Reliable Simulation-Based Inference with Balanced Neural Ratio Estimation

NASC seminar, Pepperdine University

April 18, 2023

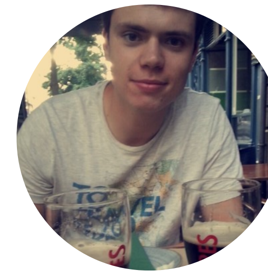
Gilles Louppe
g.louppe@uliege.be

Kyle Cranmer

Johann Brehmer

Joeri Hermans

Antoine Wehenkel



Norman Marlier

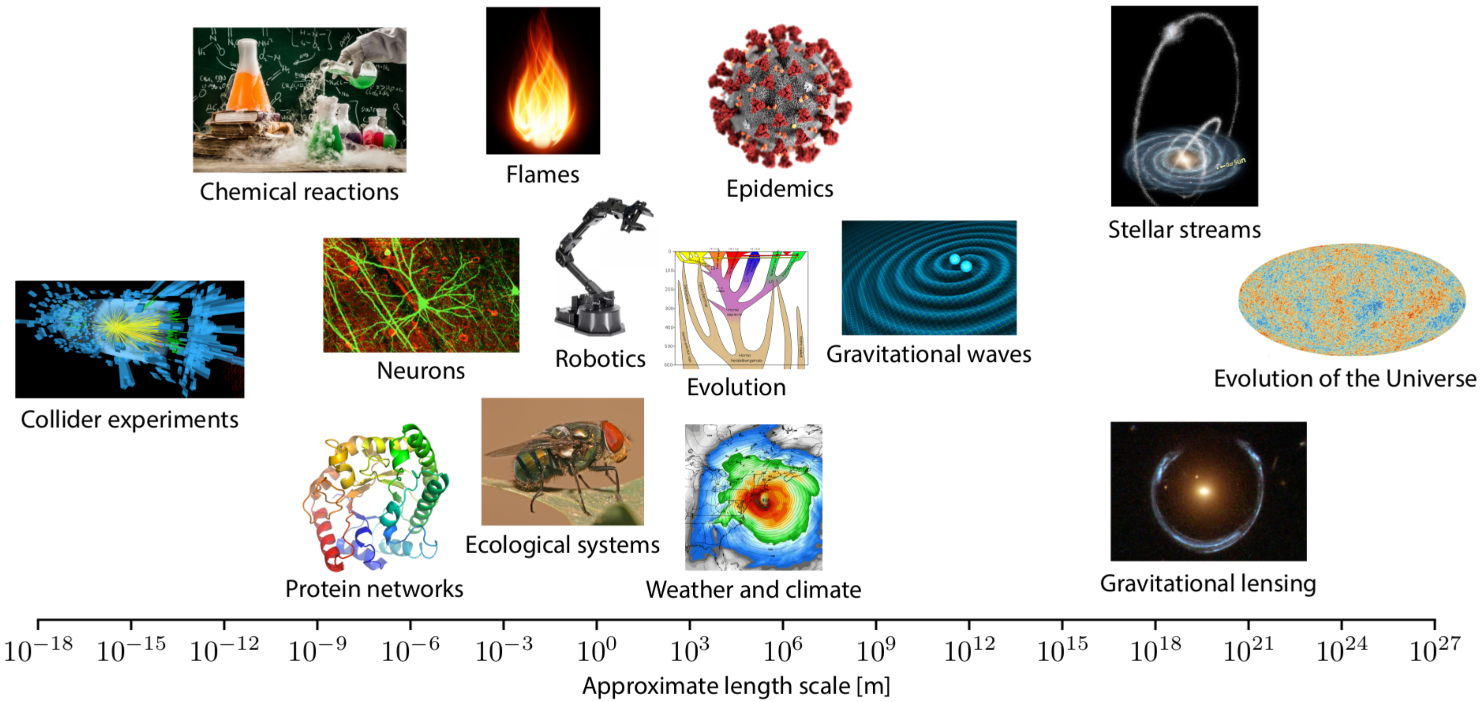
Siddharth Mishra-Sharma

Christoph Weniger

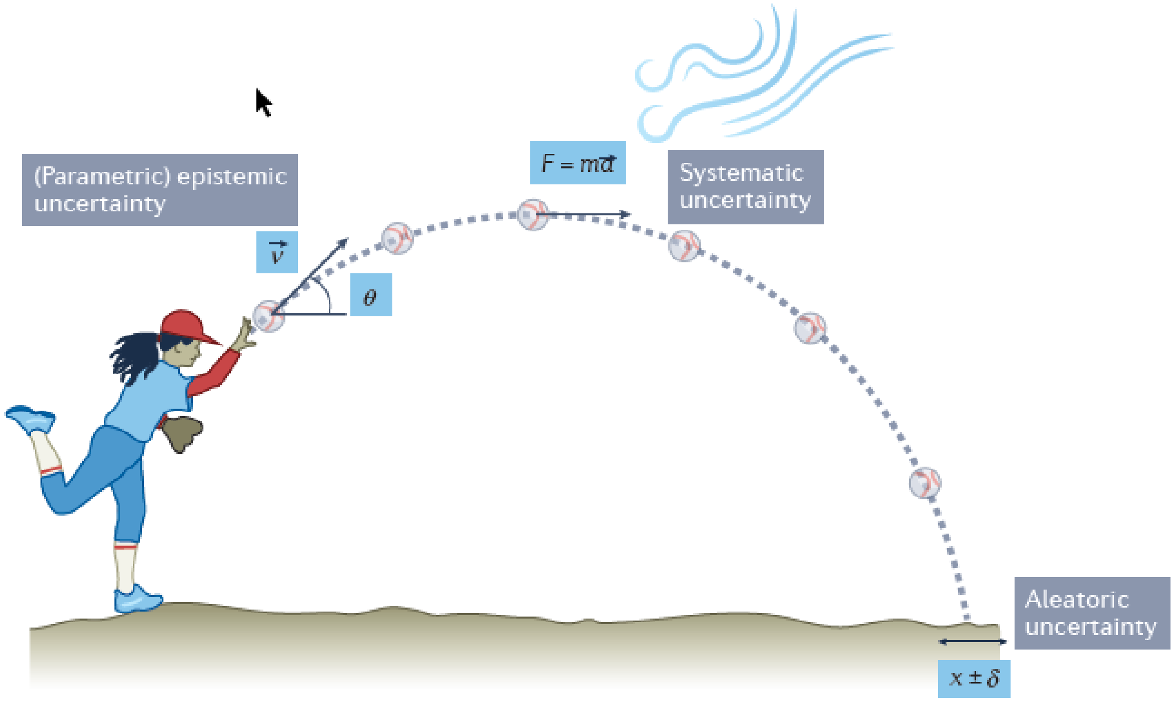
Arnaud Delaunoy

Malavika Vasist

Francois Rozet


Chemical reactions
Flames
Epidemics
Stellar streams
Neurons
Robotics
Evolution
Gravitational waves
Evolution of the Universe
Collider experiments
Ecological systems
Protein networks
Weather and climate
Gravitational lensing
10^{-18} 10^{-15} 10^{-12} 10^{-9} 10^{-6} 10^{-3} 10^{0} 10^{3} 10^{6} 10^{9} 10^{12} 10^{15} 10^{18} 10^{21} 10^{24} 10^{27}
Approximate length scale [m]


