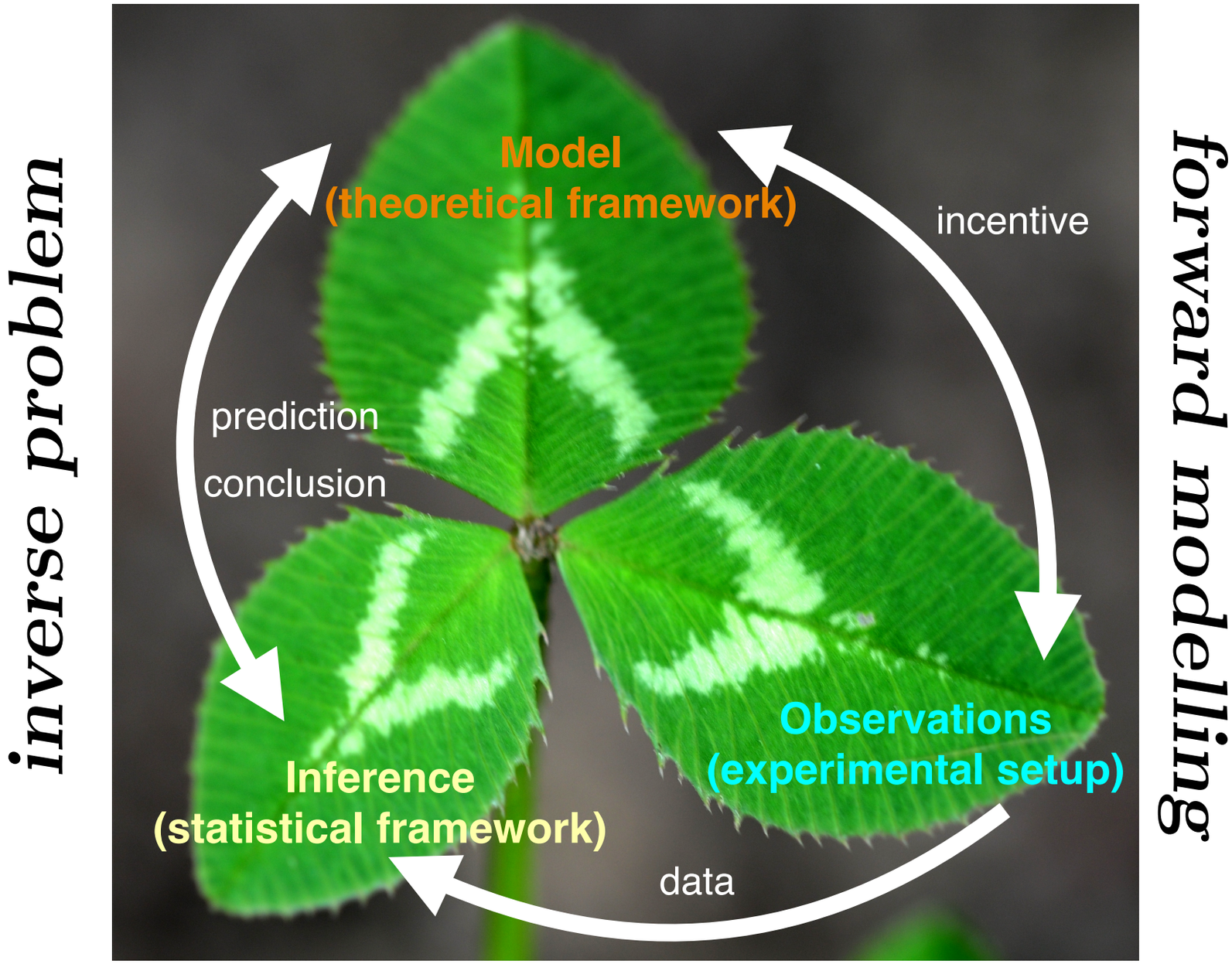
If you can simulate it, you can infer it: simulation-based inference in γ-ray astronomy

Christopher Eckner (christopher.eckner@iac.es)
UNDARK Postdoctoral Researcher
Instituto de Astrofísica de Canarias, Tenerife, Spain



2nd SMASHing Workshop | University of Nova Gorica
22nd of October 2026 | Vipava, Slovenia

The holy trinity of scientific progress

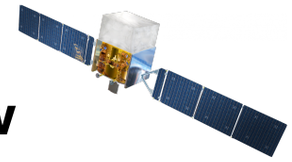


Lab experiments:
Complexity of models and impact of systematics controlled with experimental design.

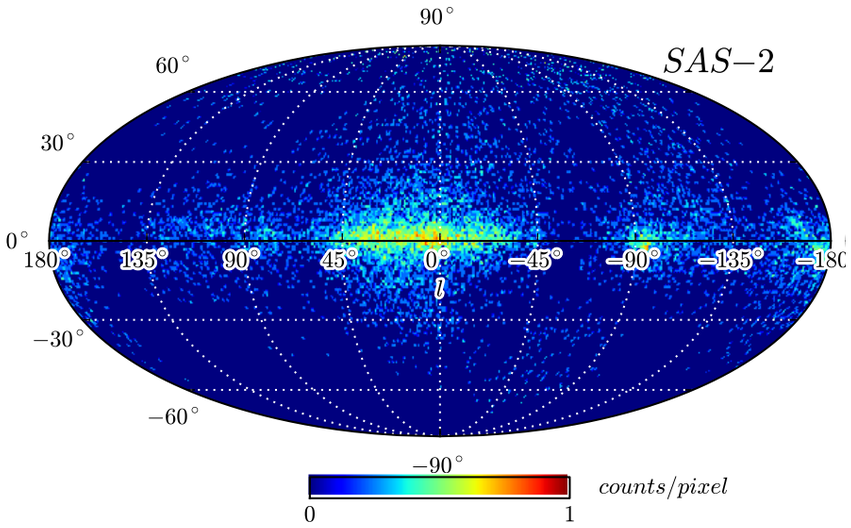
Cosmology/astrophysics:
There is exactly one universe with known and unknown physics. We have to deal with the data we get.

