Towards reliable simulation-based inference

CAp-RFIAP joint conference
July 7, 2022

Gilles Louppe
g.louppe@uliege.be

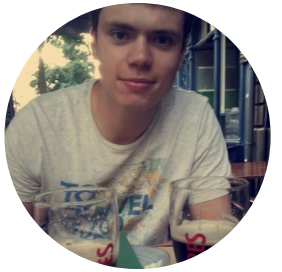
LIÈGE université

Kyle Cranmer

Johann Brehmer

Joeri Hermans

Antoine Wehenkel



Norman Marlier

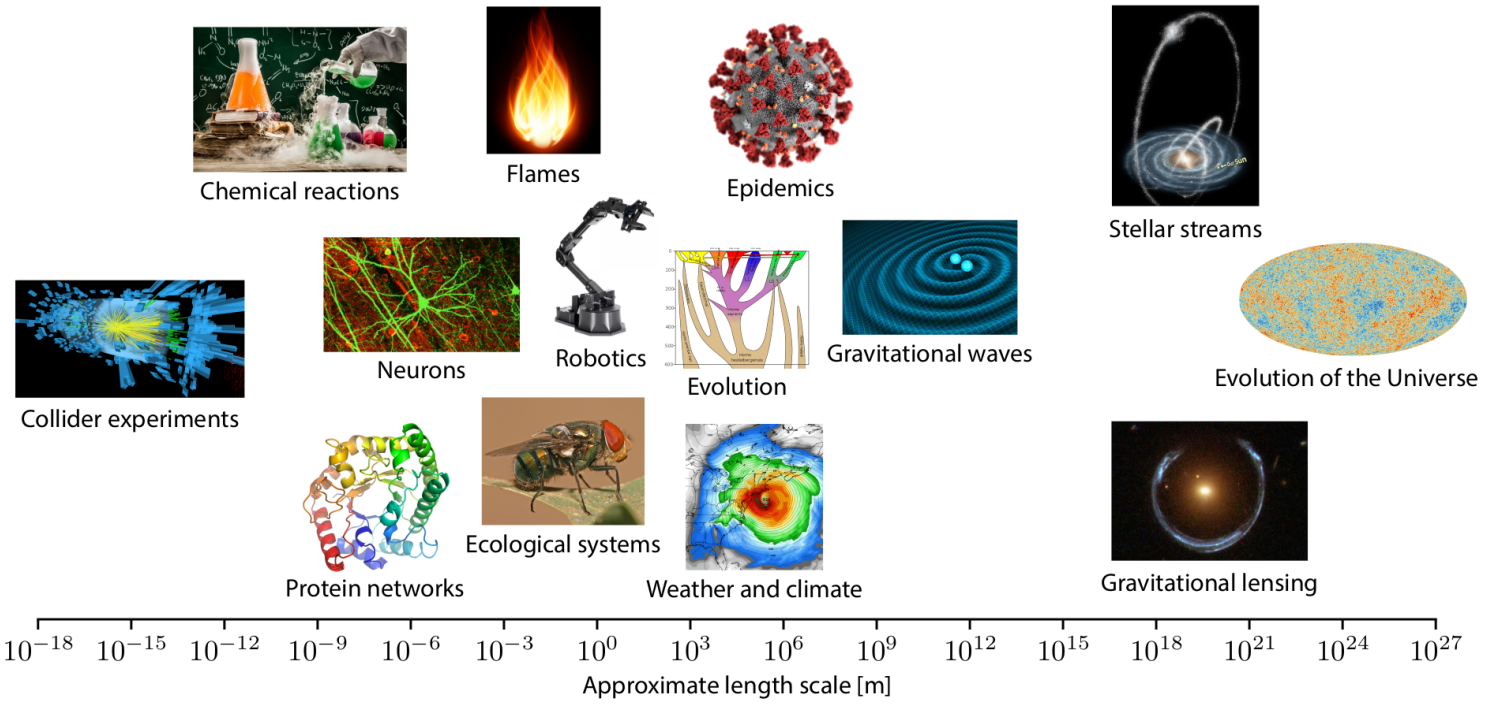
Siddharth Mishra-Sharma

Christoph Weniger

Arnaud Delaunoy

Malavika Vasist

Francois Rozet


Chemical reactions
Flames
Epidemics
Stellar streams
Collider experiments
Neurons
Robotics
Evolution
Gravitational waves
Evolution of the Universe
Protein networks
Ecological systems
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]