CSN
LAD 0
SF 2
3RD
PITCHES 40
0-0
2 OUT
GIANTS

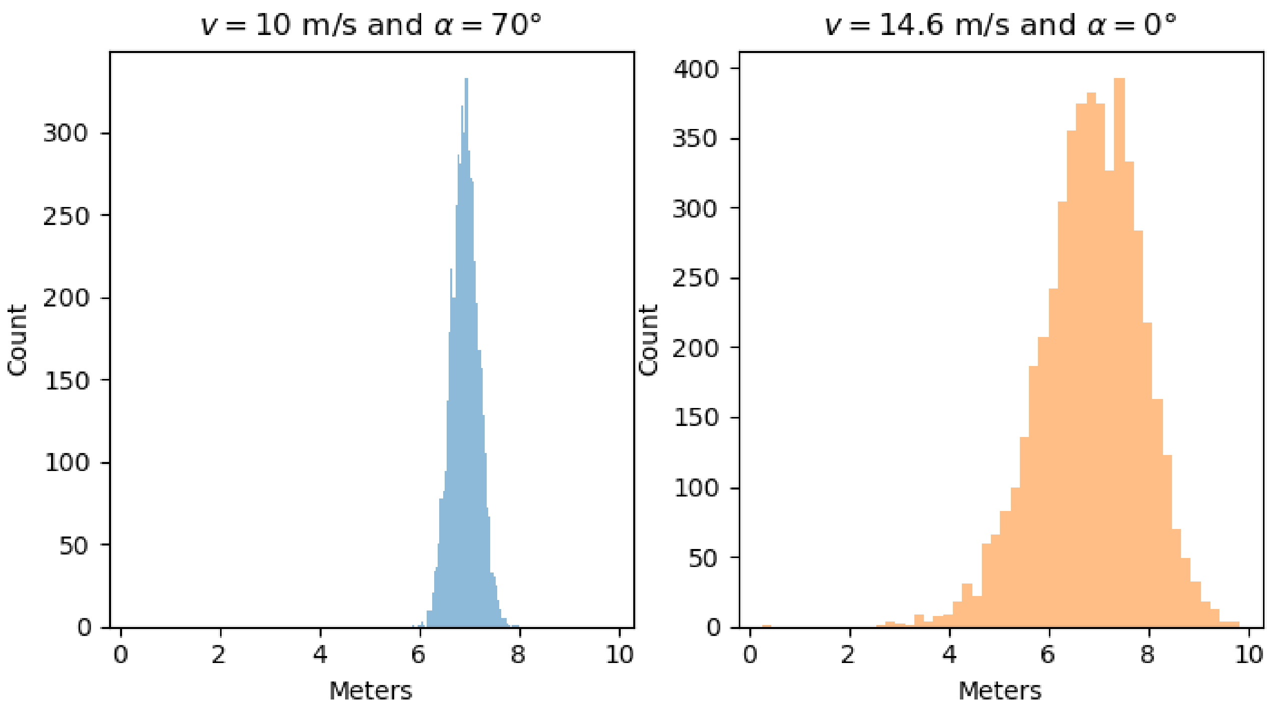

(Parametric) epistemic uncertainty
$\vec{v}$
θ
$F = m\vec{a}$
Systematic uncertainty
Aleatoric uncertainty
$x \pm \delta$

$$v_x = v\cos(\alpha), \quad v_y = v\sin(\alpha),$$

$$\frac{dx}{dt} = v_x, \quad \frac{dy}{dt} = v_y, \frac{dv_y}{dt} = -G.$$

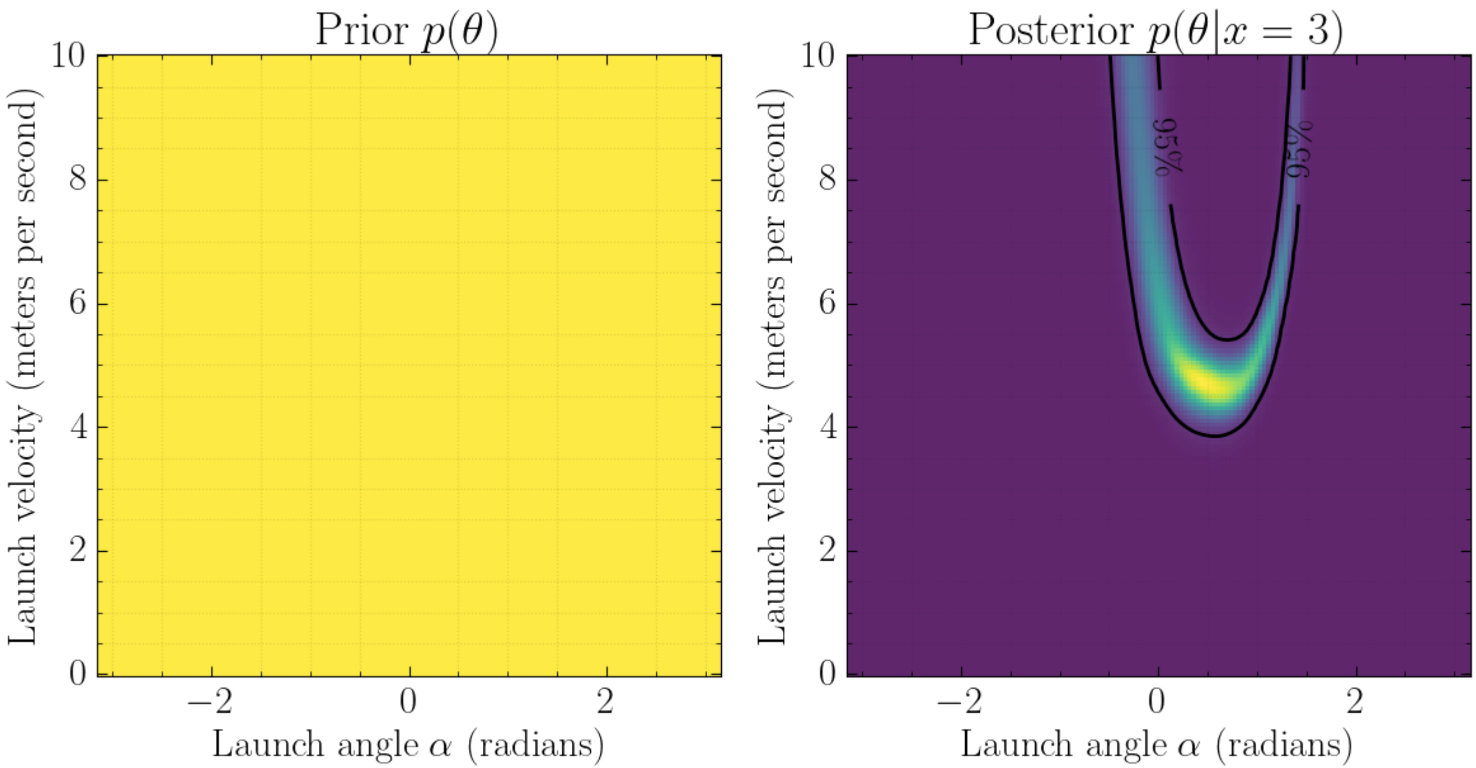
```python
def simulate(v, alpha, dt=0.001):
    v_x = v * np.cos(alpha)  # x velocity m/s
    v_y = v * np.sin(alpha)  # y velocity m/s
    y = 1.1 + 0.3 * random.normal()
    x = 0.0

    while y > 0: # simulate until ball hits floor
        v_y += dt * -G  # acceleration due to gravity
        x += dt * v_x
        y += dt * v_y

    return x + 0.25 * random.normal()
```

The computer simulator defines the likelihood function $p(x|\theta)$ implicitly.