Fundamental physics with observational data is hard

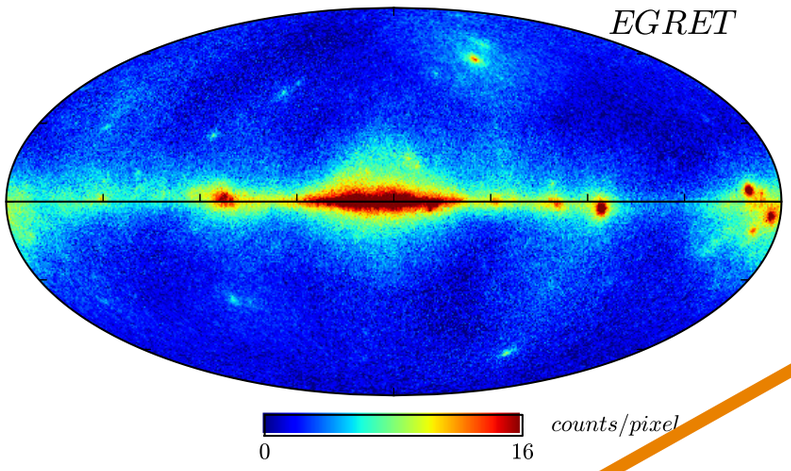
The high-energy gamma-ray sky seen over the decades (space-borne telescopes).



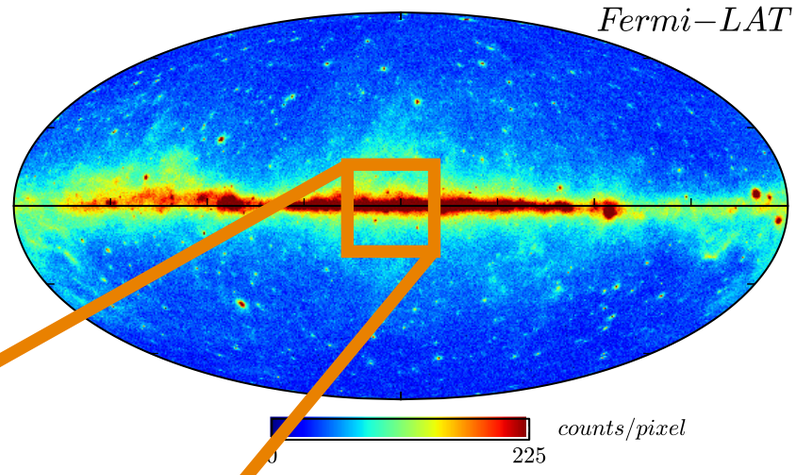
1972-73



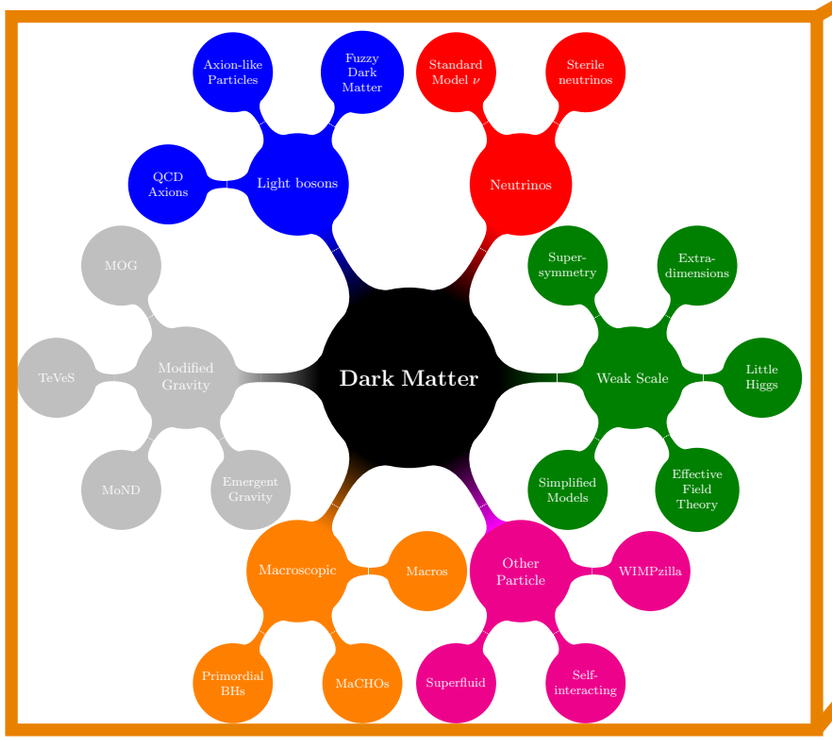
1991-2000



2008-now



[Fermi-LAT collaboration, ApJS 223 (2016) 2]



[Bertone, Tait, *Nature* 562 (2018) 7725]

Signatures of fundamental physics are potentially hiding there!
→ How to deal with the **complexity** of all the astrophysics?
There is a lot to model …

What does implicit likelihood mean?

A way to handle the complexity is by embracing it! Can’t we just solve the inverse problem by simulating the full physics and learn the likelihood implicitly from simulations?

What does implicit likelihood mean?

A way to handle the complexity is by embracing it! Can't we just solve the inverse problem by simulating the full physics and learn the likelihood implicitly from simulations?

Yes!

From Bayes theorem to simulations

Let's approach this question from the perspective of Bayesian statistics: a mathematical framework for continuously updating your existing beliefs about the probability of an event as you gather new evidence.

