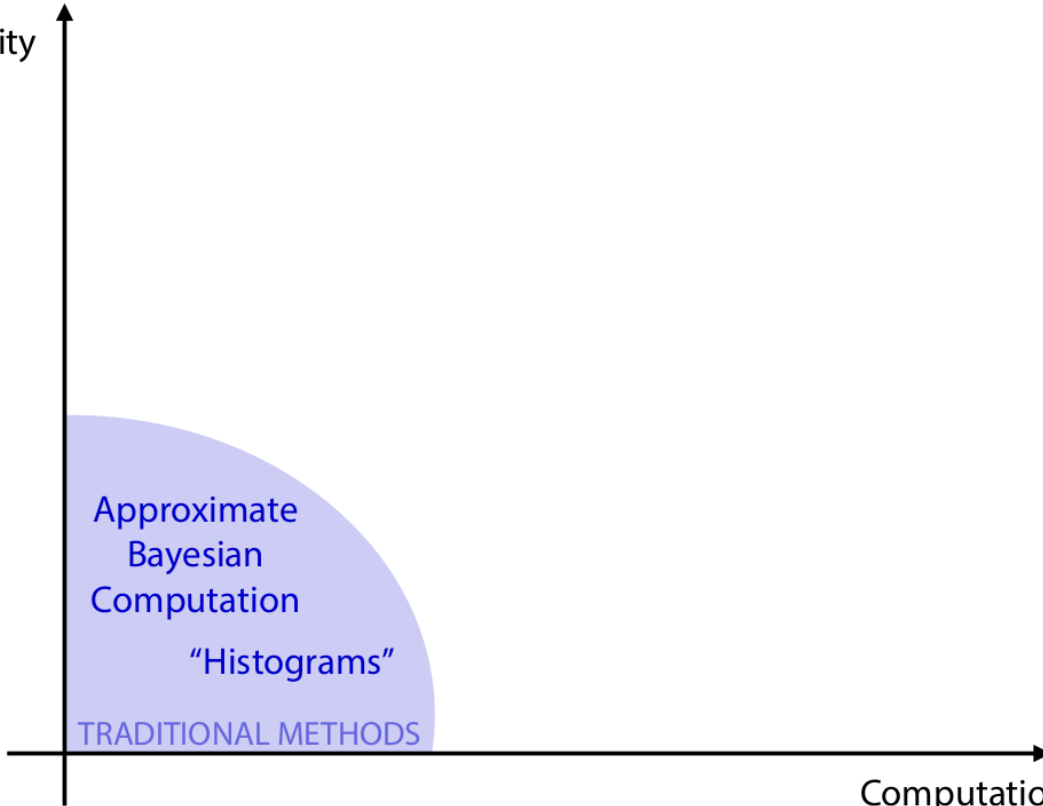
Amidst this ongoing revolution, the landscape of simulation-based inference is changing rapidly. In this review we aim to provide the reader with a high-level overview of the basic ideas behind both old and new inference techniques. Rather than discussing the algorithms in technical detail, we focus on the current frontiers of research and comment on some ongoing developments that we deem particularly exciting.

Simulation-Based Inference

Simulators. Statistical inference is performed within the context of a statistical model, and in simulation-based inference the simulator itself defines the statistical model. For the purpose of this paper, a simulator is a computer program that takes as input a vector of parameters θ , samples a series of internal states or latent variables $z_i \sim p_i(z_i|\theta, z_{<i})$, and finally produces a data vector $x \sim p(x|\theta, z)$ as output. Programs that involve random samplings and are interpreted as statistical models are known as probabilistic programs, and simulators are an example. Within this general formulation, real-life simulators can vary substantially:

- The parameters θ describe the underlying mechanistic model and thus affect the transition probabilities $p_i(z_i|\theta, z_{<i})$. Typically the mechanistic model is interpretable by a domain scientist and θ has relatively few components and a fixed dimensionality. Examples include coefficients found in the Hamiltonian of a physical system, the virulence and incubation rate of a pathogen, or fundamental constants of Nature.
- The latent variables z that appear in the data-generating process may directly or indirectly correspond to a physically meaningful state of a system, but typically this state is unobservable in practice. The structure of the latent space varies substantially between simulators. The latent variables may be continuous or discrete and the dimensionality of the latent space may be