What parameter values θ are the most plausible?

$$p(\theta|x_{\mathrm{obs}}) = \frac{p(x_{\mathrm{obs}}|\theta)p(\theta)}{p(x_{\mathrm{obs}})}$$

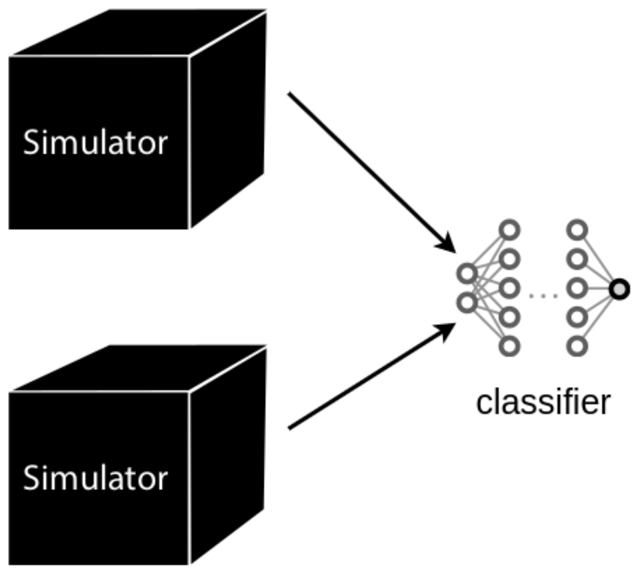

Prior $p(\theta)$
Posterior $p(\theta|x = 3)$
Launch velocity (meters per second)
Launch angle α (radians)
95%
95%
10
8
6
4
2
0
−2
0
2

Neural ratio estimation (NRE)

The likelihood-to-evidence $r(x|\theta) = \frac{p(x|\theta)}{p(x)} = \frac{p(x,\theta)}{p(x)p(\theta)}$ ratio can be learned, even if neither the likelihood nor the evidence can be evaluated:

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

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



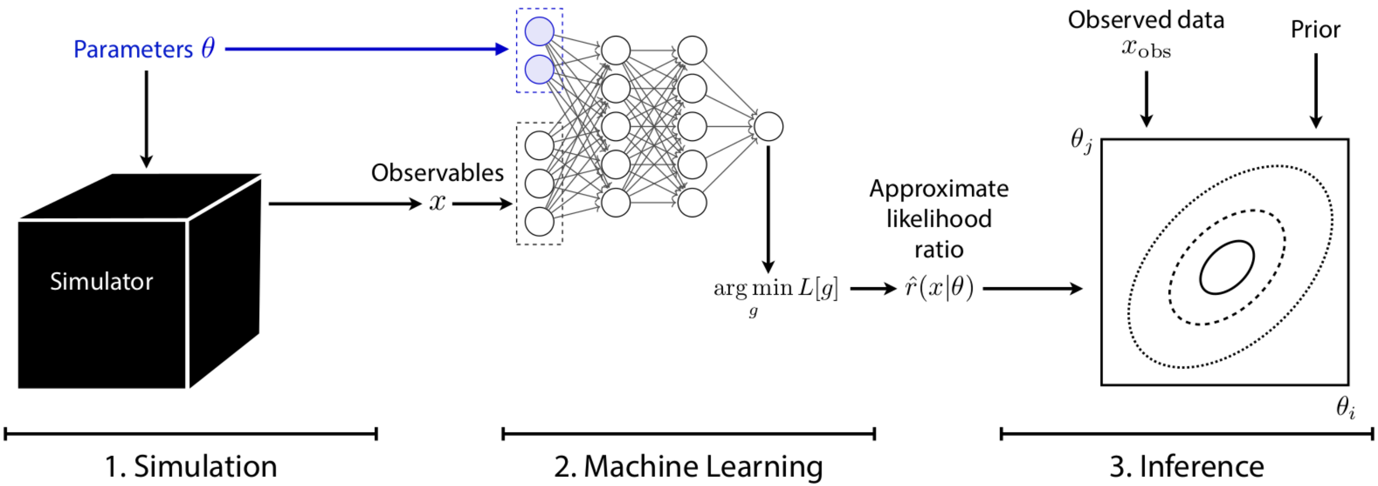
$\hat{r}(x|\theta)$

The solution d found after training approximates the optimal classifier

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

Therefore,

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



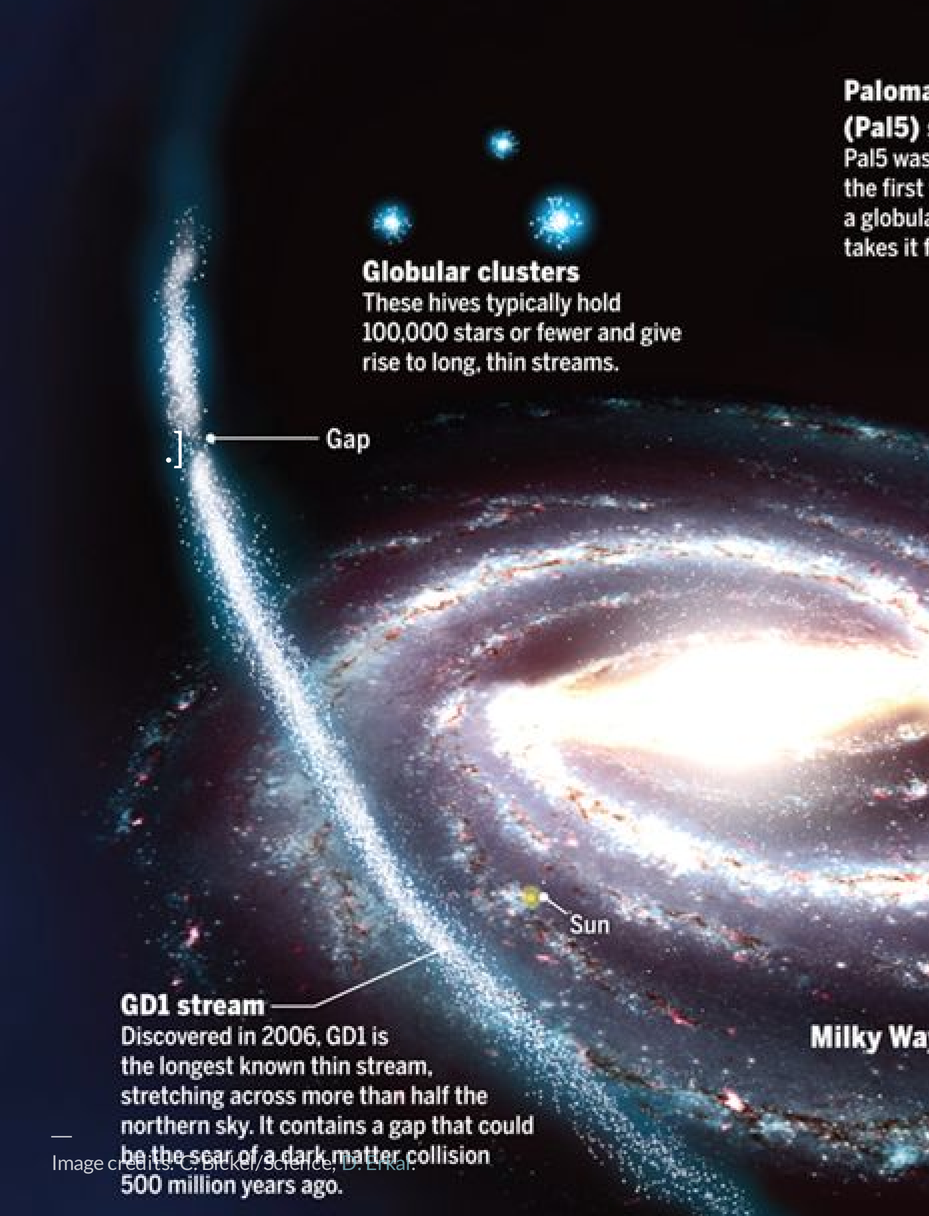
Run simulator and save data

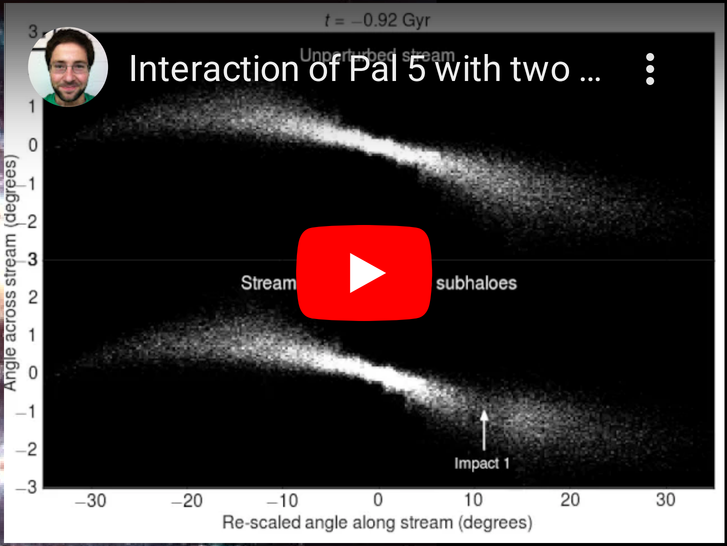
Train NN classifier, interpret as likelihood ratio estimator

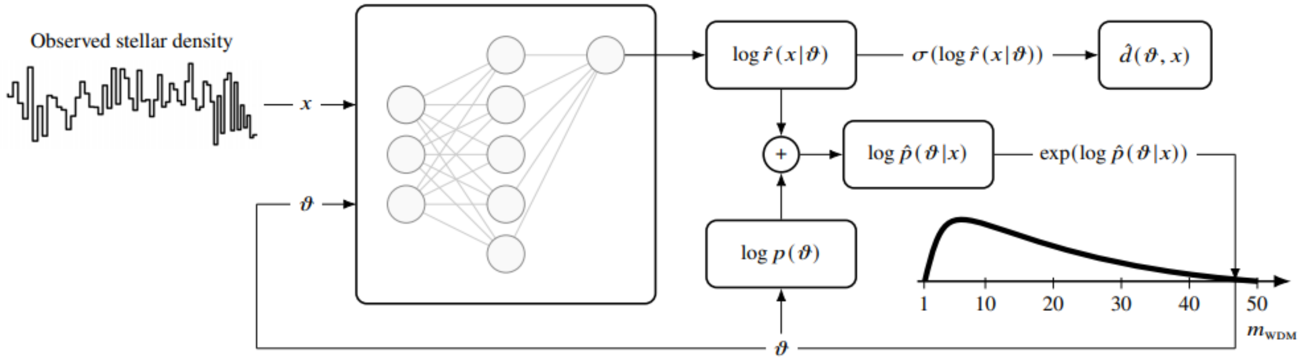
Amortized: cheap to repeat for new data

$$p(\theta|x) \approx r(x|\theta)p(\theta)$$

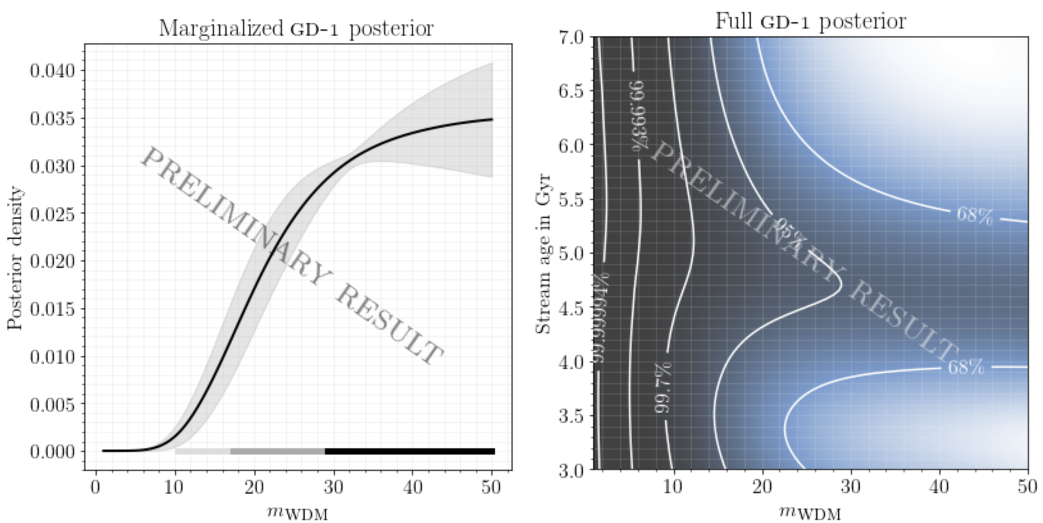
Constraining dark matter with stellar streams (Bayesian)







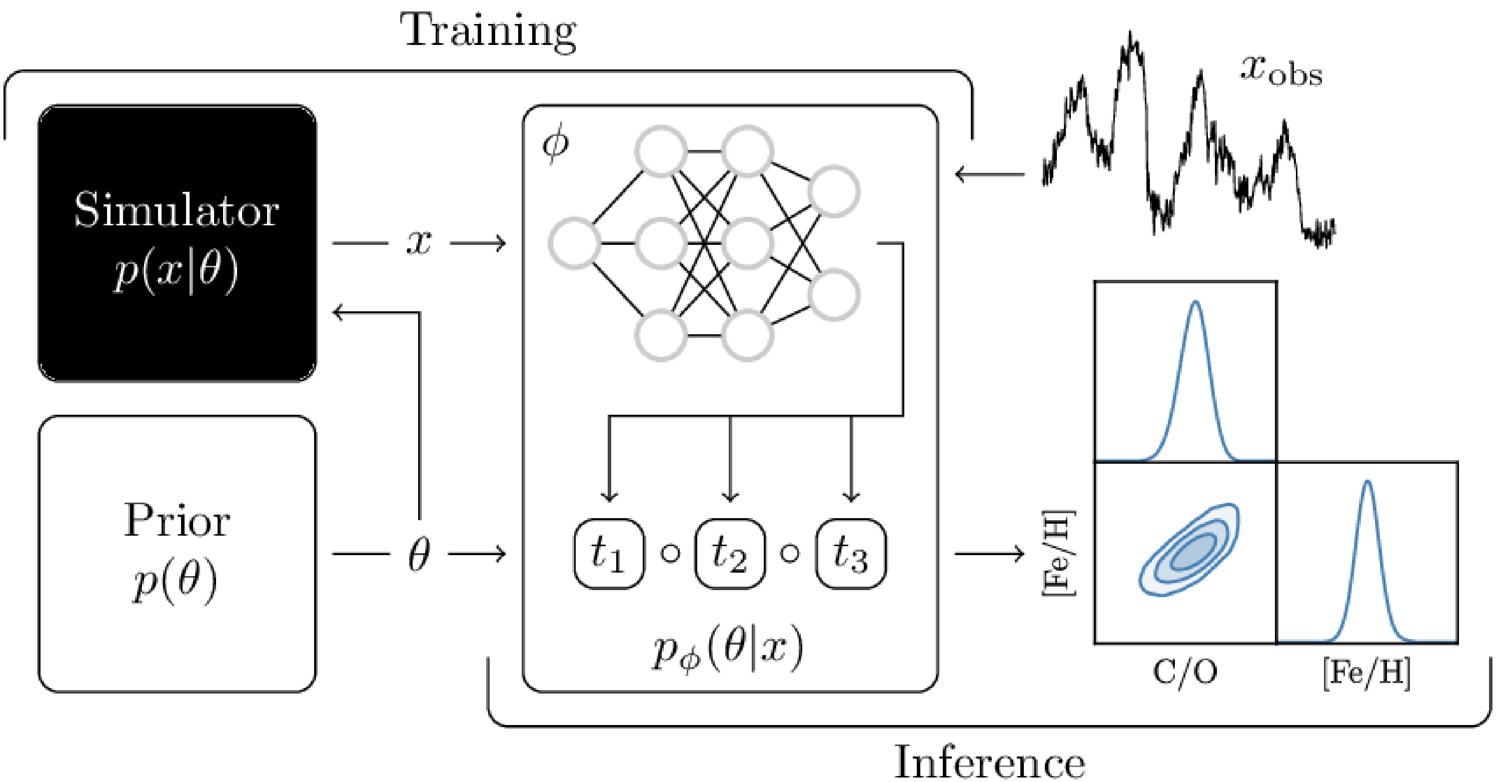
Observed stellar density
x
ϑ
$\log \hat{r}(x|\vartheta)$
$\sigma(\log \hat{r}(x|\vartheta))$
$\hat{d}(\vartheta, x)$
$\log \hat{p}(\vartheta|x)$
$\exp(\log \hat{p}(\vartheta|x))$
$\log p(\vartheta)$
1 10 20 30 40 50
m_{WDM}