Given a model with parameters $\Theta \in \mathbb{R}^{D_\Theta}$ generating data $x \in \mathbb{R}^{D_x}$ via a likelihood $p(x \mid \Theta)$, the posterior distribution is:

$$p(\Theta \mid x) = \frac{p(x \mid \Theta) p(\Theta)}{p(x)}$$

Bayes theorem

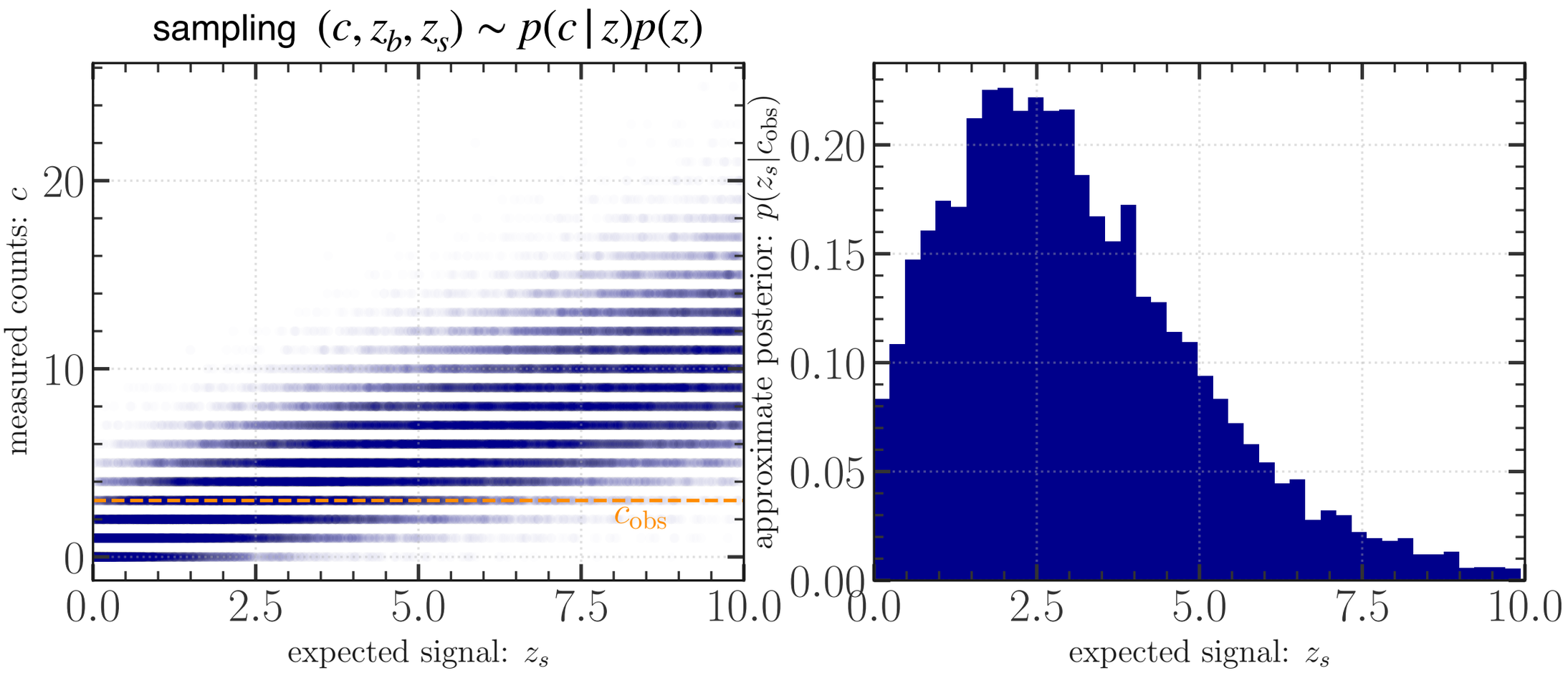
$p(x, \Theta) = p(x \mid \Theta) p(\Theta)$ **is called the joint distribution of x and Θ.**

From Bayes theorem to simulations

Idea: If we can sample $(x, \Theta) \sim p(x \mid \Theta)p(\Theta)$ by running the simulator, we can approximate the posterior by rejection.

Example: *counting experiment with background z_b and signal z_s so that the total observed number of counts c is*

$$z_b \sim \mathcal{U}(0,2), \quad z_s \sim \mathcal{U}(0,10), \quad c \sim \mathrm{Pois}(z_b + z_s)$$



We observe $c = c_{\mathrm{obs}}$ implying the posterior for the signal (marginalised over z_b)**:**

$$p(z_s \mid c_{\mathrm{obs}}) = \left\{ z_s \mid (c, z_b, z_s) \in \mathscr{D} \wedge c = c_{\mathrm{obs}} \right\}$$

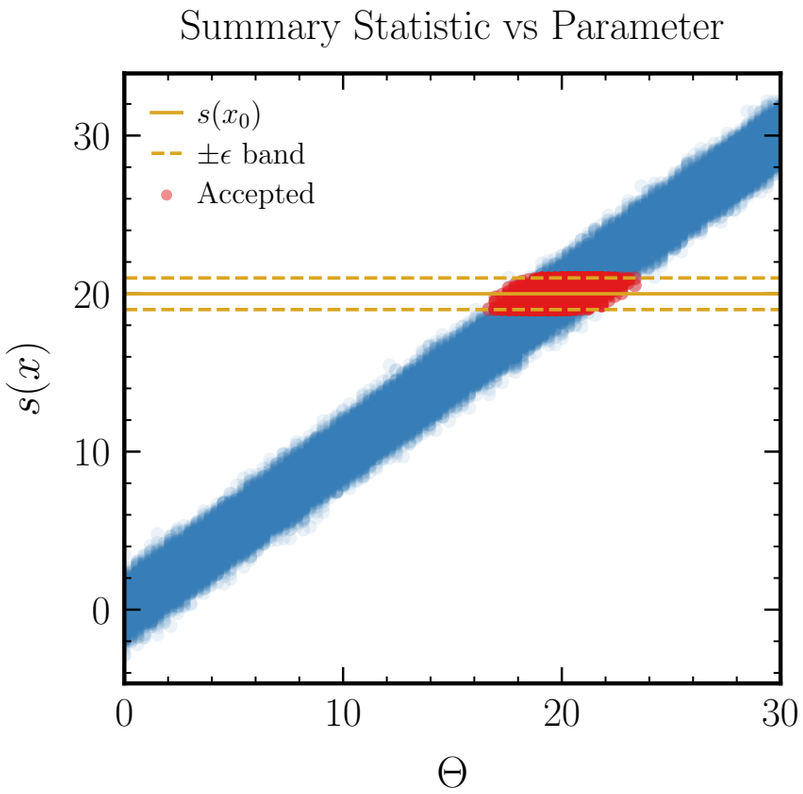
Approximate Bayesian Computation (ABC)