Data
dimensionality

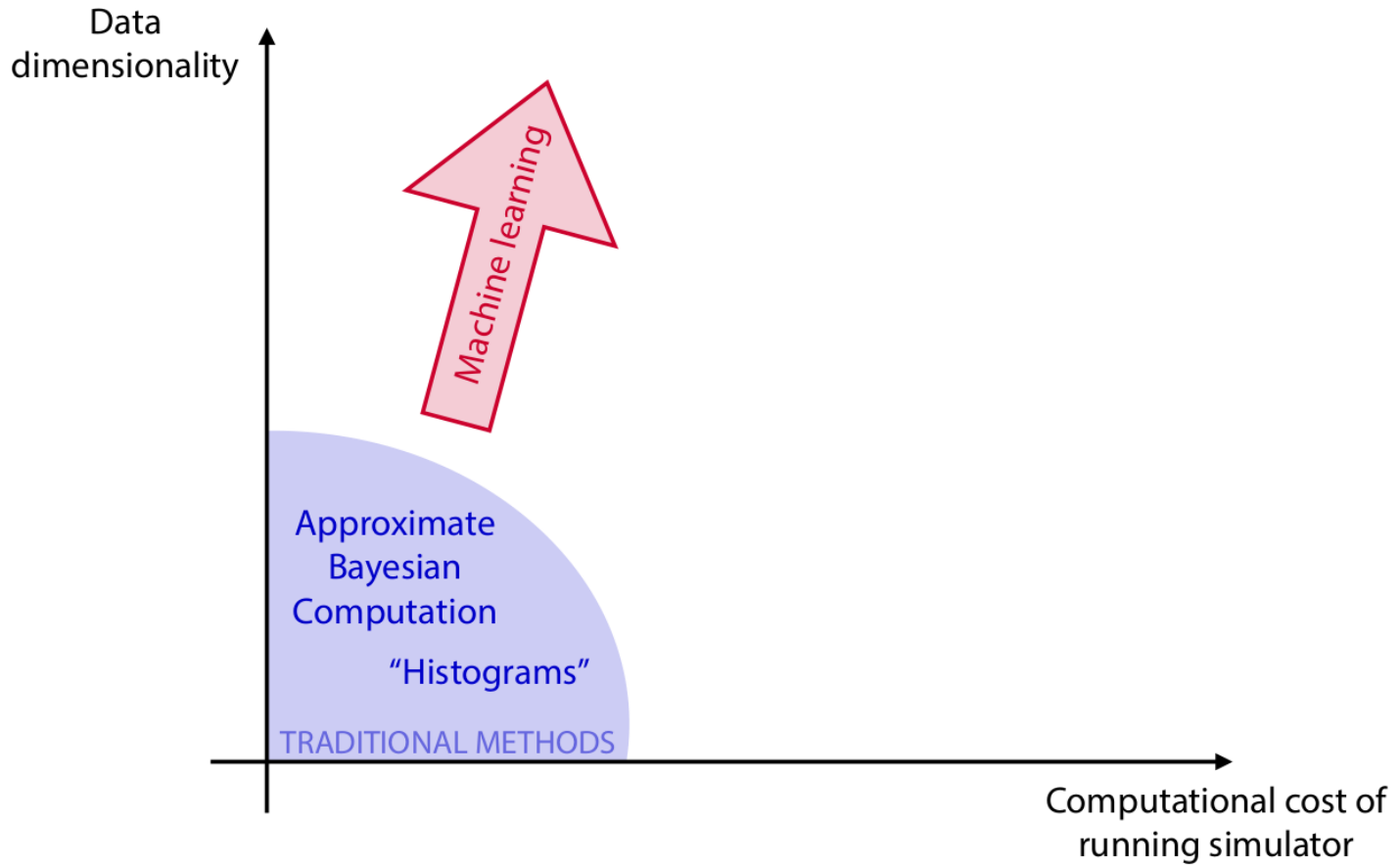


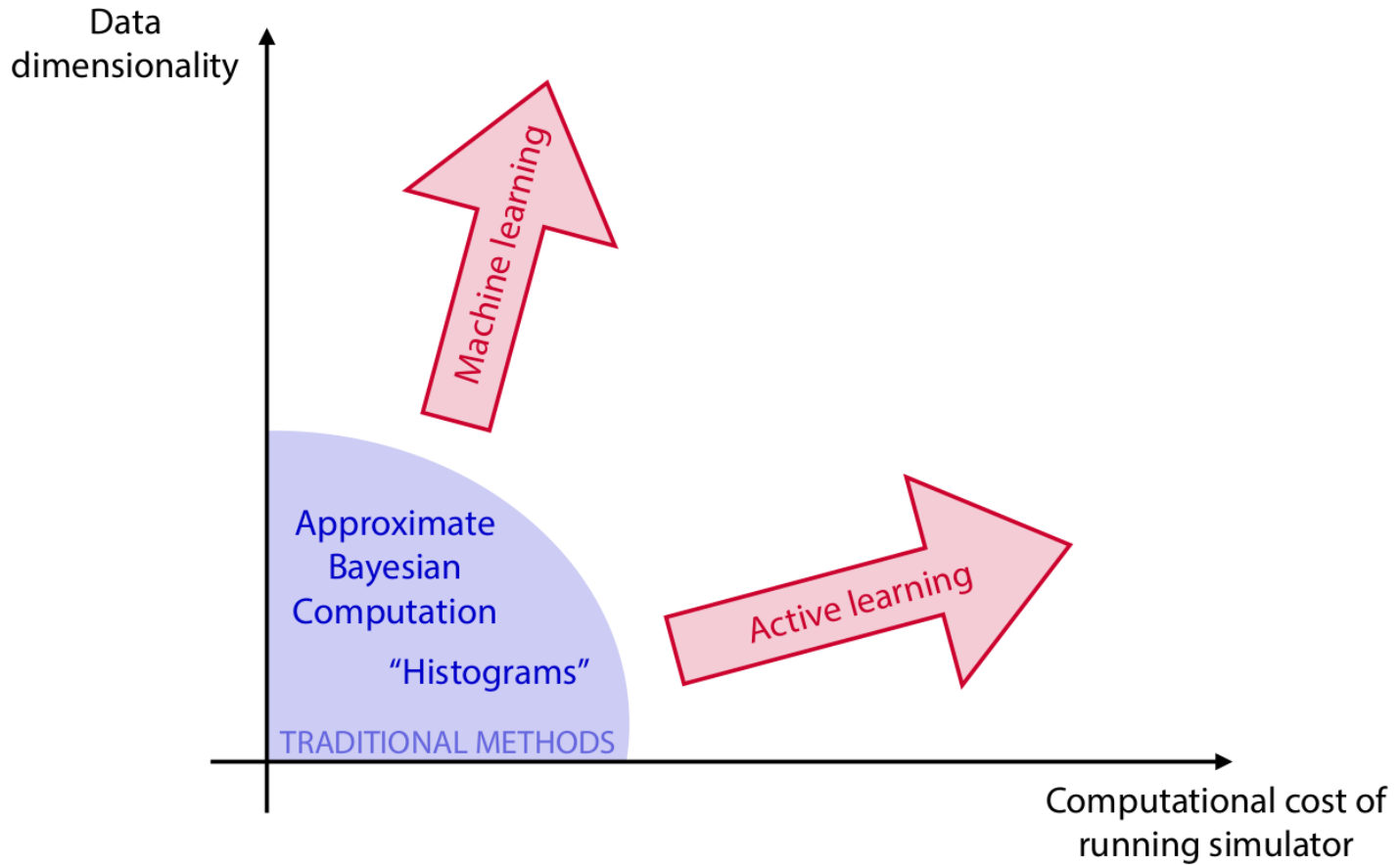