Marginalized GD-1 posterior
Posterior density
m_{WDM}
PRELIMINARY RESULT
Full GD-1 posterior
Stream age in Gyr
99.993%
99.99994%
99.7%
95%
68%
68%

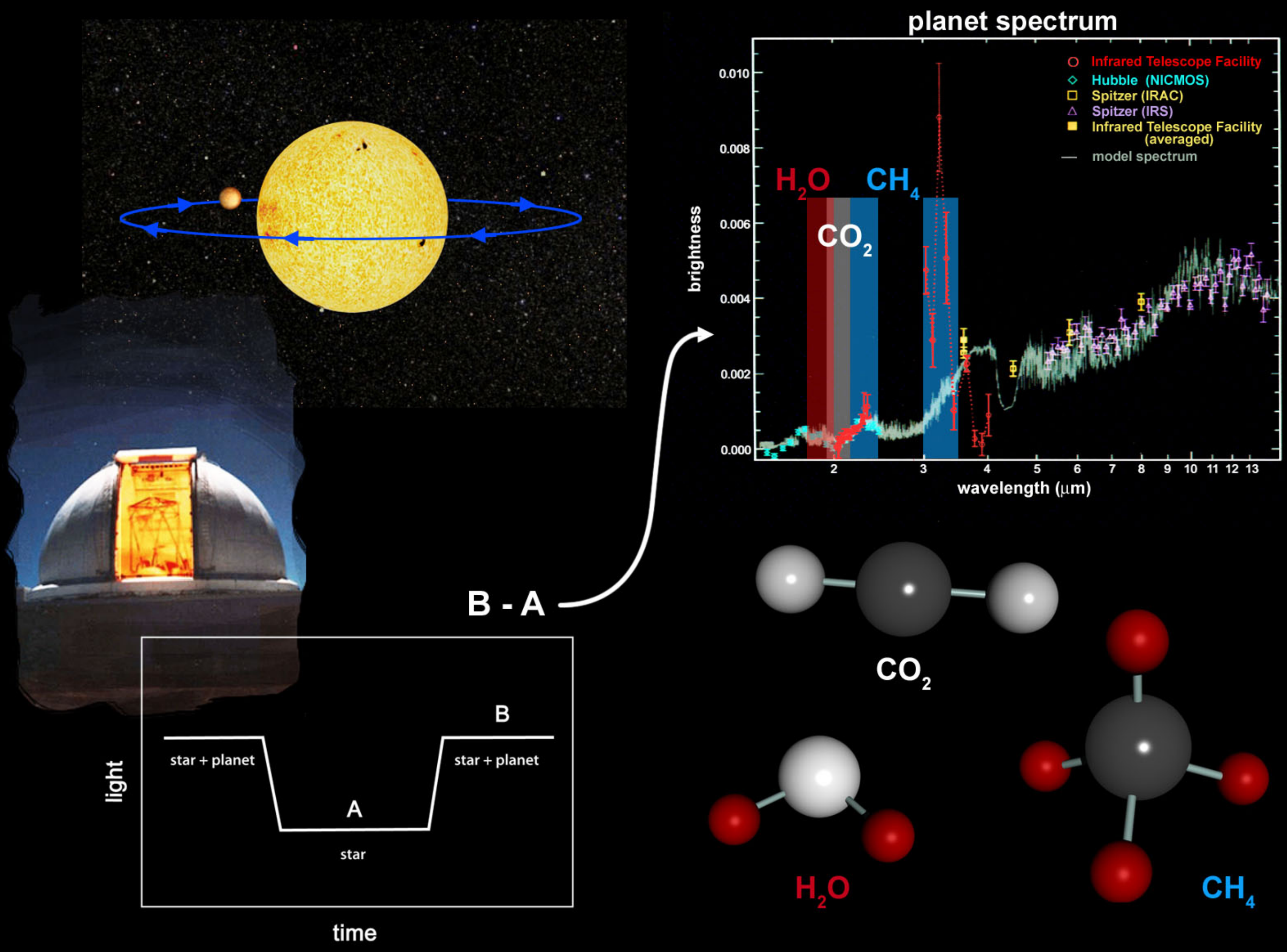
Preliminary results for GD-1 suggest a **preference for CDM over WDM**.

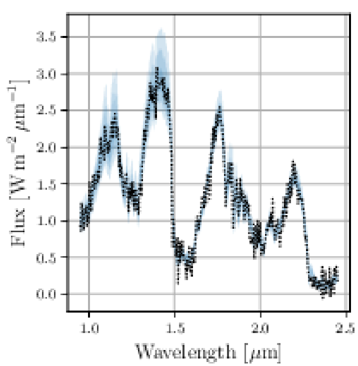
Neural Posterior Estimation (NPE)