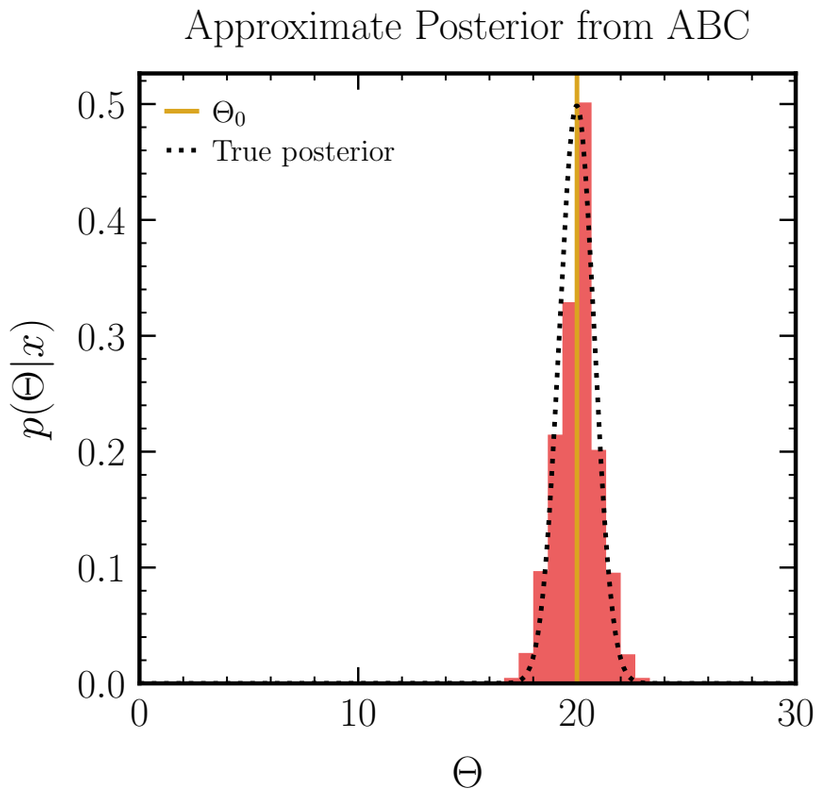
This approach is long-known in the statistics community and goes under the name of Approximate Bayesian Computation (ABC).

You get the ABC posterior by:

$$p_{\mathrm{ABC}}(\boldsymbol{\Theta} \mid \boldsymbol{x}) = \frac{\int_{d(\boldsymbol{x}_0,\boldsymbol{x})<\epsilon} \mathrm{d}\boldsymbol{x}\; p(\boldsymbol{x} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{\int_{d(\boldsymbol{x}_0,\boldsymbol{x})<\epsilon} \mathrm{d}\boldsymbol{x}\; p(\boldsymbol{x})}$$

- Needs to define a metric $d(x, x_0)$ on data space.
- High-dimensional data must be compressed via **summary statistic** $s(x)$**.**





Why don't we hear more often about this?

- **posterior calibration:** ABC posteriors are conservative (too broad) by design.
- **choice of summary statistic:** Handcrafted summaries $s(x)$ may not retain all information from x, optimal summary statistics are hard to find.
- **curse of dimensionality:** The more dimensions your parameter space has, the more simulations you need to perform.
- **choice of** ϵ**:** Smaller tolerance means better calibration but smaller acceptance (more simulations).

What does implicit likelihood mean?

A way to handle the complexity is by embracing it! Can't we just solve the inverse problem by simulating the full physics and learn the likelihood implicitly from simulations?

Yes!

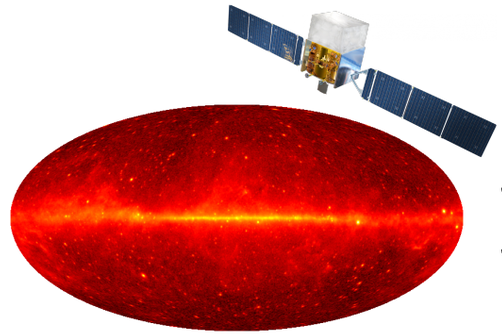
→ Using something the contemporary literature calls **simulation-based inference, ABC supercharged by neural networks.**

What does implicit likelihood mean?

A way to handle the complexity is by embracing it! Can't we just solve the inverse problem by simulating the full physics and learn the likelihood implicitly from simulations?

theory predictions & uncertainties or signal and background
+
instrument effects



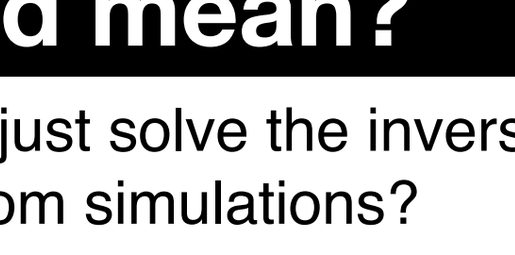


*set of **realistic** model simulations*

***fast** simulations*
& theoretical priors

simulation-based inference (SBI)

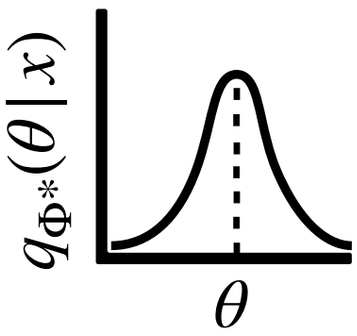
(neural posterior estimation)



normalising flows q_Φ

Bayesian posteriors

$$p(\theta \mid x) = \frac{p(x \mid \theta) p(\theta)}{p(x)}$$



$$D_{\mathrm{KL}}(p \,||\, q_\Phi) = \int p(x) \, \ln \frac{p(x)}{q_\Phi(x)} \, \mathrm{d}x$$

minimise optimisation objective D_{KL}

Key benefits:

- **Efficient:** Handles high-dimensional parameter spaces. Marginalisation over nuisance parameters.
- **Elegant:** You can learn your summary statistic alongside the posterior via an embedding network.
- **Robust:** Calibration of posteriors enabled by amortisation. Employ model mis-specification checks.

Where does SBI beat the likelihood $\mathcal{L}$?