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

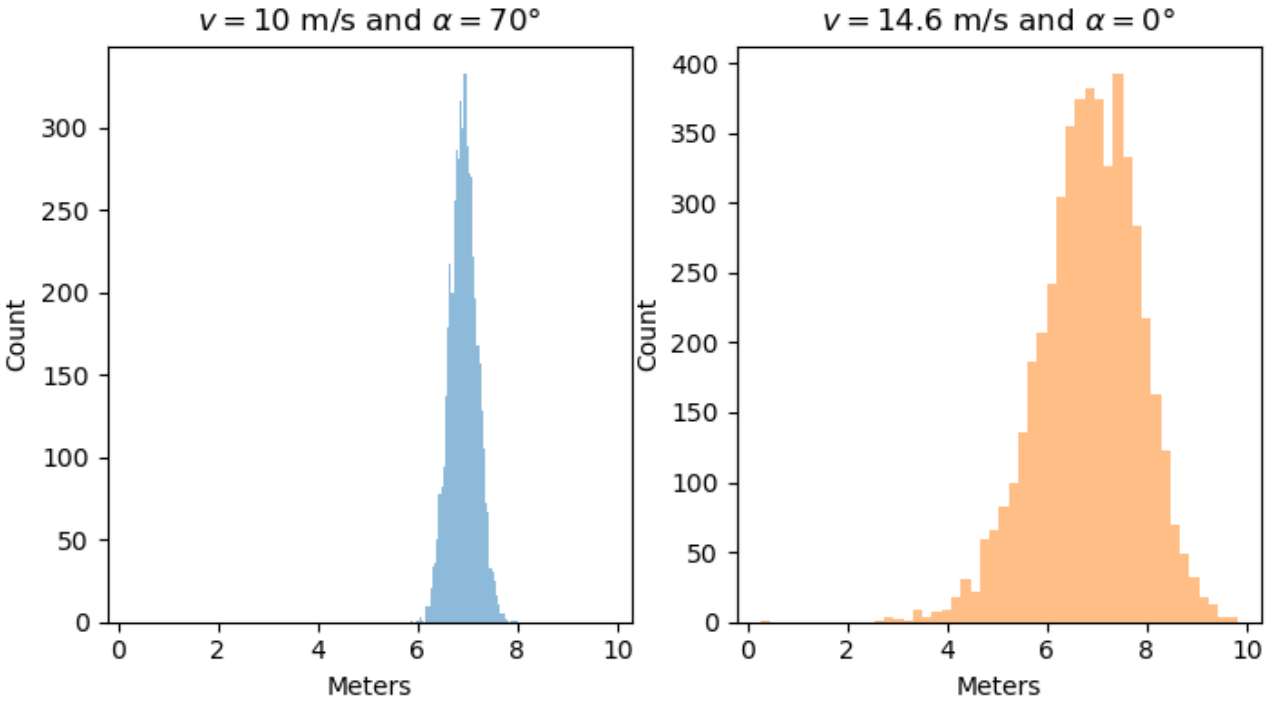
$$\frac{dx}{dt} = v_x,\ \ \frac{dy}{dt} = v_y, \frac{dv_y}{dt} = -G.$$

```
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 plausible given the observation x?

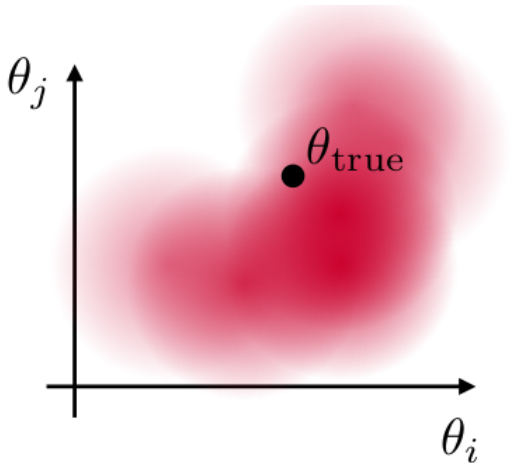
Bayesian inference

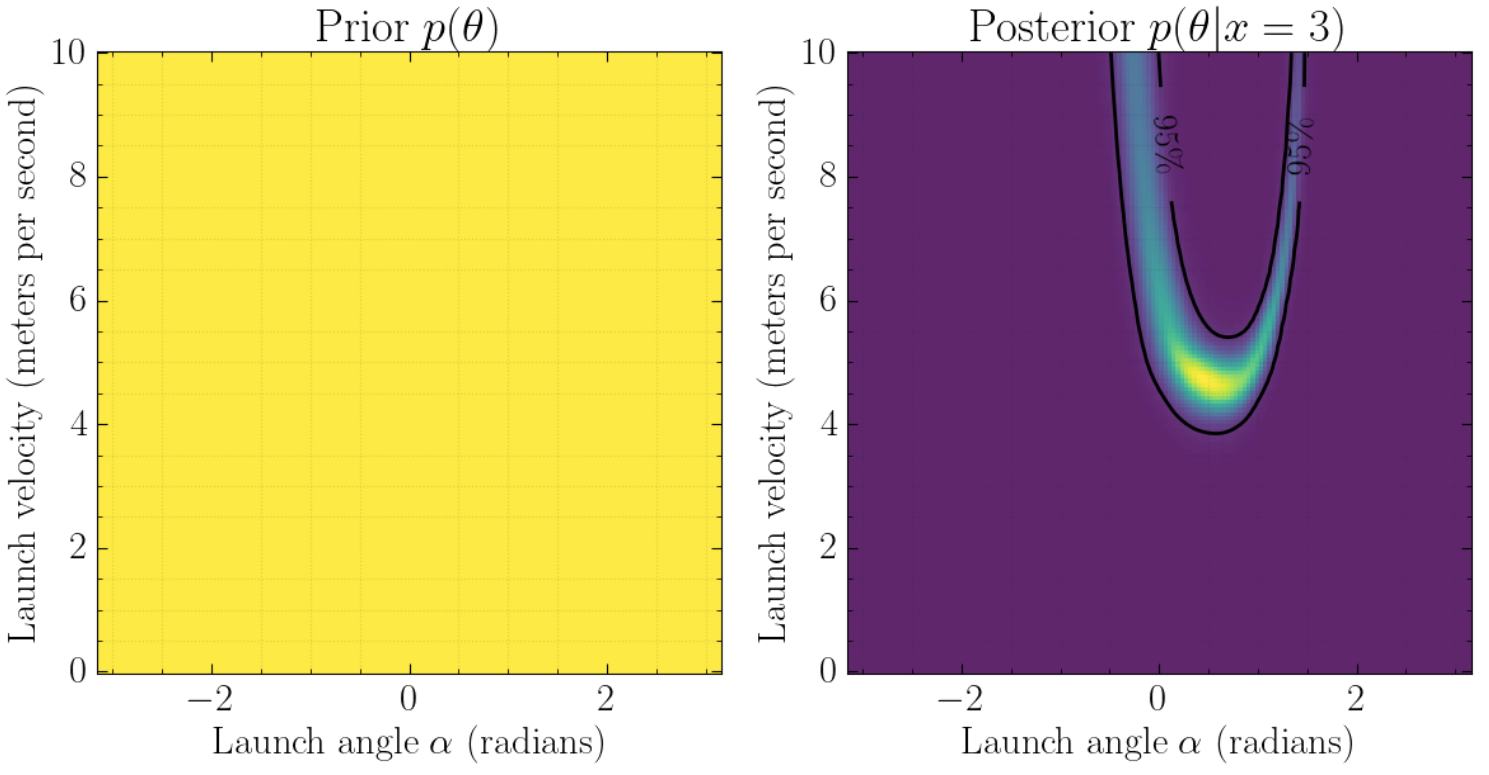
Start with

- a simulator that can generate N samples $x_i \sim p(x_i|\theta_i)$,
- a prior model $p(\theta)$,
- observed data $x_{\text{obs}} \sim p(x_{\text{obs}}|\theta_{\text{true}})$.

Then, estimate the posterior

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




Prior $p(\theta)$
Launch velocity (meters per second)
Launch angle α (radians)
0
2
4
6
8
10
−2
0
2

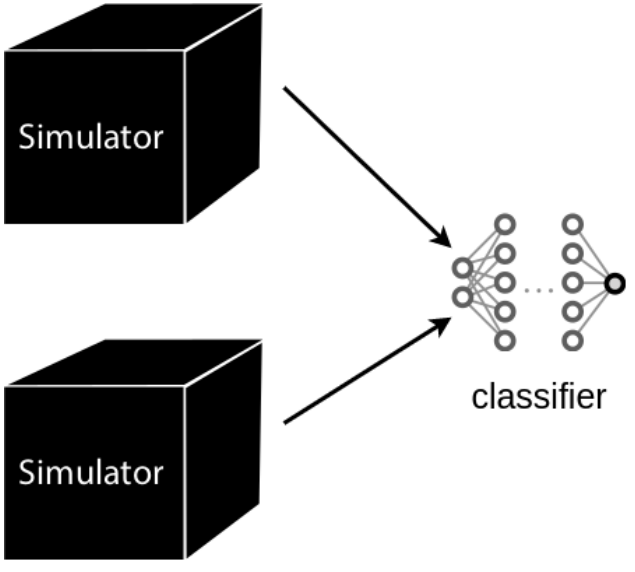
Posterior $p(\theta|x = 3)$
95%
95%
Launch velocity (meters per second)
Launch angle α (radians)
0
2
4
6
8
10
−2
0
2

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:

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