Approximate
Bayesian
Computation

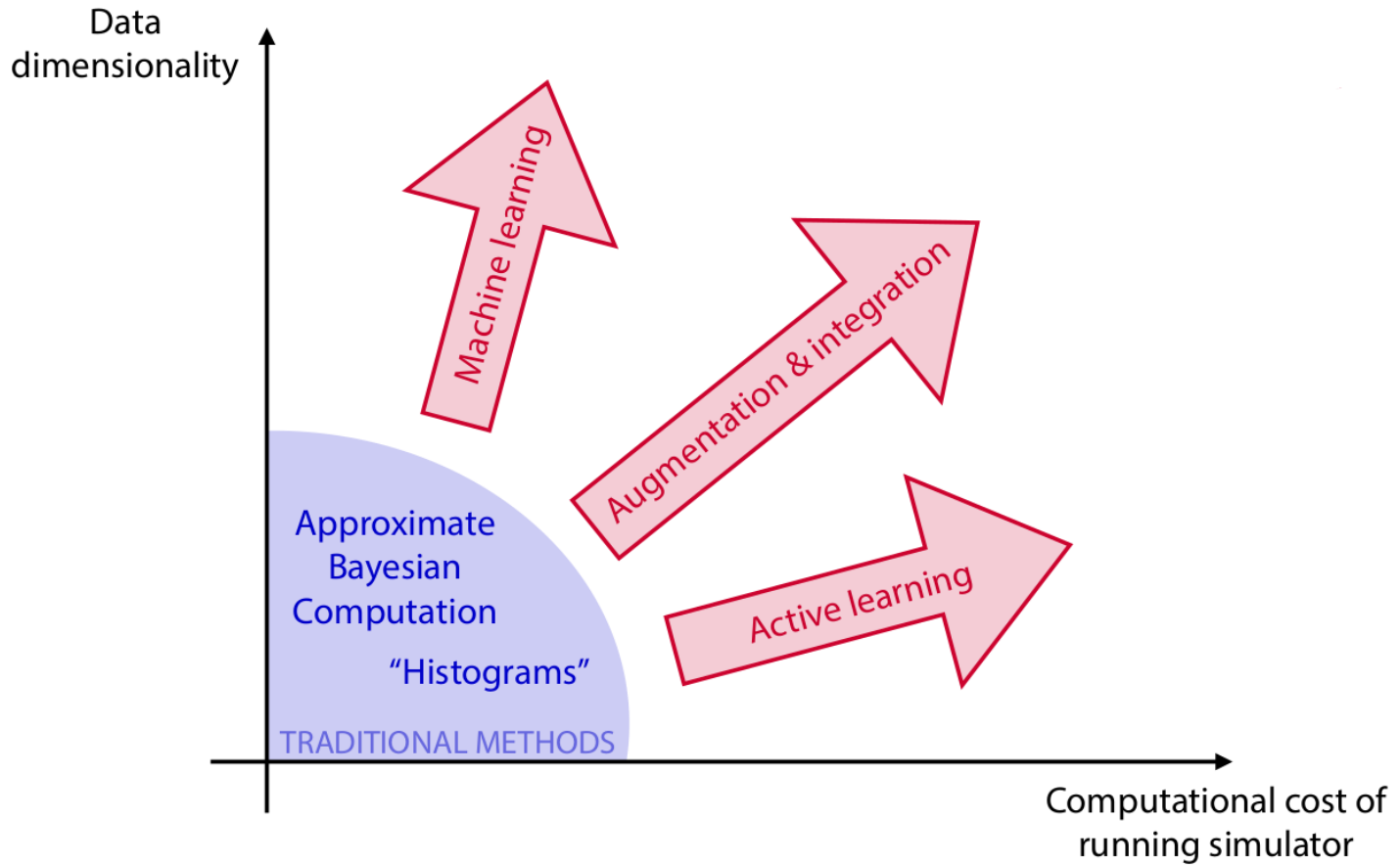
"Histograms"