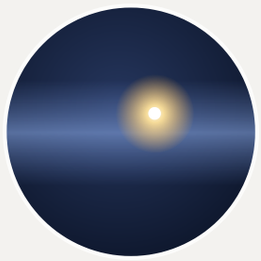
Given data d arising from parameters of interest θ and nuisance parameters η:

Likelihood requires many evaluations: $p(d \mid \theta, \eta)$

SBI needs to sample data once: $d \sim p(d \mid \theta, \eta)\pi(\theta, \eta)$

Example:

Single gravitational-wave event analysis with LIGO/VIRGO.



$\mathcal{L}$ Likelihood territory

$\mathcal{L}(\theta, \eta) = \prod_k \mathrm{Pois}\big(n_k \mid \mu_k(\theta, \eta)\big)$

one target · template background · few nuisances

SBI not needed: profiling is fast & optimal

single-source spectra · template fits

background: fixed template ↑ | stochastic realisation z ↑

1 target — targets / events sharing one model, each with its own systematics η_i — $N \gg 1$ →

Where does SBI beat the likelihood $\mathcal{L}$?

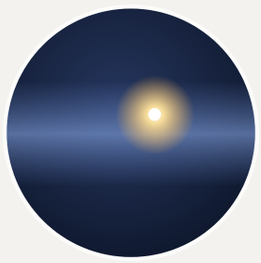
Given data d arising from parameters of interest θ and nuisance parameters η:

Likelihood requires many evaluations: $p(d \mid \theta, \eta)$

SBI needs to sample data once: $d \sim p(d \mid \theta, \eta)\pi(\theta, \eta)$

Later examples:
1. **Source detection with SBI (published work with: N. Anau Montel, F. Calore, F. List, C. Weniger)**
2. Millisecond pulsars in globular clusters [talk by F. Xotta (after coffee break)]
3. Searching for axions in multiple blazars [talk by P. Bhattacharjee (Friday)]

background: fixed template → stochastic realisation z



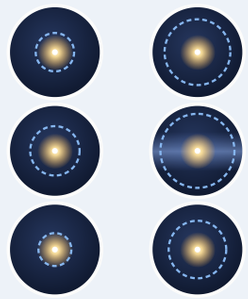
$\mathcal{L}$ Likelihood territory

$\mathcal{L}(\theta, \eta) = \prod_k \mathrm{Pois}\big(n_k \mid \mu_k(\theta, \eta)\big)$

one target · template background · few nuisances

SBI not needed: profiling is fast & optimal

single-source spectra · template fits



1 Many targets, one model

$\prod_i \int d\eta_i \; p(d_i \mid \theta, \eta_i)\, p(\eta_i)$

shared θ, own systematics η_i → N marginalisations

SBI amortised: train once, apply to all N

dSph stacking (J-factors) · event-wise IRFs (IACTs) · source populations

1 target — targets / events sharing one model, each with its own systematics η_i — $N \gg 1$

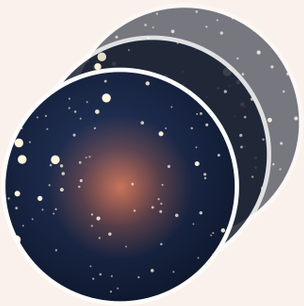
Where does SBI beat the likelihood $\mathcal{L}$?

Given data d arising from parameters of interest θ and nuisance parameters η:

Likelihood requires many evaluations: $p(d \mid \theta, \eta)$

SBI needs to sample data once: $d \sim p(d \mid \theta, \eta)\pi(\theta, \eta)$

stochastic realisation z ↑



2 Stochastic backgrounds

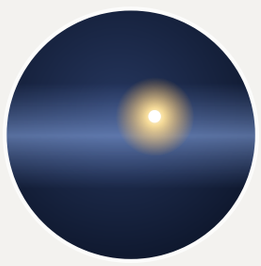
$$p(d \mid \theta) = \int dz\, p(d \mid \theta, z)\, p(z)$$

z = one random sky realisation, $\dim z \gtrsim 10^3$

SBI just sample z, no $p(d \mid \theta)$ needed

GCE: DM vs. unresolved PS · high-latitude dN/dS · diffuse-model ensembles

Later examples:
Analysing future LISA data of gravitational waves
[Talk by G. Puleo (after coffee break)]



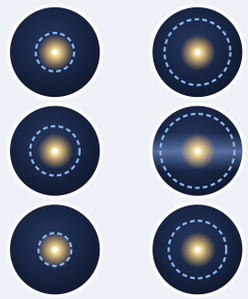
$\mathcal{L}$ Likelihood territory

$$\mathcal{L}(\theta, \eta) = \prod_k \mathrm{Pois}\big(n_k \mid \mu_k(\theta, \eta)\big)$$

one target · template background · few nuisances

SBI not needed: profiling is fast & optimal

single-source spectra · template fits



1 Many targets, one model

$$\prod_i \int d\eta_i\; p(d_i \mid \theta, \eta_i)\, p(\eta_i)$$

shared θ, own systematics η_i → N marginalisations

SBI amortised: train once, apply to all N

dSph stacking (J-factors) · event-wise IRFs (IACTs) · source populations

background: fixed template ↑

1 target — **targets / events sharing one model, each with its own systematics η_i** — $N \gg 1$ →

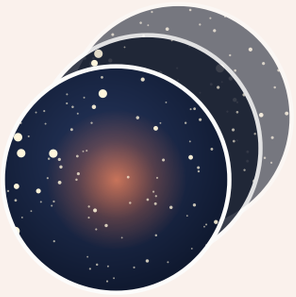
Where does SBI beat the likelihood $\mathcal{L}$?