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



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

Credits: Cranmer et al, 2015; Hermans et al, 2020.

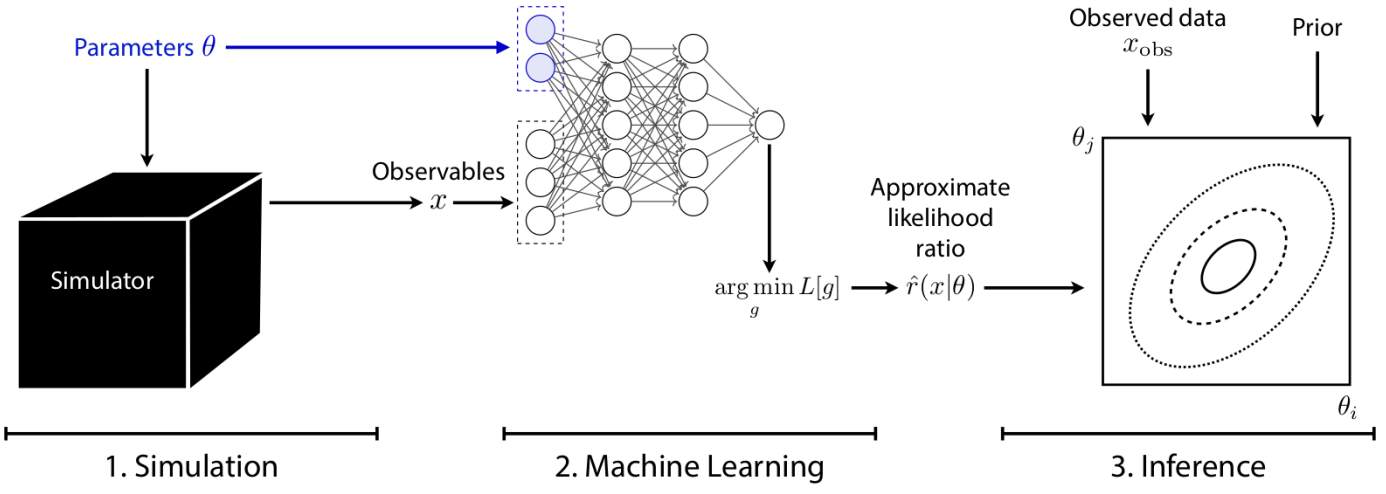
The solution d found after training approximates the optimal classifier

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

Therefore,

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

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



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

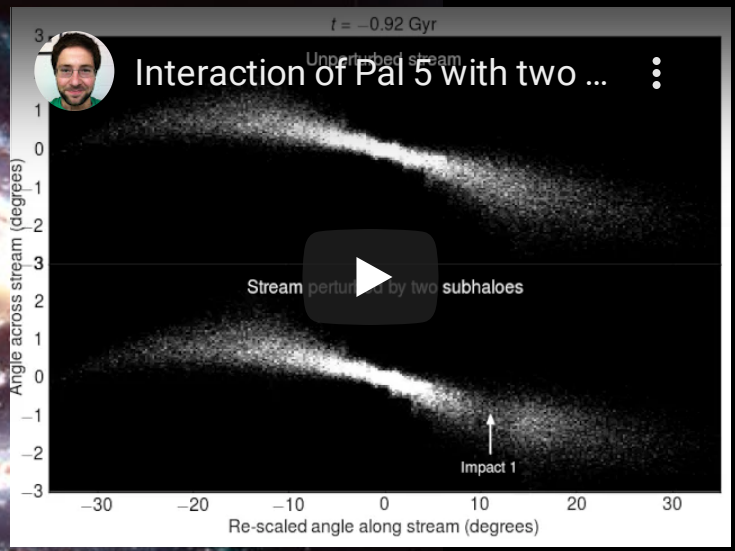
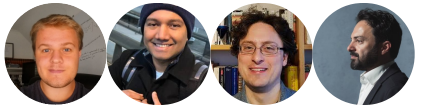
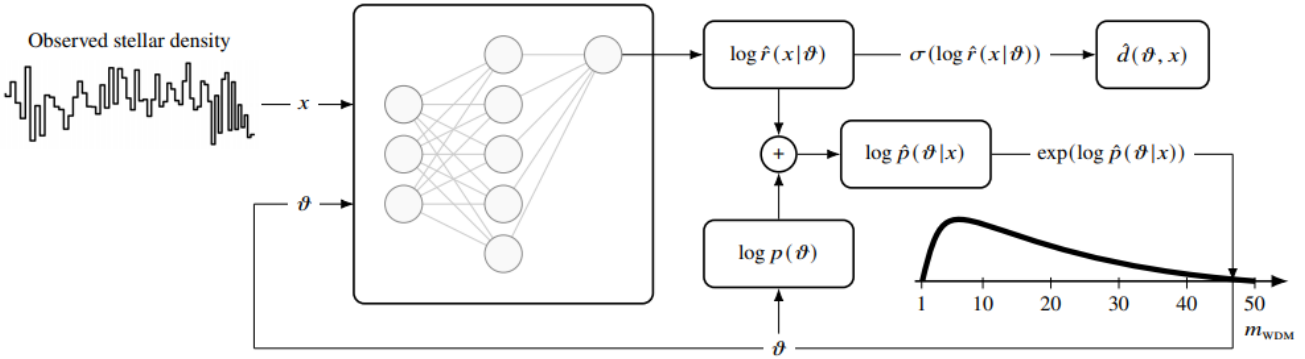
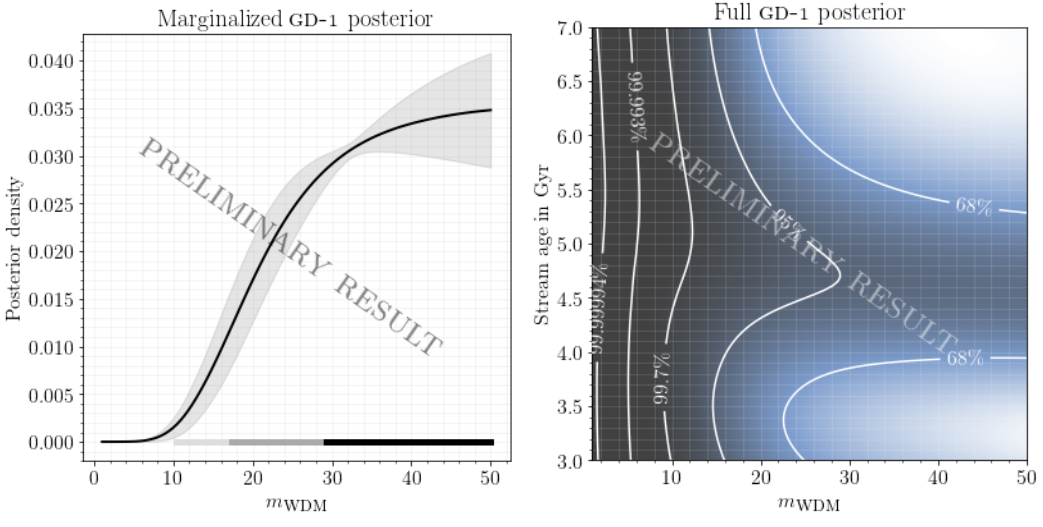


Image credits: C. Bickel/Science; D. Erkal.







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

—

Credits: Hermans et al, 2021.

Neural Posterior Estimation (NPE)

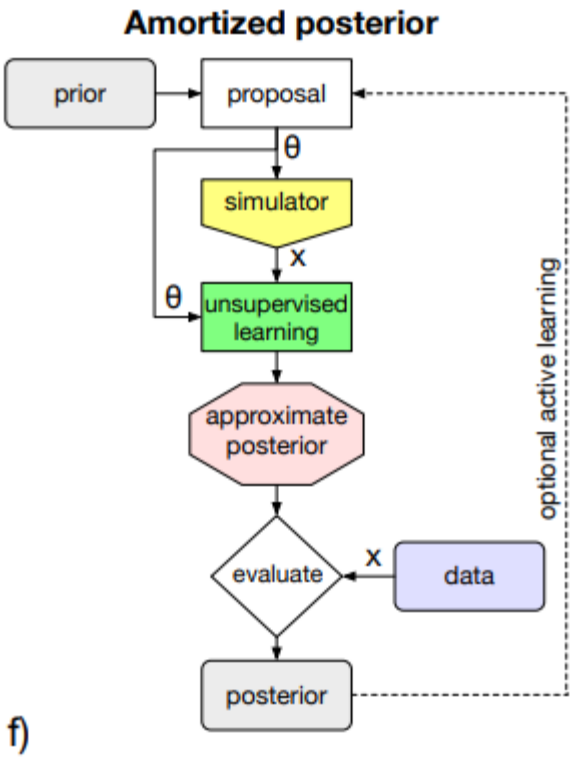
Use variational inference to directly estimate the posterior:

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

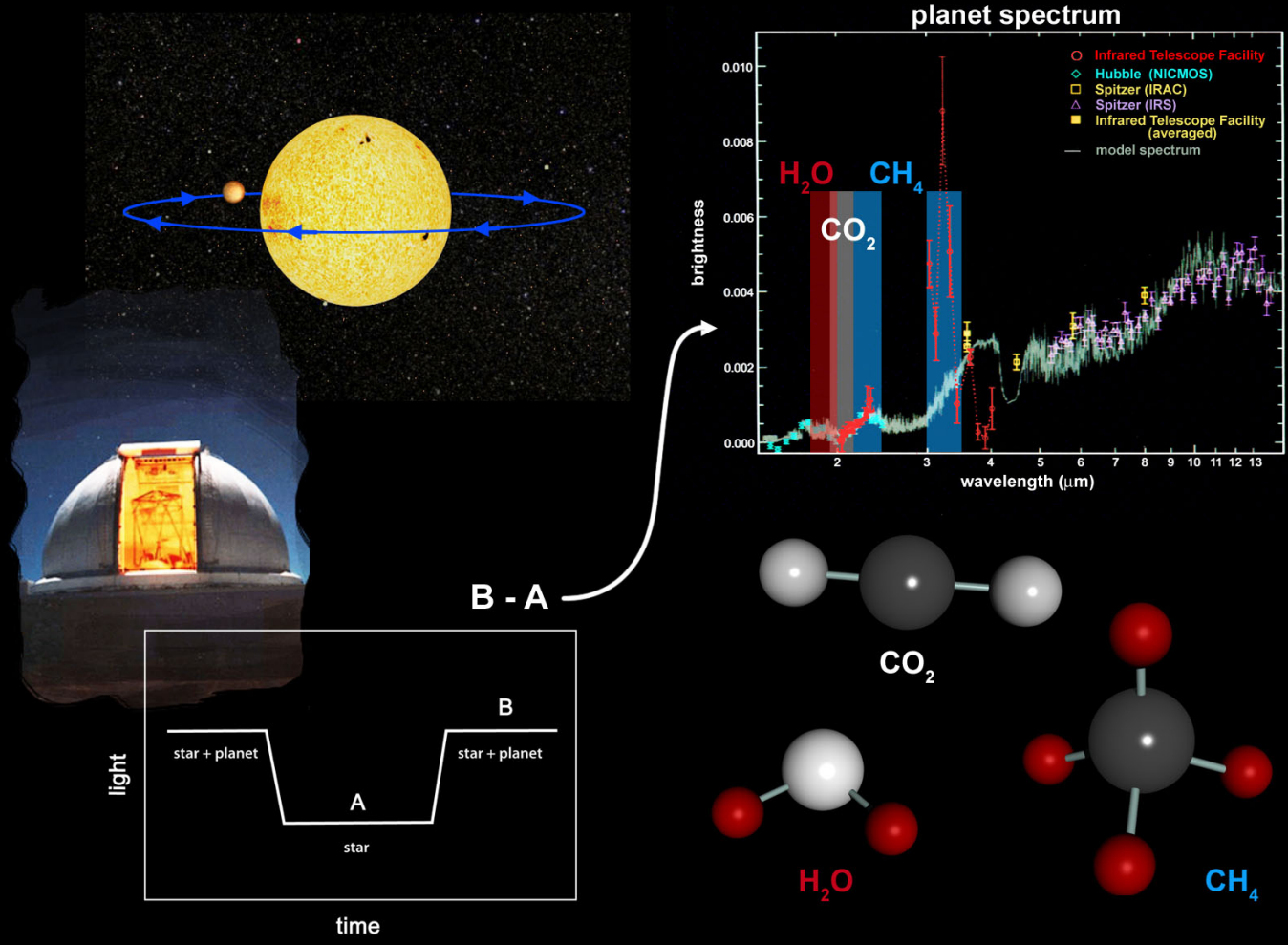