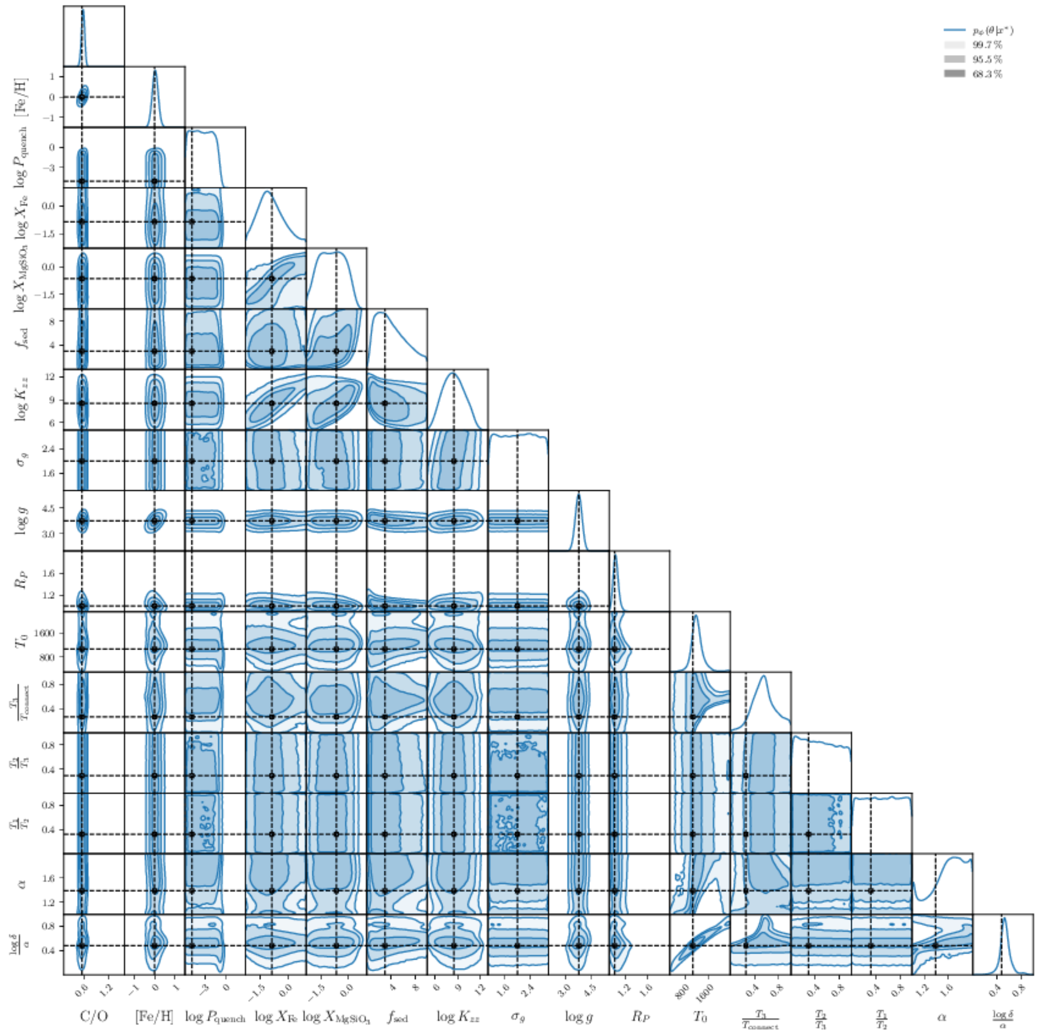
$$\min_{q_\phi} \mathbb{E}_{p(x)} \left[\mathrm{KL}(p(\theta|x) || q_\phi(\theta|x)) \right]$$

Exoplanet atmosphere characterization (Bayesian)





Flux [W m^{-2} μm^{-1}]
Wavelength [μm]

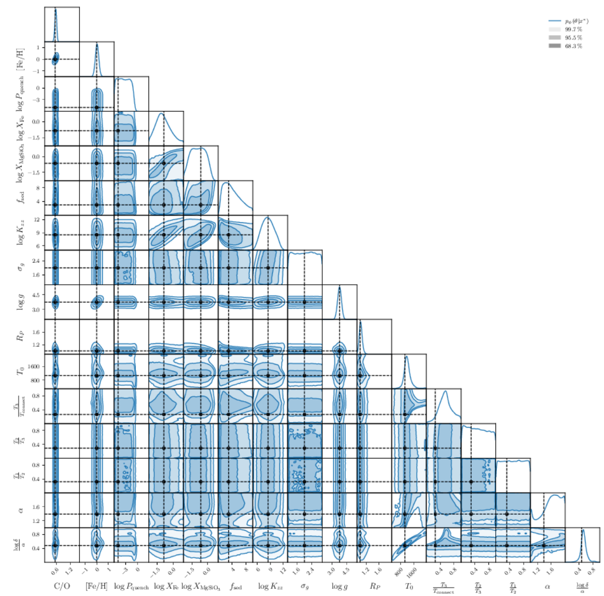


$p_\phi(\theta|x^*)$
99.7 %
95.5 %
68.3 %
C/O
[Fe/H]
$\log P_{\rm quench}$
$\log X_{\rm Fe}$
$\log X_{\rm MgSiO_3}$
$f_{\rm sed}$
$\log K_{zz}$
σ_g
$\log g$
R_P
T_0
$\frac{T_3}{T_{\rm connect}}$
$\frac{T_2}{T_3}$
$\frac{T_1}{T_2}$
α
$\frac{\log \delta}{\alpha}$

Computational faithfulness

$$\hat{p}(\theta|x) = \text{sbi}(p(x|\theta), p(\theta), x)$$

We must make sure our approximate simulation-based inference algorithms can (at least) actually realize faithful inferences on the (expected) observations.

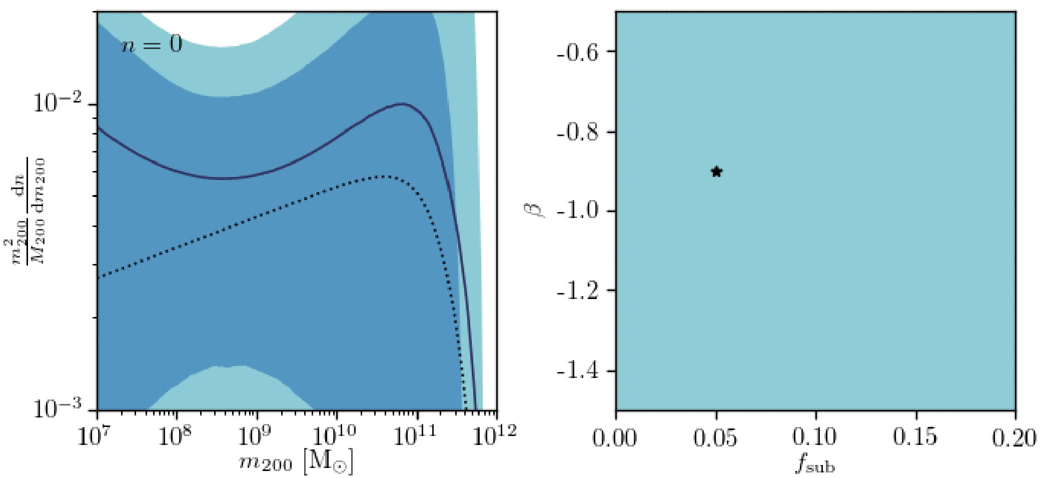


How do we know this is good enough?

Mode convergence:

The maximum a posteriori estimate converges towards the nominal value θ^* for an increasing number of independent and identically distributed observables $x_i \sim p(x|\theta^*)$:

$$
\begin{aligned}
& \lim_{N \to \infty} \arg\max_\theta p(\theta | \{x_i\}_{i=1}^N) \\
= & \lim_{N \to \infty} \arg\max_\theta p(\theta) \prod_{x_i} r(x_i|\theta) = \theta^*
\end{aligned}
$$



—

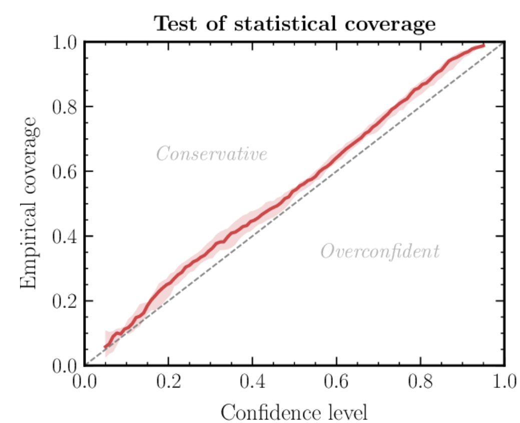
Credits: Brehmer et al, 2019.