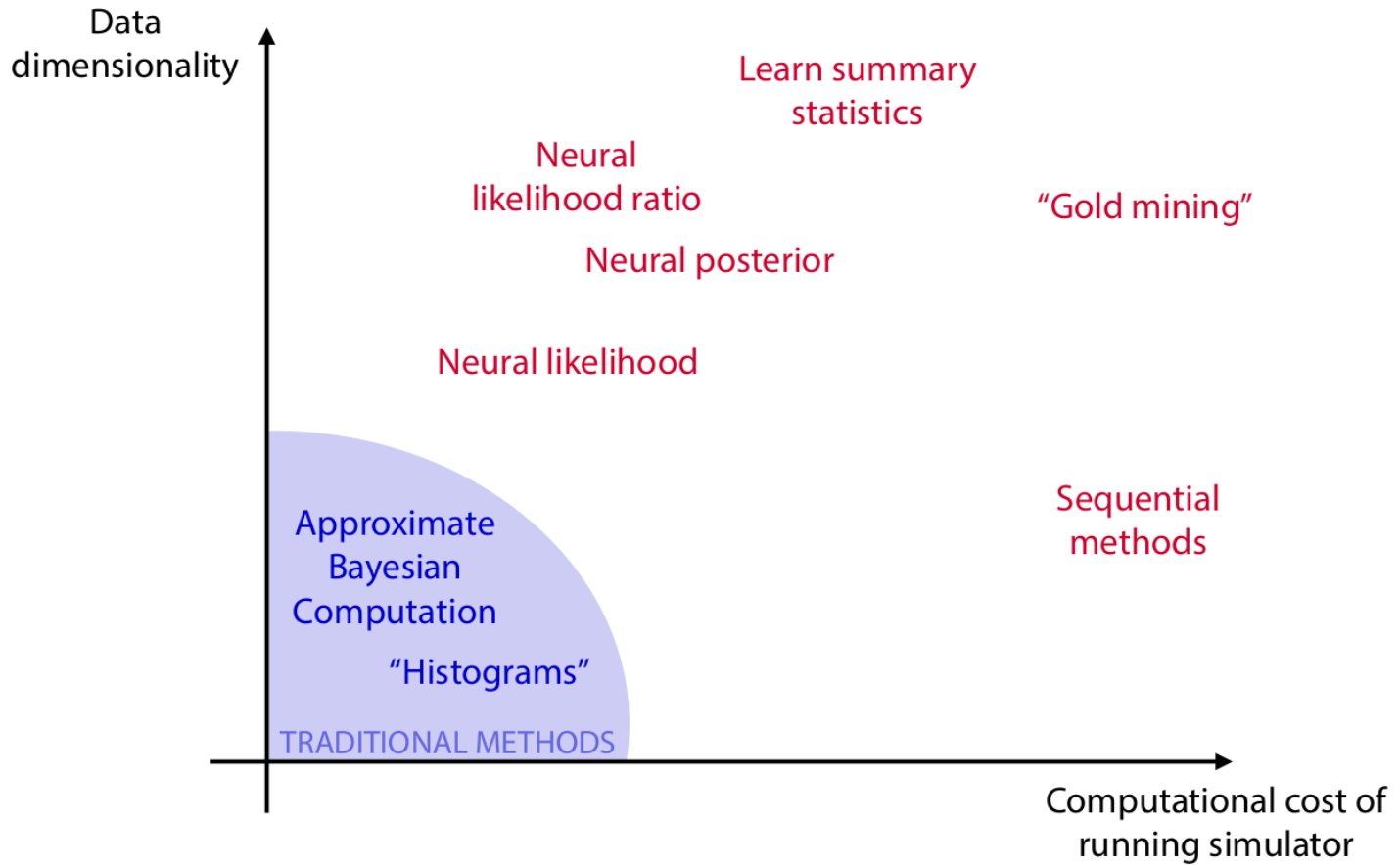
TRADITIONAL METHODS

Computational cost of
running simulator



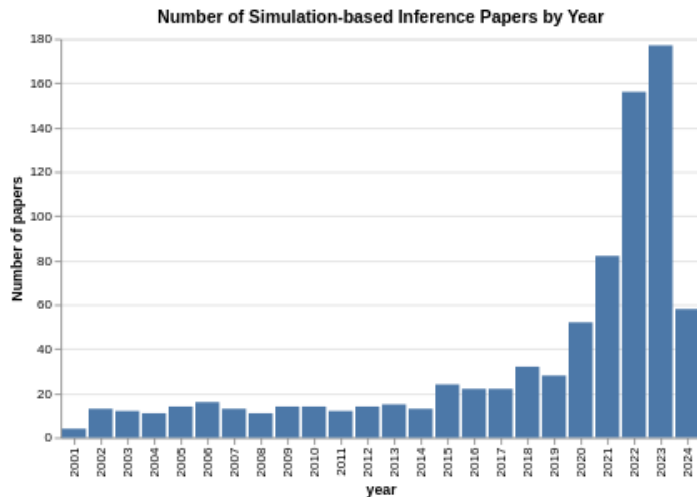






2024

The adoption of simulation-based inference has been growing steadily since then, with new algorithms and applications pushing the boundaries of what is possible.



Papers

The list is automatically compiled each day. Should you observe any inaccuracies or concerns, kindly bring them to our attention. Additionally, if you believe a new paper aligns with the topic, feel free to submit it. Visualize the annual growth in the number of publications.

Sort by Category

Sort by Year

Sort by Journal

Total (825)

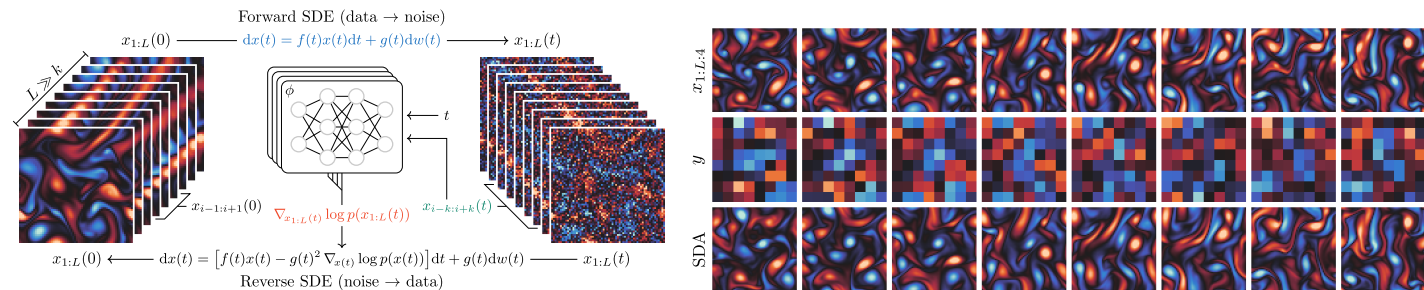
- Statistics (207)
- Computer Science (114)
- Astrophysics (100)
- Mathematics (57)
- Education (53)
- Economics (46)
- Physics (44)
- Quantitative Biology (32)
- Quantitative Finance (21)
- Astronomy (18)
- Engineering (14)
- Genetics (13)
- Epidemiology (11)
- Medicine (11)
- Geography (8)
- Social Science (7)
- Ecology (6)
- Evolutionary biology (6)
- Environmental Science (4)
- Cognitive Science (4)
- Robotics (4)
- Systems biology (4)
- Electrical Engineering and Systems Science (3)
- Systems Science (3)
- Bioinformatics (3)
- Bioinformatics (3)
- Biophysics (3)
- Mathematical (1)
- Geology (1)
- Musiology (1)
- statistical inference (1)
- Chemistry (1)
- Uncategorized (4)