Given data d arising from parameters of interest θ and nuisance parameters η:

Likelihood requires many evaluations: $p(d \mid \theta, \eta)$

SBI needs to sample data once: $d \sim p(d \mid \theta, \eta)\pi(\theta, \eta)$

stochastic realisation z ↑



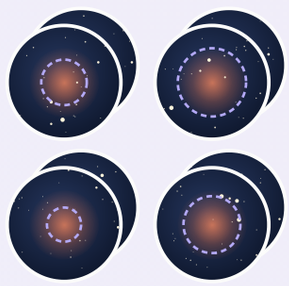
2 Stochastic backgrounds

$p(d \mid \theta) = \int dz\, p(d \mid \theta, z)\, p(z)$

z = one random sky realisation, $\dim z \gtrsim 10^3$

SBI just sample z, no $p(d \mid \theta)$ needed

GCE: DM vs. unresolved PS · high-latitude dN/dS · diffuse-model ensembles



3 Both at once

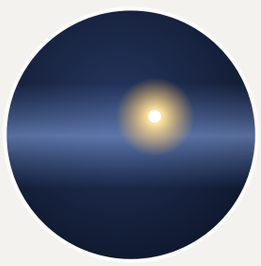
★ SBI sweet spot

$\prod_i \int d\eta_i\, dz_i\, p(d_i \mid \theta, \eta_i, z_i)\, p(\eta_i, z_i)$

N × high-dim. integrals: out of reach for explicit $\mathcal{L}$

SBI amortised marginal posteriors for θ

dSphs: J-factors × sub-threshold blazars · sky patches × own PS populations



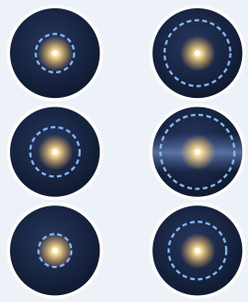
$\mathcal{L}$ Likelihood territory

$\mathcal{L}(\theta, \eta) = \prod_k \mathrm{Pois}\big(n_k \mid \mu_k(\theta, \eta)\big)$

one target · template background · few nuisances

SBI not needed: profiling is fast & optimal

single-source spectra · template fits



1 Many targets, one model

$\prod_i \int d\eta_i\, p(d_i \mid \theta, \eta_i)\, p(\eta_i)$

shared θ, own systematics η_i → N marginalisations

SBI amortised: train once, apply to all N

dSph stacking (J-factors) · event-wise IRFs (IACTs) · source populations

background: fixed template ↑

1 target — **targets / events sharing one model, each with its own systematics η_i** — $N \gg 1$ →

Lat example:
Field-level inference on cosmological parameters via multiple tracers.
[Talk by G. Scelfo (after coffee break)]

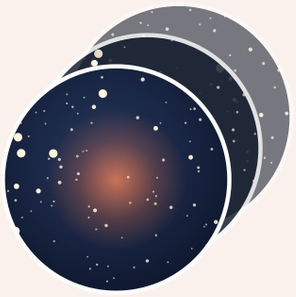
Where does SBI beat the likelihood $\mathcal{L}$?

Given data d arising from parameters of interest θ and nuisance parameters η:

Likelihood requires many evaluations: $p(d \mid \theta, \eta)$

SBI needs to sample data once: $d \sim p(d \mid \theta, \eta)\pi(\theta, \eta)$

stochastic realisation z ↑



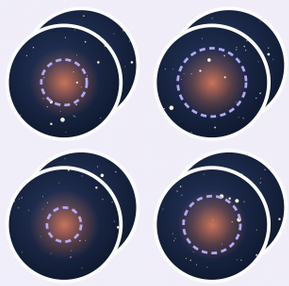
2 Stochastic backgrounds

$p(d \mid \theta) = \int dz\, p(d \mid \theta, z)\, p(z)$

z = one random sky realisation, $\dim z \gtrsim 10^3$

SBI just sample z, no $p(d \mid \theta)$ needed

GCE: DM vs. unresolved PS · high-latitude dN/dS · diffuse-model ensembles



3 Both at once

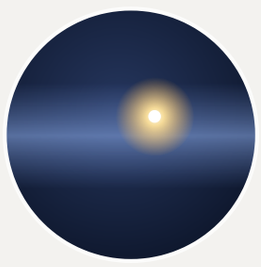
★ SBI sweet spot

$\prod_i \int d\eta_i\, dz_i\, p(d_i \mid \theta, \eta_i, z_i)\, p(\eta_i, z_i)$

N × high-dim. integrals: out of reach for explicit $\mathcal{L}$

SBI amortised marginal posteriors for θ

dSphs: J-factors × sub-threshold blazars · sky patches × own PS populations



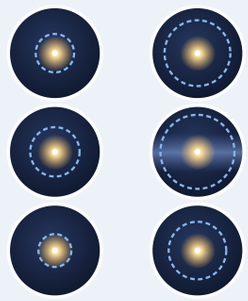
$\mathcal{L}$ Likelihood territory

$\mathcal{L}(\theta, \eta) = \prod_k \mathrm{Pois}\big(n_k \mid \mu_k(\theta, \eta)\big)$

one target · template background · few nuisances

SBI not needed: profiling is fast & optimal

single-source spectra · template fits



1 Many targets, one model

$\prod_i \int d\eta_i\, p(d_i \mid \theta, \eta_i)\, p(\eta_i)$

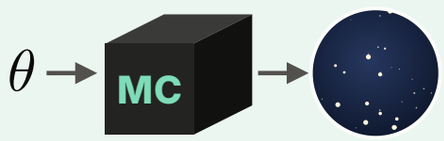
shared θ, own systematics η_i → N marginalisations

SBI amortised: train once, apply to all N

dSph stacking (J-factors) · event-wise IRFs (IACTs) · source populations

background: fixed template ↓

1 target — targets / events sharing one model, each with its own systematics η_i — $N \gg 1$ →



4 Black-box forward model

applies anywhere on the map ↑

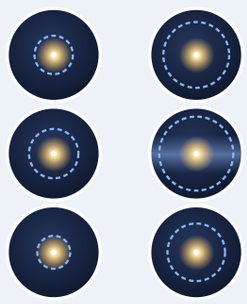
signal or instrument exists only as a simulator: no closed-form $\mathcal{L}$, and Wilks' theorem often fails

ALP–γ mixing in turbulent B-fields · IGMF pair cascades · IACT air-shower + detector MC

example today

How do you work with SBI in practice?