$$= \min_{q_\phi} \mathbb{E}_{p(x)} \mathbb{E}_{p(\theta|x)} \left[\log \frac{p(\theta|x)}{q_\phi(\theta|x)} \right]$$

$$= \max_{q_\phi} \mathbb{E}_{p(x,\theta)} \left[\log q_\phi(\theta|x) \right]$$

where q_ϕ is a neural density estimator, such as a normalizing flow.

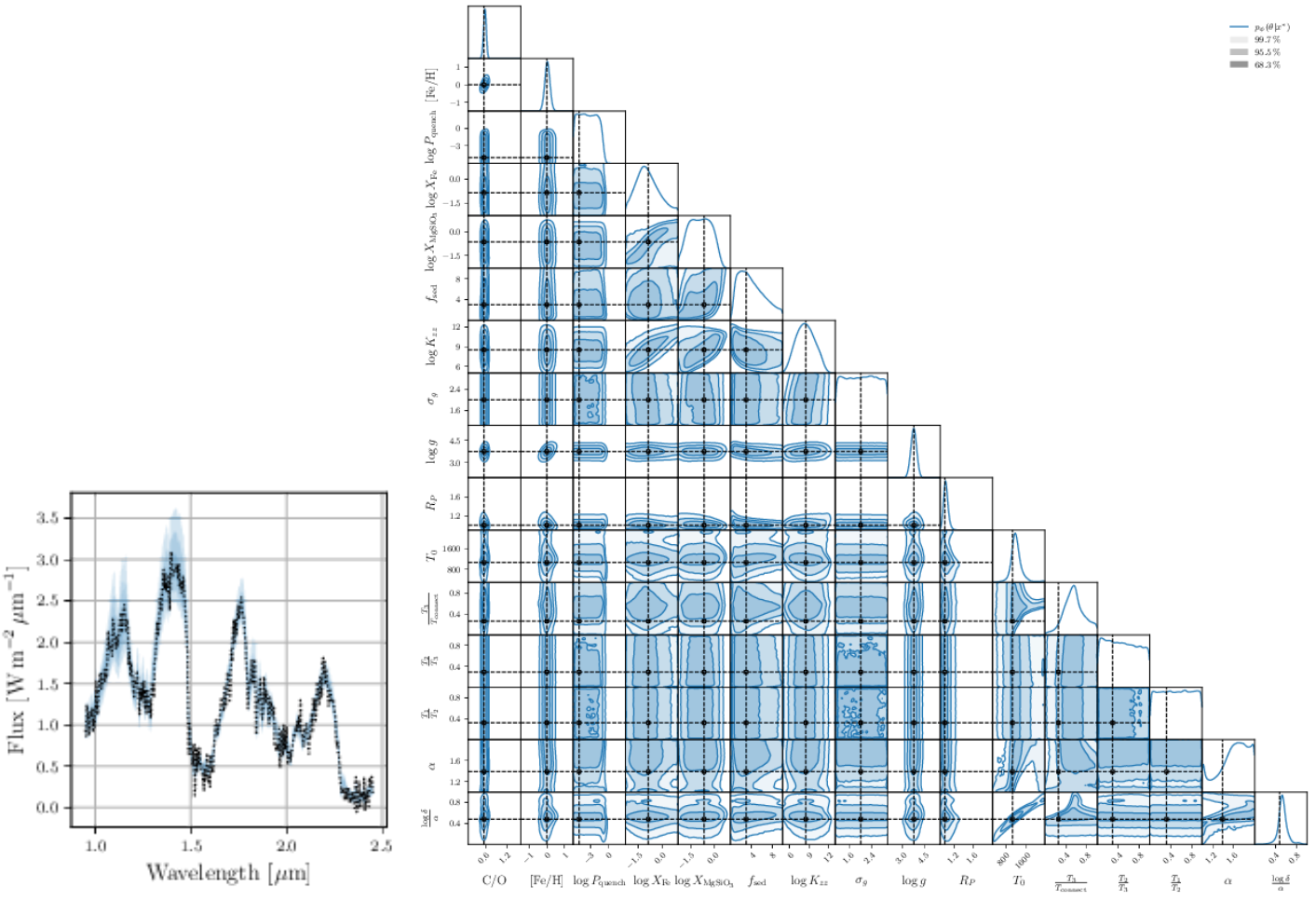


Exoplanet atmosphere characterization



—

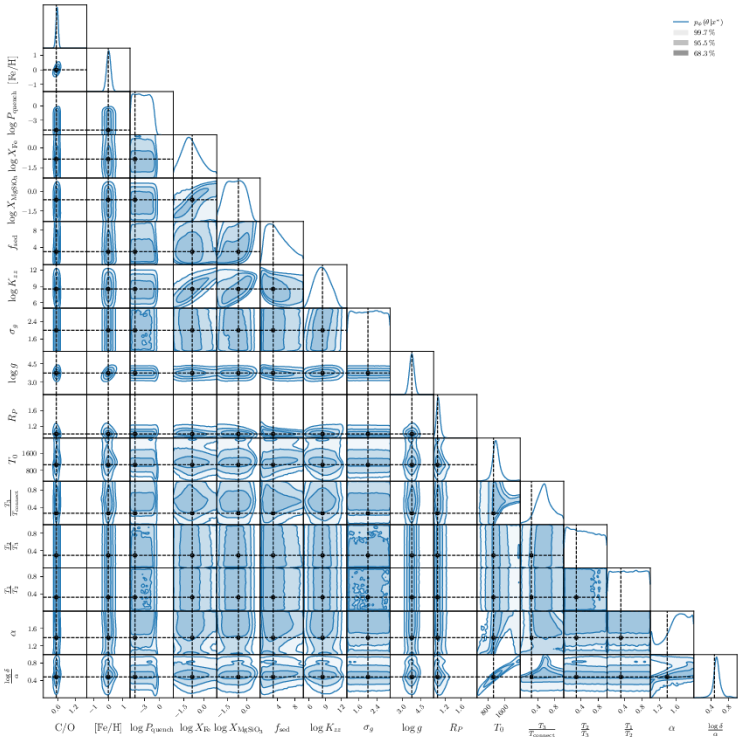
Credits: NSA/JPL-Caltech, 2010.


Flux [W m^{-2} μm^{-1}]
Wavelength [μm]
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
$\frac{T_3}{T_{\rm conn}}$
$\frac{T_2}{T_3}$
$\frac{T_1}{T_2}$
α
$\frac{\log \delta}{\alpha}$

Computational faithfulness

$$\hat{p}\left(\theta|x\right) = \mathrm{sbi}(p(x|\theta), p(\theta), x)$$

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

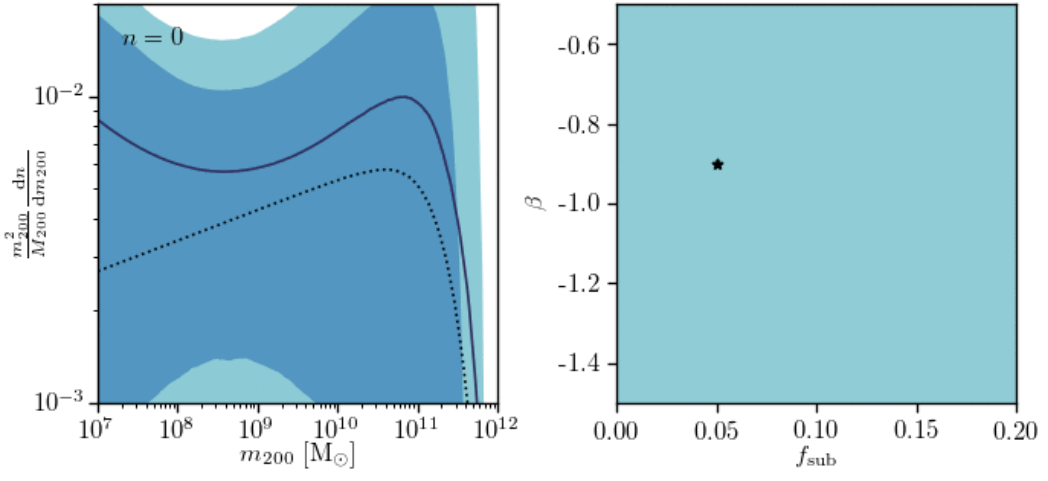


How do we know this is good enough?

Mode convergence:

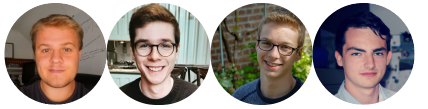
The maximum a posteriori estimate converges towards the nominal value θ^* for an increasing number of independent and identically distributed observables $x_i \sim p(x|\theta^*)$:

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



—

Credits: Brehmer et al, 2019.

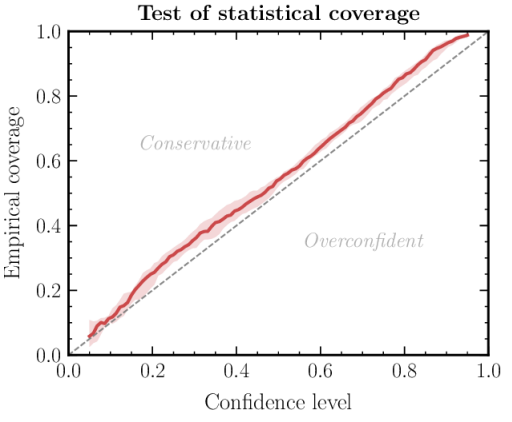


A common observation at the root of several other diagnostics is to check for the **self-consistency** of the Bayesian joint distribution,

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

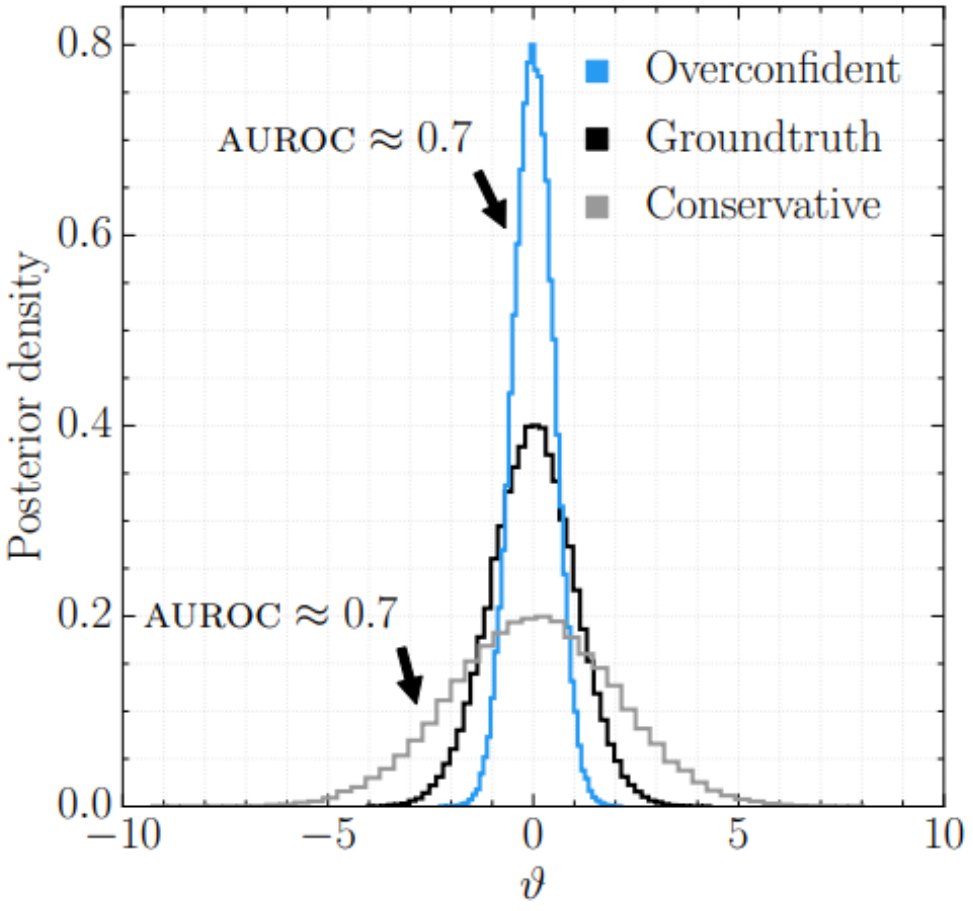
Coverage diagnostic:

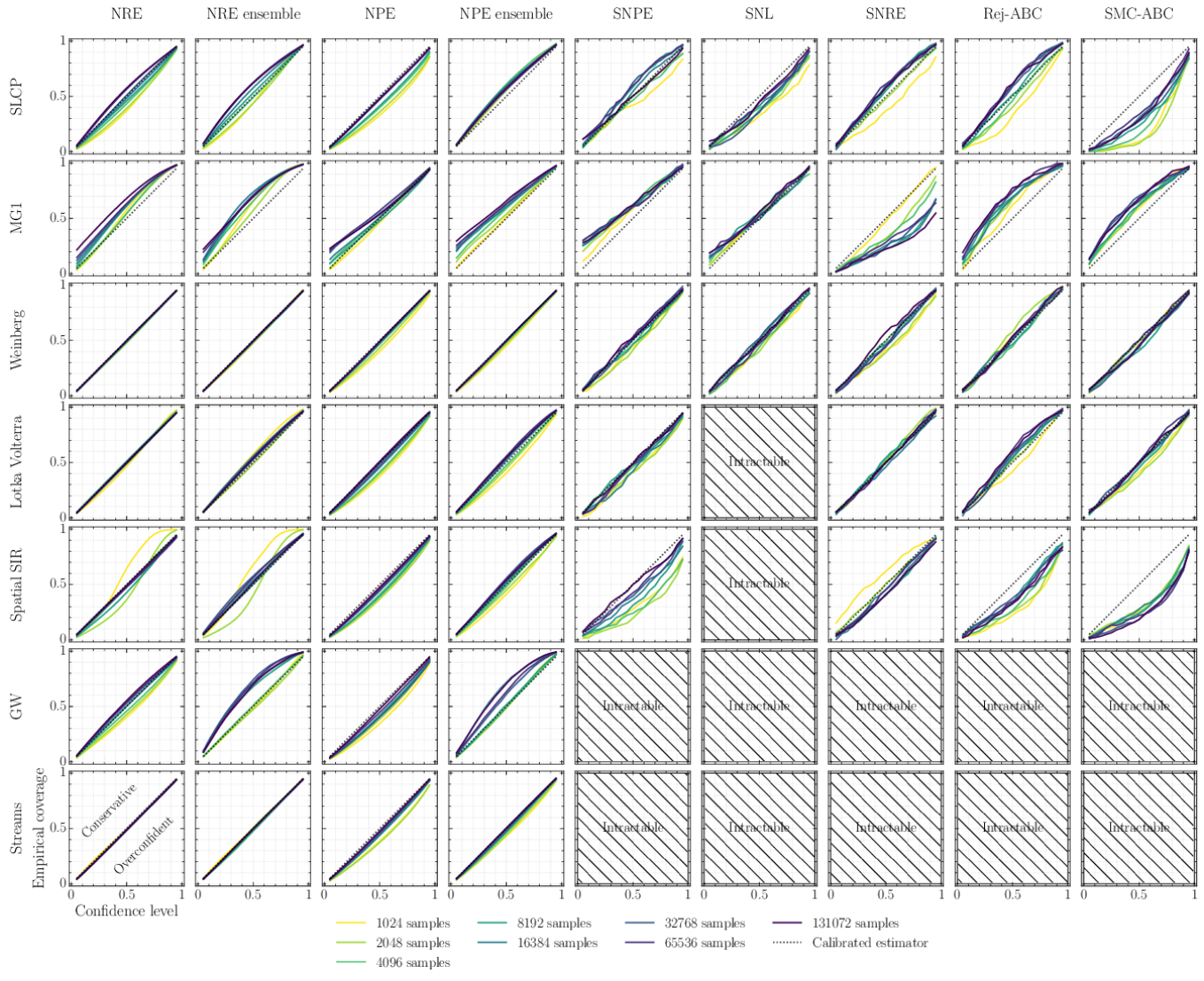
- For $x, \theta \sim p(x, \theta)$, compute the $1 - \alpha$ credible interval based on $\hat{p}\left(\theta|x\right)$.
- If the fraction of samples for which θ is contained within the interval is larger than the nominal coverage probability $1 - \alpha$, then the approximate posterior $\hat{p}\left(\theta|x\right)$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



—

Credits: Hermans et al, 2021; Siddharth Mishra-Sharma, 2021.

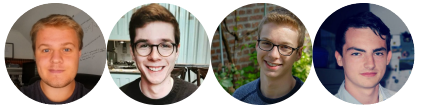

Overconfident
Groundtruth
Conservative
AUROC ≈ 0.7