Statistics

- Modelling Sampling Distributions of Test Statistics with Autograd, AA Kadhim, HB Prosper - arXiv preprint arXiv:2405.02488, 2024 - arxiv.org
- Preconditioned Neural Posterior Estimation for Likelihood-free Inference, X Wang, RP Kelly, D3 Warne, C Drovandi - arXiv preprint arXiv ..., 2024 - arxiv.org
- A variational neural Bayes framework for inference on intractable posterior distributions, E Maceda, EC Hector, A Lenzi, B3 Reich - arXiv preprint arXiv:2404.10899, 2024 - arxiv.org
- Increased perceptual reliability reduces membrane potential variability in cortical neurons, B von Hünenbein, J Jordan, M Oude Lohuis. - bioRxiv, 2024 - biorxiv.org
- How much information can be extracted from galaxy clustering at the field level?, NM Nguyen, F Schmidt, B Tucci, M Reinecke. - arXiv preprint arXiv ..., 2024 - arxiv.org
- Evolution of Analysis Techniques and Statistical Treatment, A Held - Bulletin of the American Physical Society, 2024 - APS
- Simulation-Based Inference with Quantile Regression, H Jia - arXiv preprint arXiv:2401.02413, 2024 - arxiv.org
- Direct Amortized Likelihood Ratio Estimation, AD Cobb, B Matejek, D Elenius, A Roy. - arXiv preprint arXiv ..., 2023 - arxiv.org
- On simulation-based inference for implicitly defined models, J Park - arXiv preprint arXiv:2311.09446, 2023 - arxiv.org
- Machine Learning for Mechanistic Models of Metapopulation Dynamics, J Li, EL Ionides, AA King, M Pascual, N Ning - arXiv preprint arXiv ..., 2023 - arxiv.org
- Inference on spatiotemporal dynamics for networks of biological populations, J Li, EL Ionides, AA King, M Pascual, N Ning - arXiv preprint arXiv ..., 2023 - arxiv.org
- Optimal simulation-based Bayesian decisions, J Alsing, TDP Edwards, B Wandelt - arXiv preprint arXiv:2311.05742, 2023 - arxiv.org
- Simulation based stacking, Y Yao, BRS Blancard, J Domke - arXiv preprint arXiv:2310.17009, 2023 - arxiv.org
- Calibrating Neural Simulation-Based Inference with Differentiable Coverage Probability, M Falkiewicz, N Takeishi, I Shekzadeh. - arXiv preprint arXiv ..., 2023 - arxiv.org
- Simulation-based Inference with the Generalized Kullback-Leibler Divergence, BK Miller, M Federici, C Weniger, P Forré - arXiv preprint arXiv ..., 2023 - arxiv.org
- Simulation-based Inference for Cardiovascular Models, A Wehenkel, J Behrmann, AC Miller, G Sapiro. - arXiv preprint arXiv ..., 2023 - arxiv.org
- Hierarchical Neural Simulation-Based Inference Over Event Ensembles, L Heinrich, S Mishra-Sharma, C Pollard. - arXiv preprint arXiv ..., 2023 - arxiv.org
- L-CST Local Diagnostics for Posterior Approximations in Simulation-Based Inference, J Linhart, A Gramfort, PLC Rodrigues - arXiv preprint arXiv:2306.03560, 2023 - arxiv.org
- Learning Robust Statistics for Simulation-based Inference under Model Misspecification, D Huang, A Bharti, A Souza, L Acerbi. - arXiv preprint arXiv ..., 2023 - arxiv.org



Developments in deep learning (e.g., diffusion models, transformers, GNNs, etc) have continued to **scale up simulation-based inference to higher dimensional simulation models** (both in the number of parameters θ and size of the data \mathbf{x}).



Rozet and Louppe et al (2023): "We introduce score-based data assimilation for trajectory inference. We learn a score-based generative model of state trajectories of a high-dimensional dynamical system and use it for the assimilation of noisy observations."

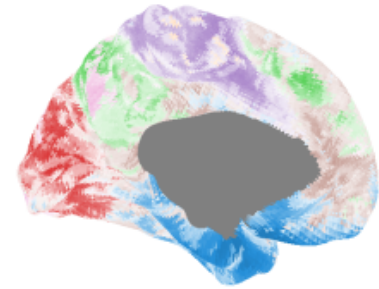
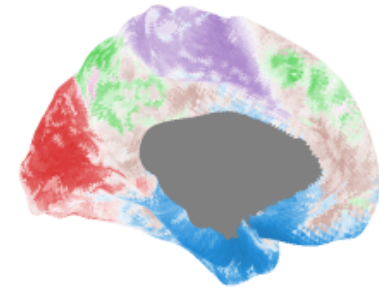
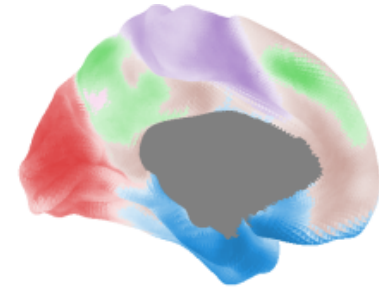
Active learning remains largely unexplored. Beyond greedy strategies, little attention has been given to the informed selection of simulations for building a training set.

Current paradigms cannot deal with expensive simulators (e.g., climate models, cosmological simulations).

Extracting side information based on θ , \mathbf{z} , \mathbf{x} remains challenging due to implementation constraints.

However, designing inference networks that **leverage domain knowledge** (e.g., symmetries, conservation laws) **or the structure of the simulator** (e.g., hierarchical models) has shown promising results.

Rouillard et al (2024): "We demonstrate the ability of PAVI to tackle large neuroimaging hierarchical inference problems. For each of the 59000 vertices of every of the 1000 subjects, we infer a probabilistic label to belong to one of the 7 functional networks. This amounts to inferring over 400 million latent variables."





A case study

Hermans et al, "Constraining dark matter with stellar streams", 2021.



Can we constrain the nature of dark matter from cosmological observations?

Constraining dark matter with stellar streams

Palomar 5 (Pal5) stream

Pal5 was discovered in 2001 as the first thin stream formed from a globular cluster. Its current orbit takes it far over the galactic center.

Globular clusters

These hives typically hold 100,000 stars or fewer and give rise to long, thin streams.

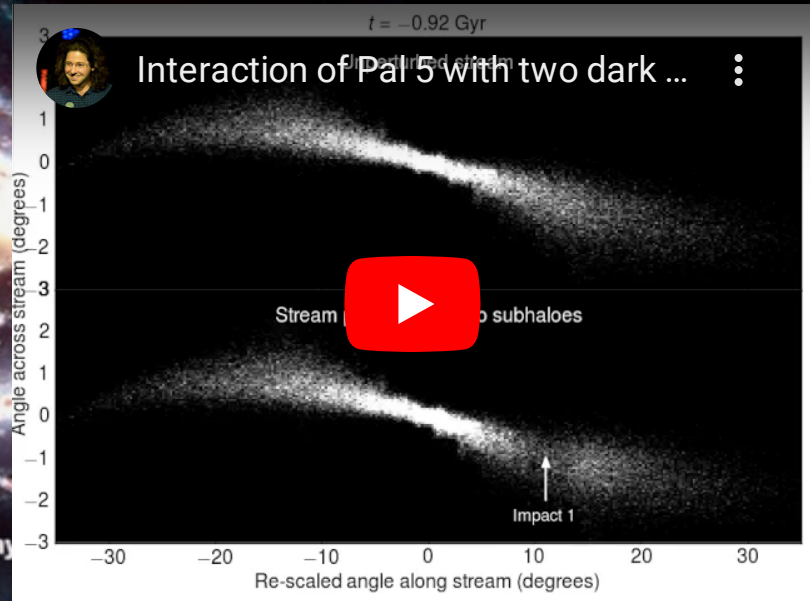
Gap

Sun

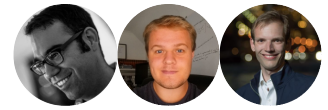
GD1 stream

Discovered in 2006, GD1 is the longest known thin stream, stretching across more than half the northern sky. It contains a gap that could be the scar of a dark matter collision 500 million years ago.