— Some examples from gamma-ray astrophysics —



1 Many targets, one model

$$\prod_i \int d\eta_i \; p(d_i \mid \theta, \eta_i)\, p(\eta_i)$$

shared θ, own systematics η_i → N marginalisations

SBI amortised: train once, apply to all N

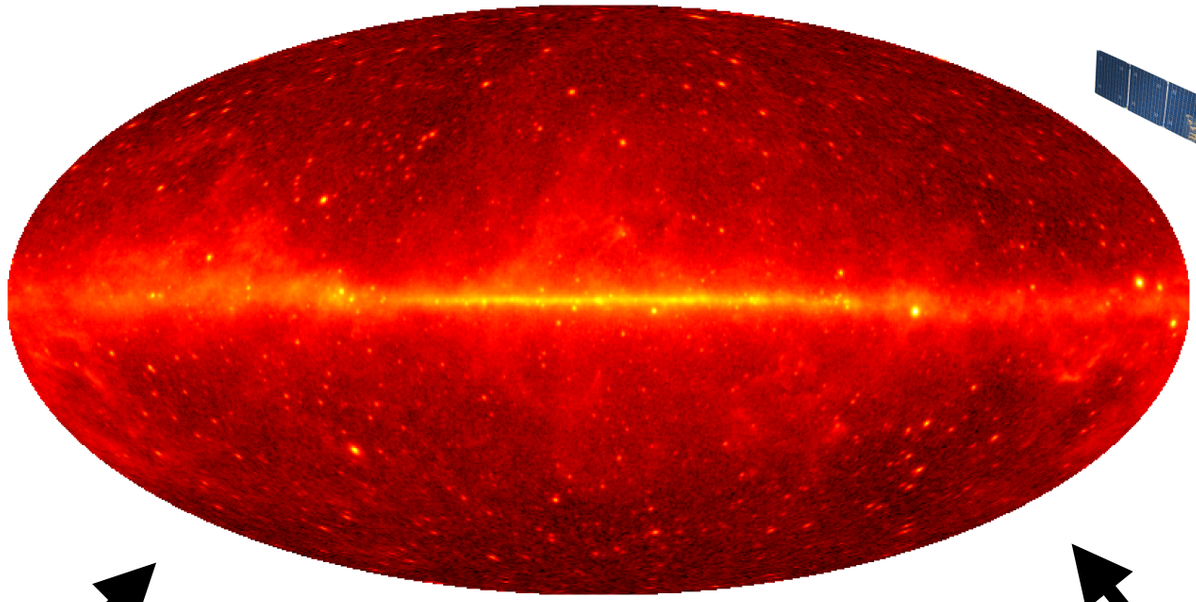
dSph stacking (J-factors) | event-wise IRFs (IACTs) | source populations

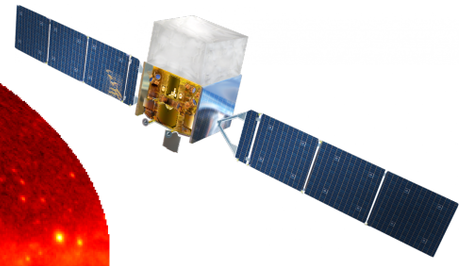
Detecting GeV-bright gamma-ray sources in the high-latitude sky seen by the Fermi LAT

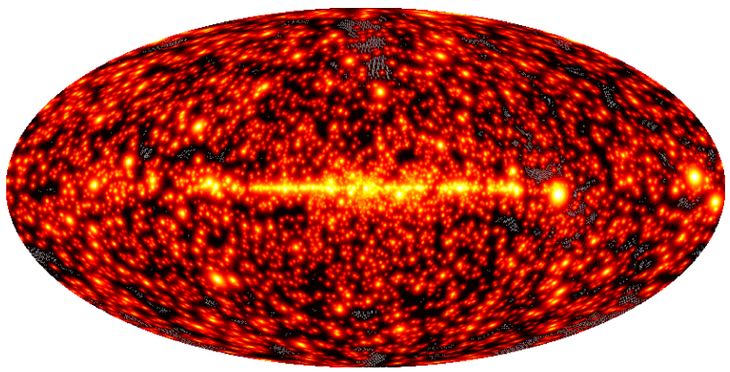
published study with: Noemi Anau Montel, Francesca Calore, Florian List, Christoph Weniger

[Eckner et al.; *PRD* 112 (2025) 10, 103022]