AUROC ≈ 0.7
Posterior density
0.0
0.2
0.4
0.6
0.8
−10
−5
0
5
10
ϑ



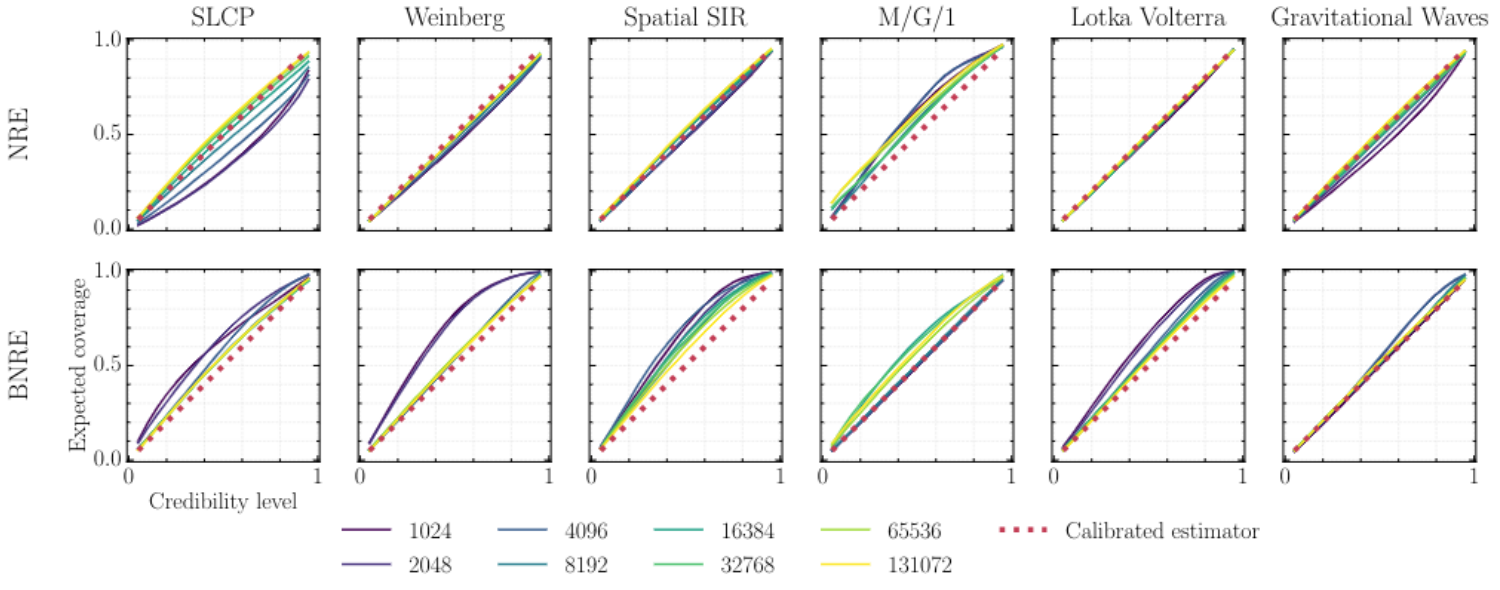
Credits: Hermans et al, 2021.

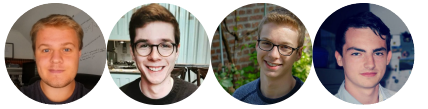
What if diagnostics fail?

Balanced NRE



Neural ratio estimation can be forced to be more **conservative**, hence increasing the reliability of the approximate posteriors and reducing the risk of false inferences.





Definition

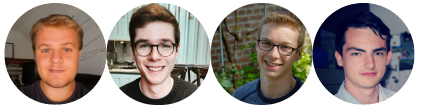
A binary classifier $\hat{d}$ is balanced if

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

Theorems 1 and 2

Any balanced classifier $\hat{d}$ satisfies

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



Ideally*, whenever $\hat{d}\left(\theta, x\right) \leq d(\theta, x)$ then

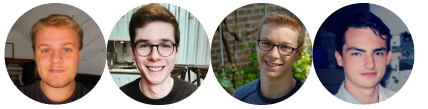
$$\frac{\hat{d}\left(\theta, x\right)}{1 - \hat{d}\left(\theta, x\right)} \leq \frac{d(\theta, x)}{1 - d(\theta, x)},$$

which is equivalent to