Milky Way

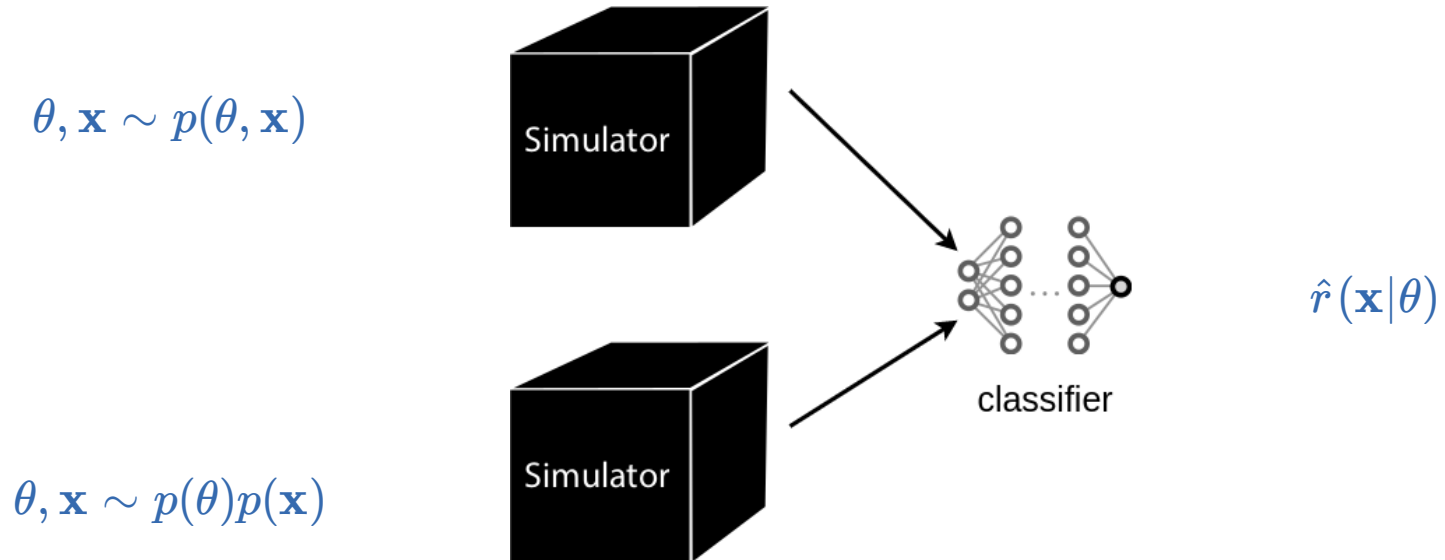


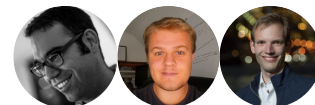
$$p(m_{\text{WDM}}, t_{\text{age}} | \text{GD1}) = \frac{p(\text{GD1} | m_{\text{WDM}}, t_{\text{age}}) p(m_{\text{WDM}}, t_{\text{age}})}{p(\text{GD-1})}$$



Neural ratio estimation (NRE)

The likelihood-to-evidence $r(\mathbf{x}|\theta) = \frac{p(\mathbf{x}|\theta)}{p(\mathbf{x})} = \frac{p(\theta, \mathbf{x})}{p(\theta)p(\mathbf{x})}$ ratio can be estimated from a binary classifier $d(\theta, \mathbf{x})$, even if neither the likelihood nor the evidence can be evaluated.



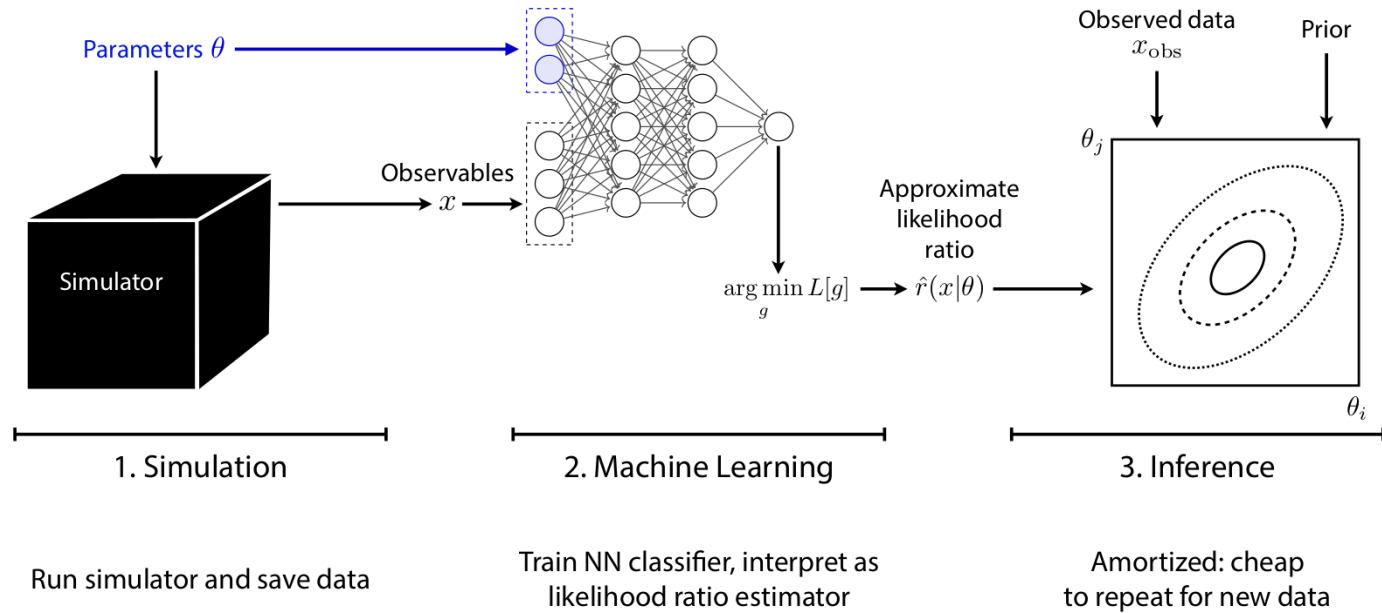
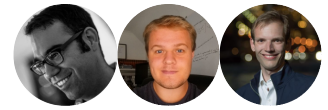