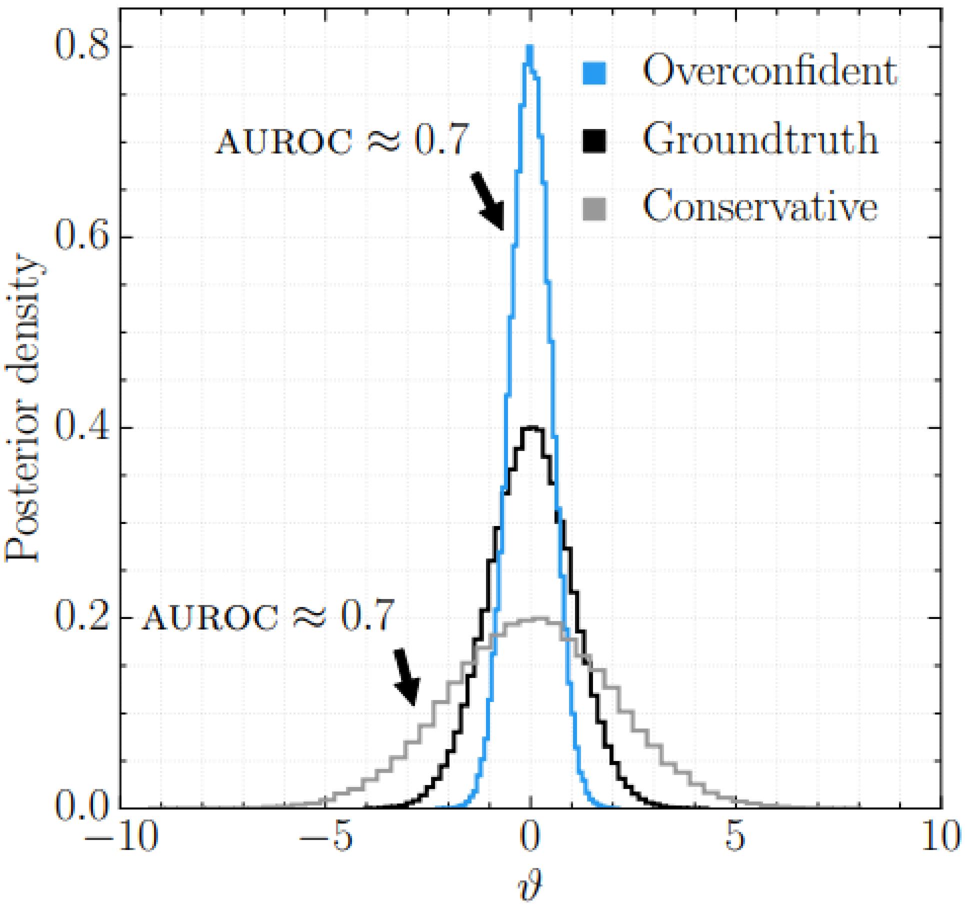
A common observation at the root of several other diagnostics is to check for the **self-consistency** of the Bayesian joint distribution,

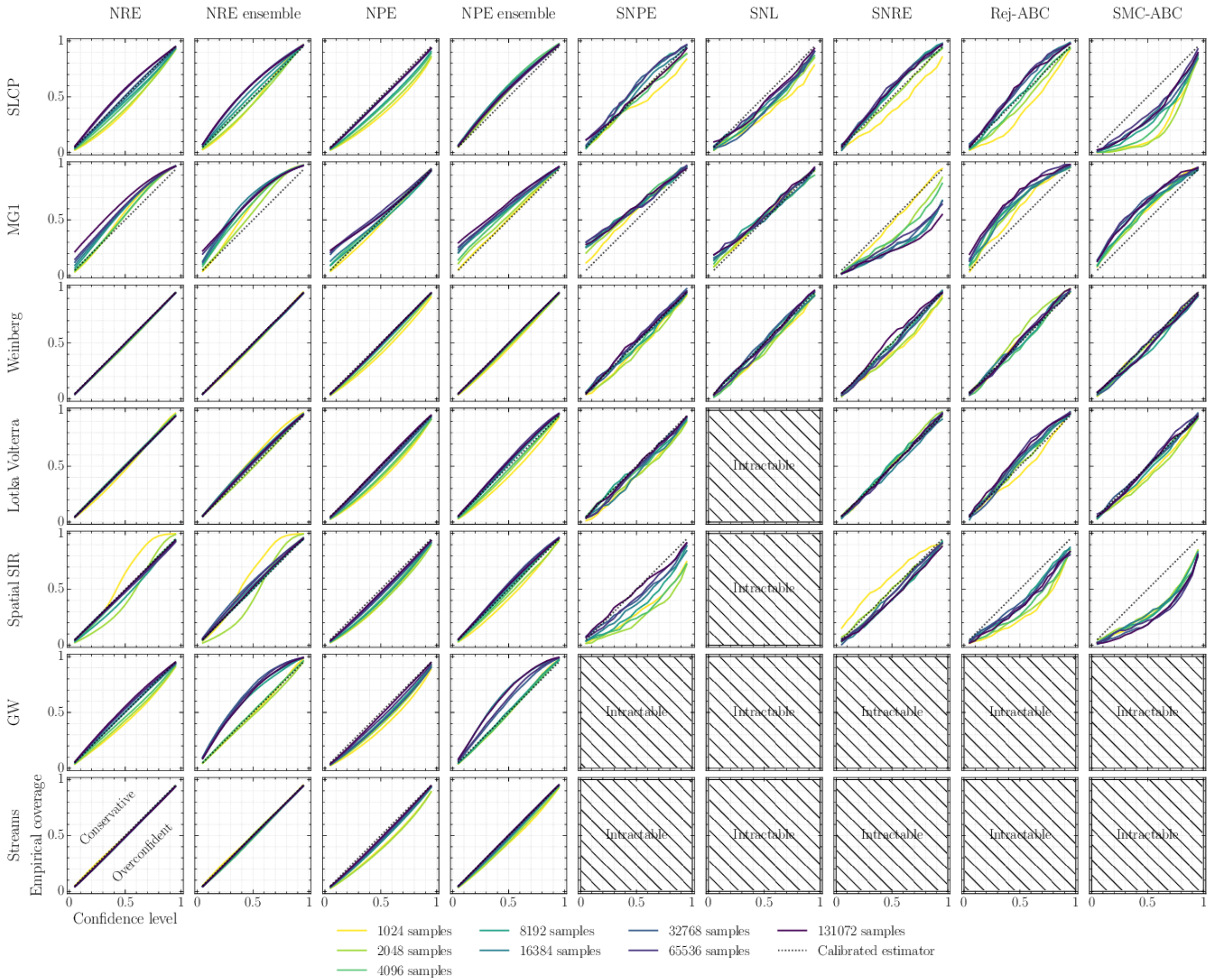
$$p(\theta) = \int p(\theta')p(x|\theta')p(\theta|x)d\theta'\,dx.$$

Coverage diagnostic:

- For $x, \theta \sim p(x, \theta)$, compute the $1-\alpha$ credible interval based on $\hat{p}(\theta|x)$.
- If the fraction of samples for which θ is contained within the interval is larger than the nominal coverage probability $1-\alpha$, then the approximate posterior $\hat{p}(\theta|x)$ has coverage.