$$\begin{aligned} & \hat{r}\left(x|\theta\right) \leq r(x|\theta) \\ \Leftrightarrow\ & p(\theta)\hat{r}\left(x|\theta\right) \leq p(\theta) r(x|\theta) \\ \Leftrightarrow\ & \hat{p}\left(\theta|x\right) \leq p(\theta|x). \end{aligned}$$

—

*: The balancing condition does not guarantee that $\hat{d}\left(\theta, x\right) \leq d(\theta, x)$ for all θ, x :-(



Algorithm 1 Training algorithm for Balanced Neural Ratio Estimation (BNRE).

Inputs: Implicit generative model $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ (simulator) and prior $p(\boldsymbol{\vartheta})$
Outputs: Approximate classifier $\hat{d}_\psi(\boldsymbol{\vartheta}, \boldsymbol{x})$ parameterized by ψ
hyper-parameters: Balancing condition strength λ (default = 100) and batch-size n

repeat

Sample data from the joint $\{\boldsymbol{\vartheta}_i,\ \boldsymbol{x}_i \sim p(\boldsymbol{\vartheta}, \boldsymbol{x}),\ y_i = 1\}_{i=1}^{n/2}$
Sample data from the marginals $\{\boldsymbol{\vartheta}_i,\ \boldsymbol{x}_i \sim p(\boldsymbol{\vartheta})p(\boldsymbol{x}),\ y_i = 0\}_{i=n/2+1}^{n}$

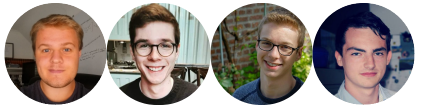