The solution d found after training approximates the optimal classifier

$$d(\theta, \mathbf{x}) \approx d^*(\theta, \mathbf{x}) = \frac{p(\theta, \mathbf{x})}{p(\theta, \mathbf{x}) + p(\theta)p(\mathbf{x})}.$$

Therefore,

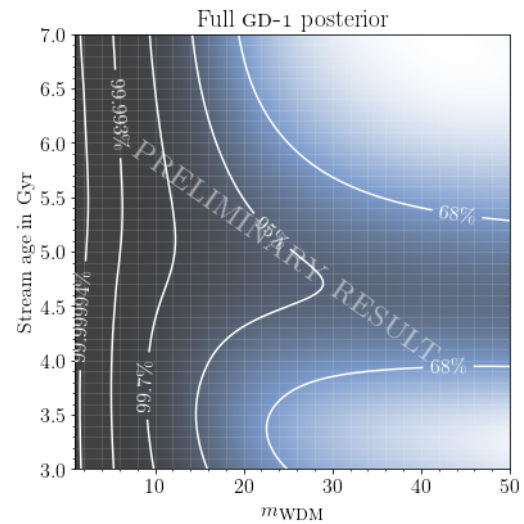
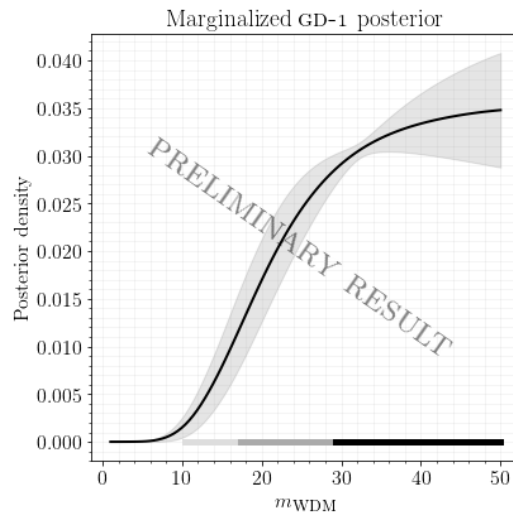
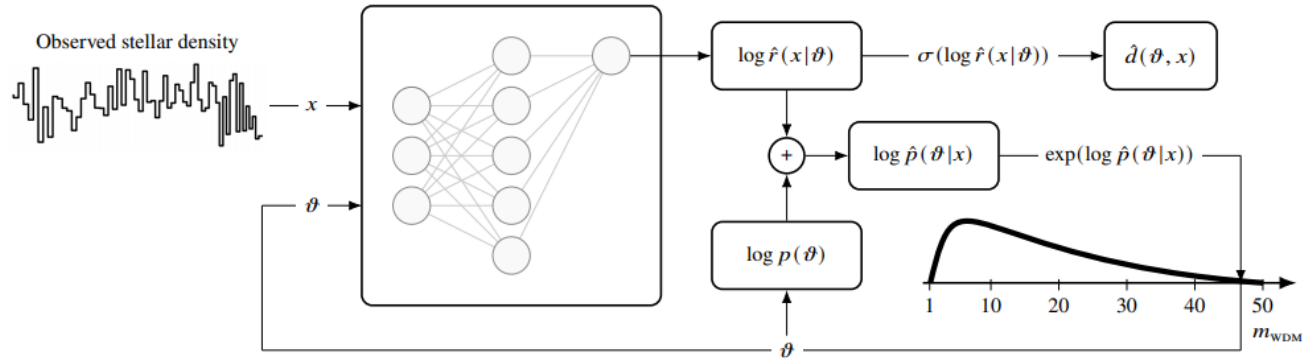
$$r(\mathbf{x}|\theta) = \frac{p(\mathbf{x}|\theta)}{p(\mathbf{x})} = \frac{p(\theta, \mathbf{x})}{p(\theta)p(\mathbf{x})} \approx \frac{d(\theta, \mathbf{x})}{1 - d(\theta, \mathbf{x})} = \hat{r}(\mathbf{x}|\theta).$$