Overconfident
Groundtruth
Conservative
AUROC ≈ 0.7
AUROC ≈ 0.7
Posterior density
0.0
0.2
0.4
0.6
0.8
−10
−5
0
5
10
ϑ



—

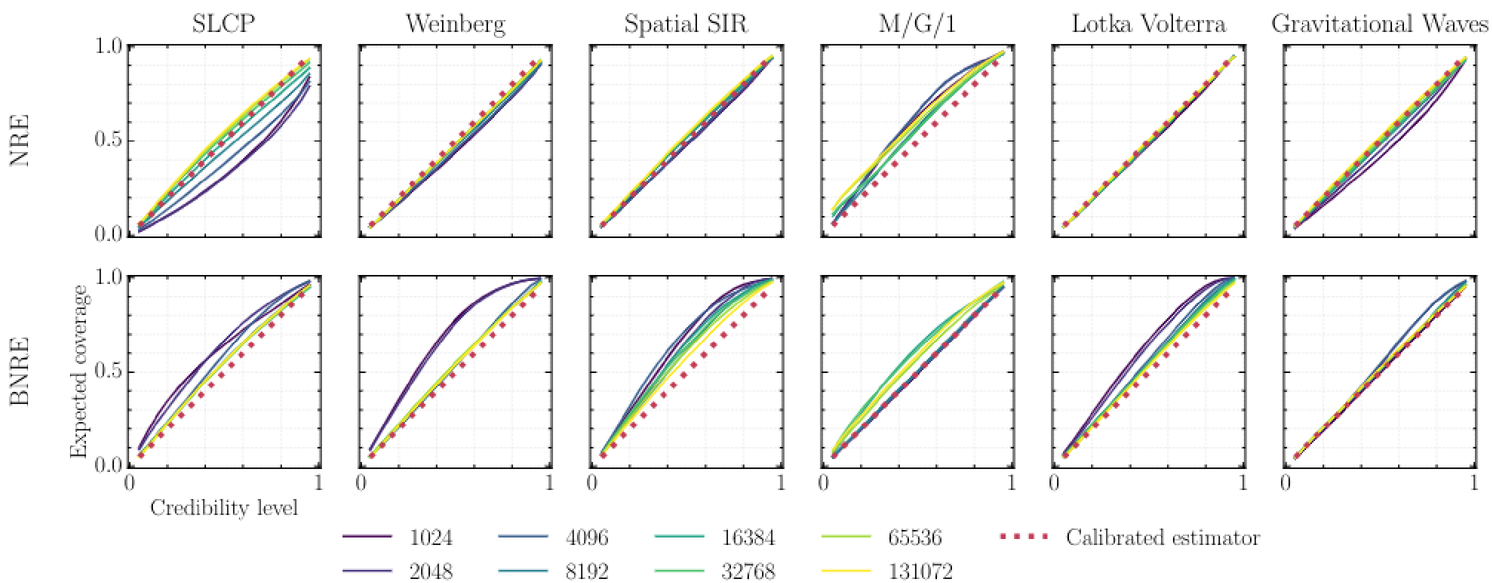
Credits: Hermans et al, 2021.

What if diagnostics fail?

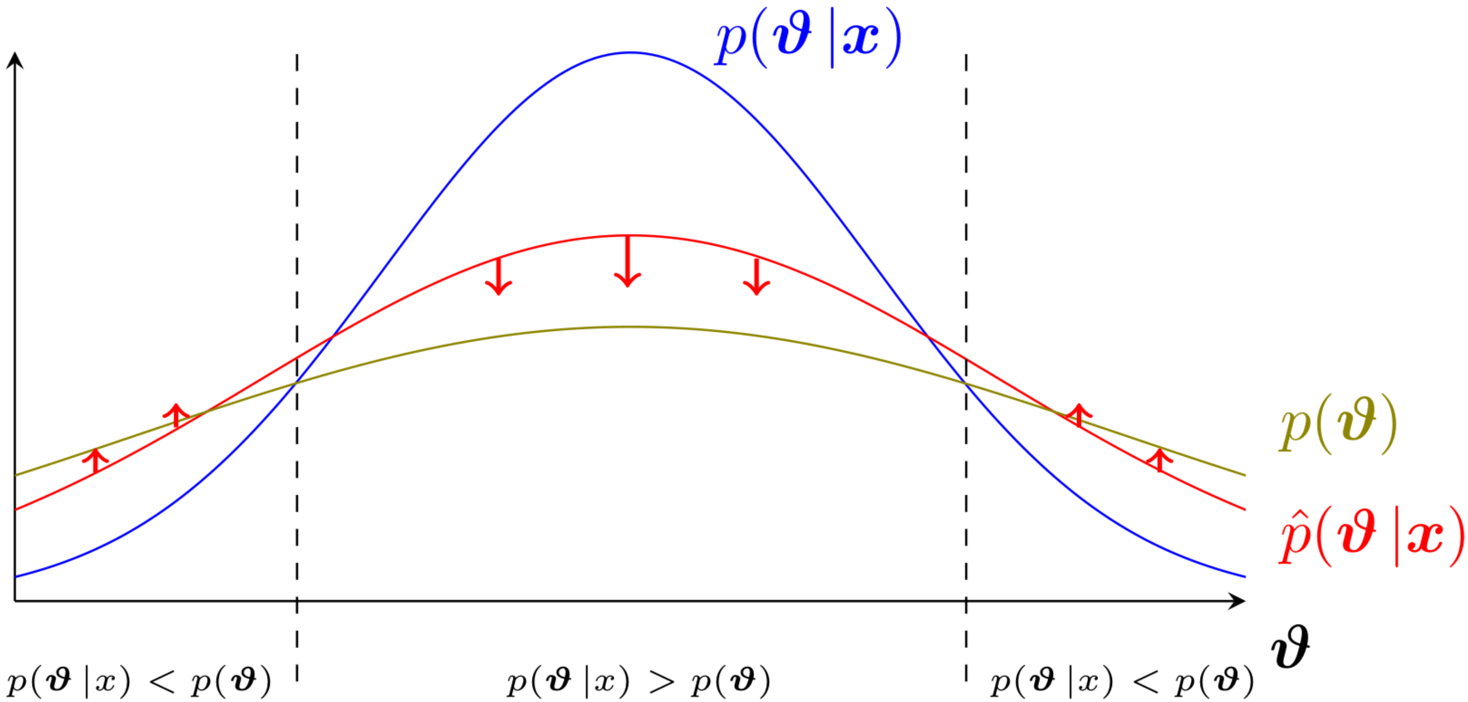
Balanced NRE

Enforce neural ratio estimation to be **conservative** by using binary classifiers $\hat{d}$ that are balanced, i.e. such that

$$\mathbb{E}_{p(\theta,x)}\left[\hat{d}\left(\theta,x\right)\right]=\mathbb{E}_{p(\theta)p(x)}\left[1-\hat{d}\left(\theta,x\right)\right].$$



Credits: Delaunoy et al, 2022.



—

Credits: Delaunoy et al, 2022.

Summary

Simulation-based inference is a major evolution in the statistical capabilities for science, enabled by advances in machine learning.

Need to reliably and efficiently evaluate the quality of the posterior approximations.

The end.