$$\mathcal{L}[\hat{d}_\psi] = -\frac{1}{n}\sum_{i=1}^{n} y_i \log \hat{d}_\psi(\boldsymbol{\vartheta}_i, \boldsymbol{x}_i) + (1 - y_i)\log(1 - \hat{d}_\psi(\boldsymbol{\vartheta}_i, \boldsymbol{x}_i))$$

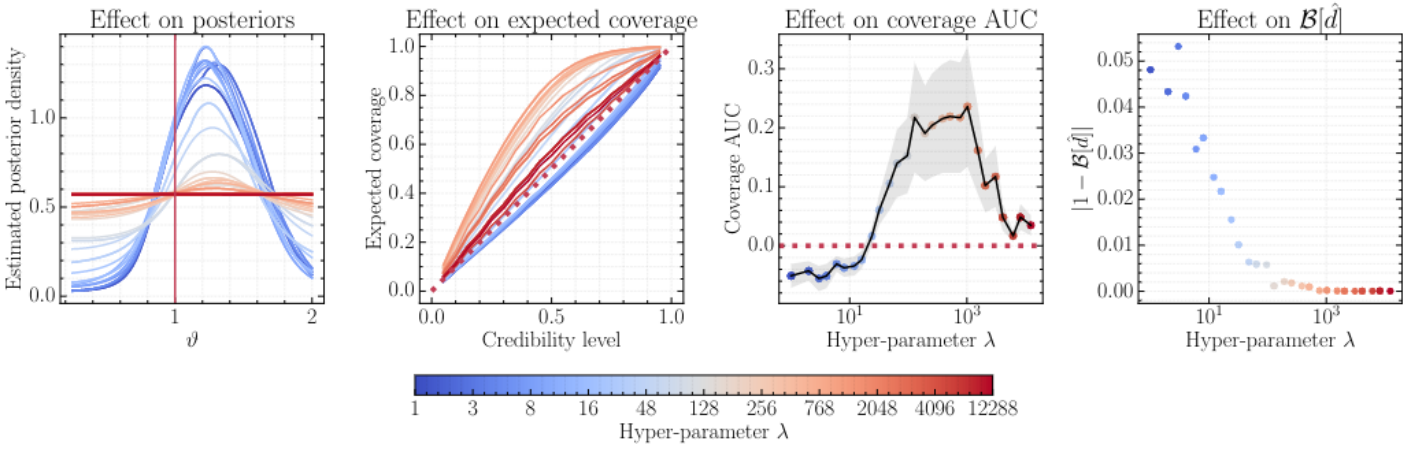
$$\mathcal{B}[\hat{d}_\psi] = \frac{2}{n}\sum_{i=1}^{n/2} \hat{d}_\psi(\boldsymbol{\vartheta}_i, \boldsymbol{x}_i) + \frac{2}{n}\sum_{i=n/2+1}^{n} \hat{d}_\psi(\boldsymbol{\vartheta}_i, \boldsymbol{x}_i)$$

$$\psi = \texttt{minimizer\_step}(\texttt{params=}\psi,\ \texttt{loss=}\mathcal{L}[\hat{d}_\psi] + \lambda(\mathcal{B}[\hat{d}_\psi] - 1)^2)$$

until convergence

return $\hat{d}_\psi(\boldsymbol{\vartheta}, \boldsymbol{x})$.




Effect on posteriors
Estimated posterior density
ϑ
Effect on expected coverage
Expected coverage
Credibility level
Effect on coverage AUC
Coverage AUC
Hyper-parameter λ
Effect on $\mathcal{B}[\hat{d}]$
$|1 - \mathcal{B}[\hat{d}]|$
Hyper-parameter λ
1 3 8 16 48 128 256 768 2048 4096 12288
Hyper-parameter λ

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.