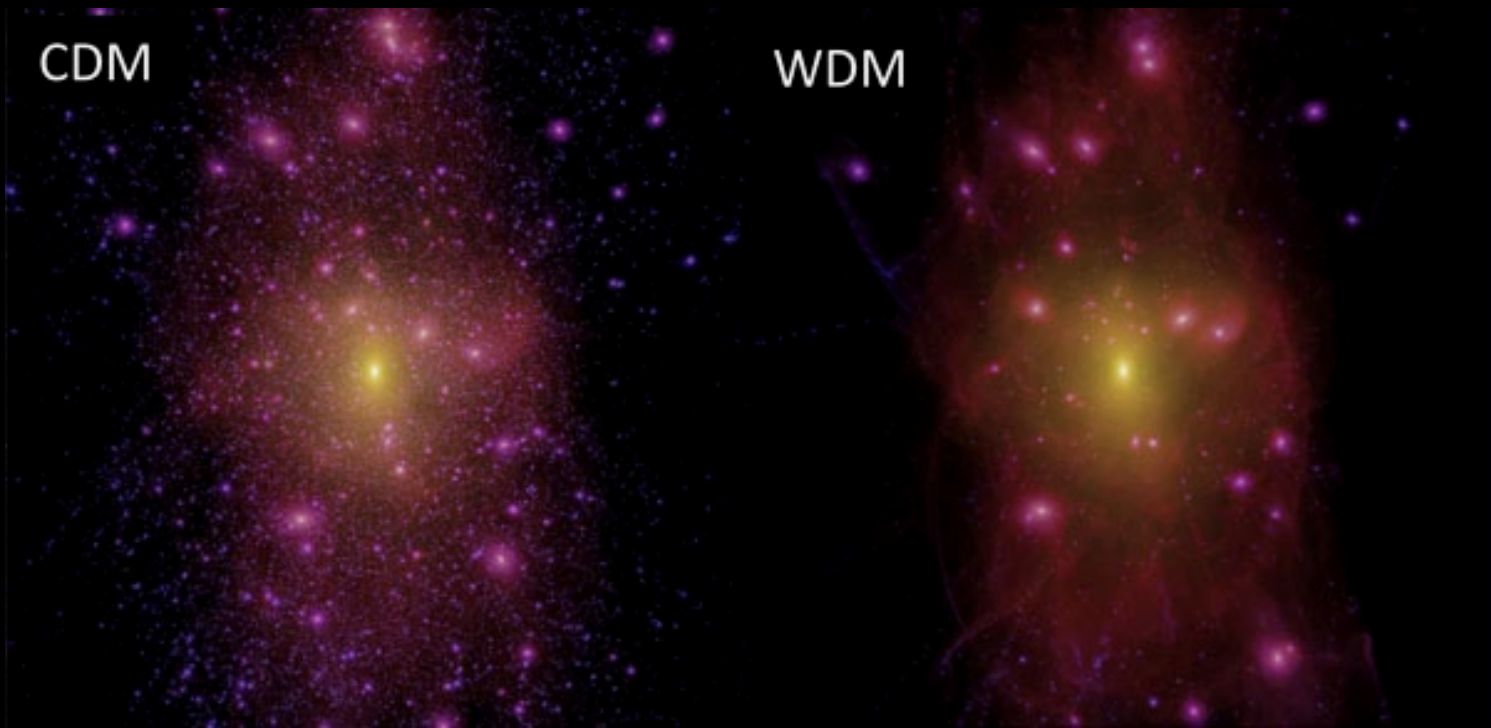


$$p(\theta|\mathbf{x}) \approx \hat{r}(\mathbf{x}|\theta)p(\theta)$$



NRE for stellar streams





Preliminary results for GD-1 suggest a preference for CDM over WDM.

Wait a minute Gilles...
I can't claim that in a paper!
Your neural network must be wrong!

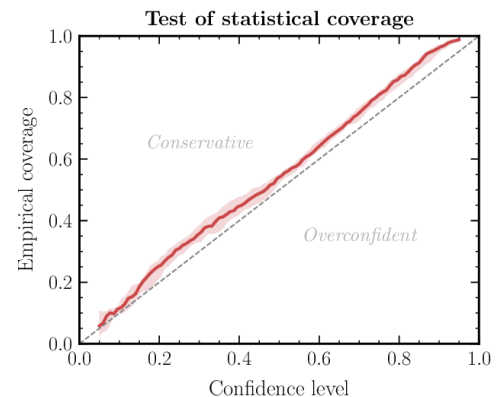
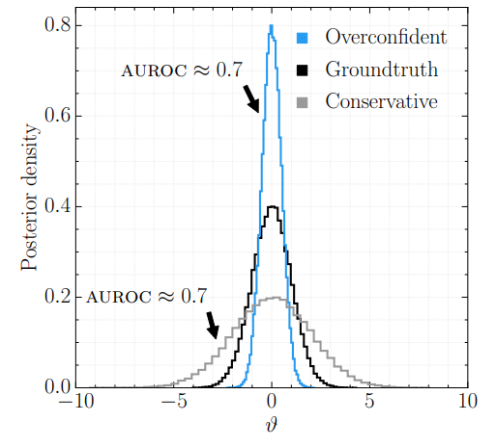


Expected coverage

$$EC(\hat{p}, \alpha) = \mathbb{E}_{p(\theta, \mathbf{x})} [\theta \in \Theta_{\hat{p}(\theta|\mathbf{x})}(\alpha)]$$

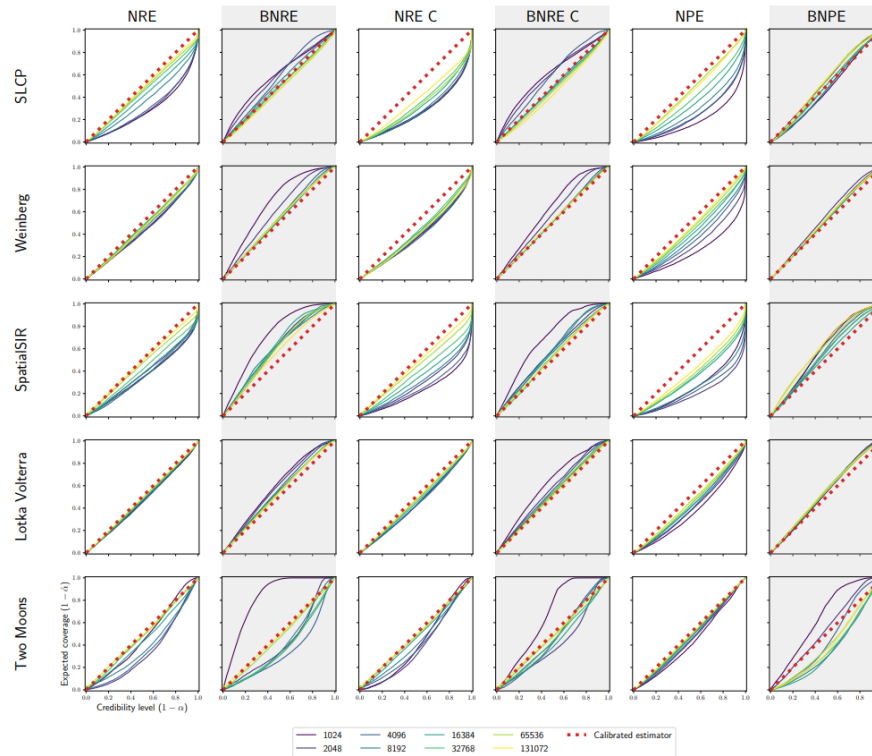
If the expected coverage is close to the nominal coverage probability α , then the approximate posterior \hat{p} is calibrated.

- If $EC < \alpha$, then the posterior is underdispersed and overconfident.
- If $EC > \alpha$, then the posterior is overdispersed and conservative.





Balancing inference for conservative posteriors



Conservative posteriors can be obtained by enforcing d to be balanced, i.e. such that $\mathbb{E}_{p(\theta, \mathbf{x})} [d(\theta, \mathbf{x})] = \mathbb{E}_{p(\theta)p(\mathbf{x})} [1 - d(\theta, \mathbf{x})]$.

Summary

Simulation-based inference is a major evolution in the statistical capabilities for science, as it enables the analysis of complex models and data without simplifying assumptions.

Obstacles remain to be overcome, such as the curse of dimensionality, the need for large amounts of data, or the necessary robustness of the inference network.

The next frontiers? Let's find out this week!