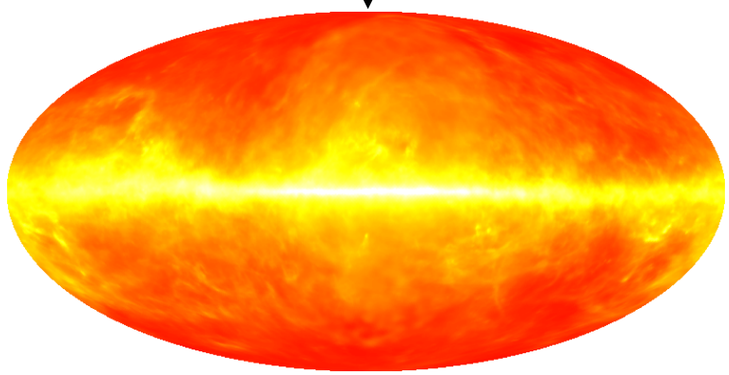
Understanding the gamma-ray sky



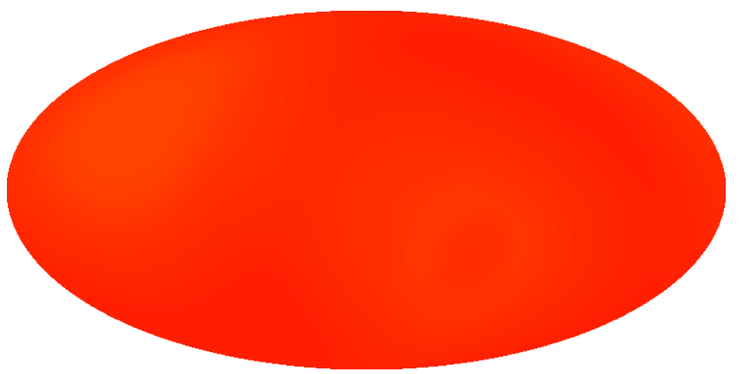




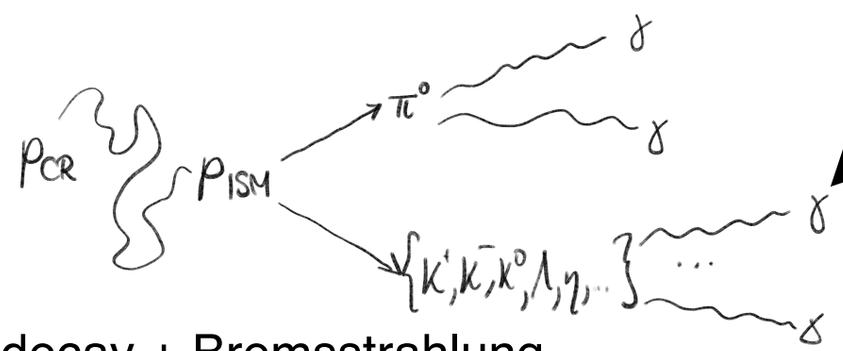
discrete sources



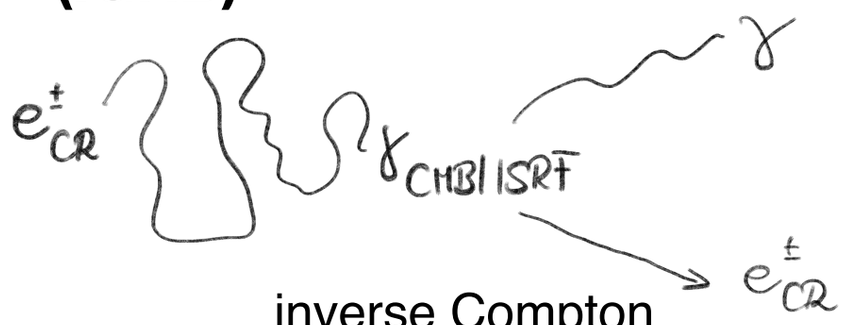
Galactic diffuse emission



isotropic γ-ray background (IGRB)

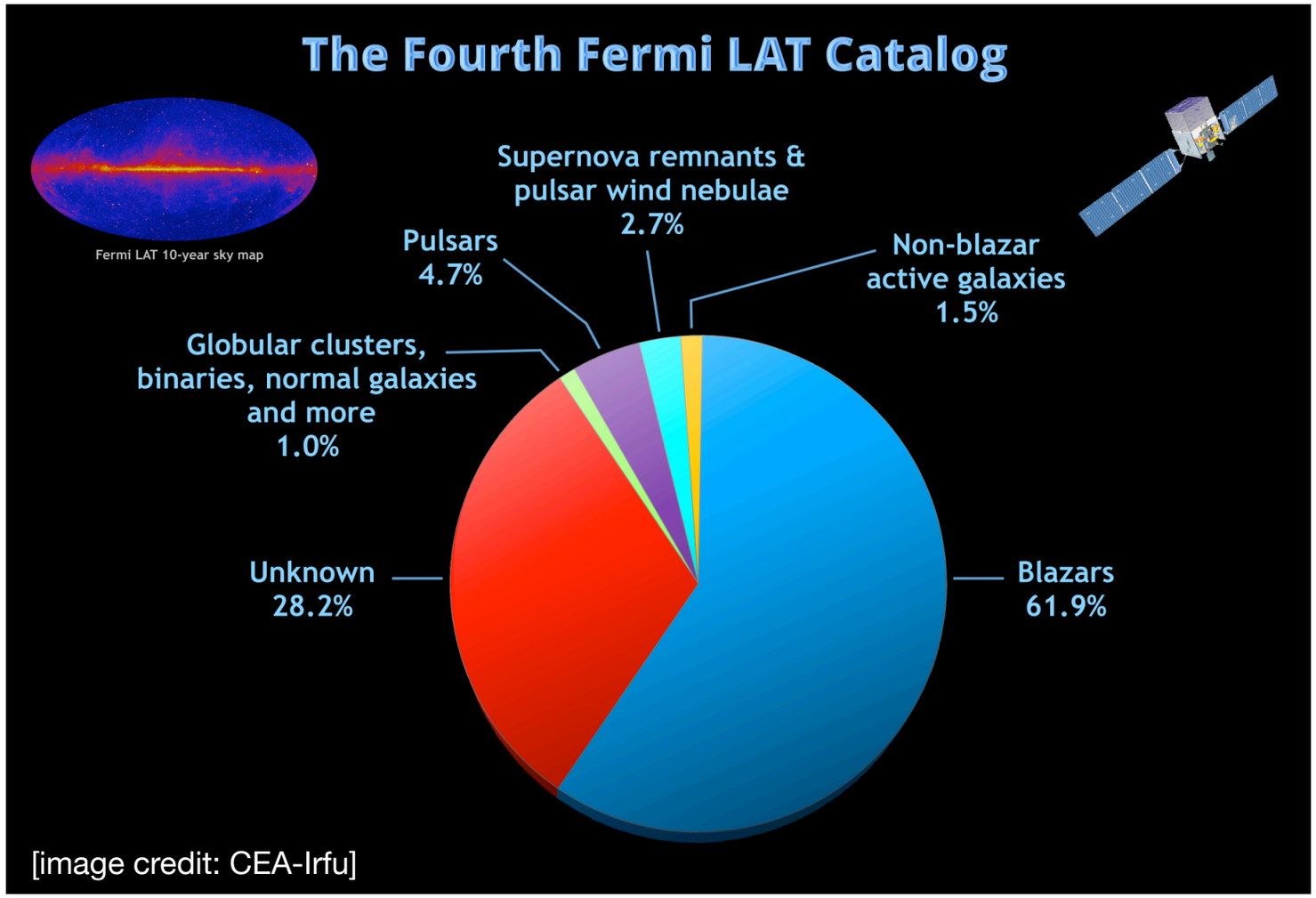


π^0-decay + Bremsstrahlung



inverse Compton

Census of known GeV-bright gamma-ray sources



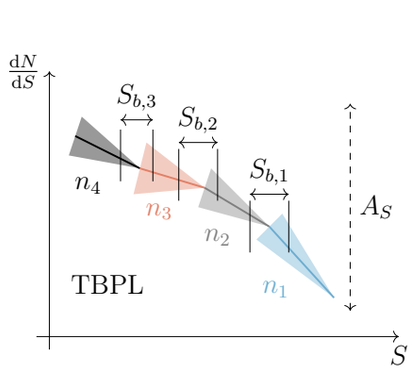
→ **The brightest of these discrete sources can be detected by likelihood approaches.**
→ **About 60 % of all detected discrete GeV-bright sources are blazars.**
→ **There remains a huge portion of undetected discrete source from the same population (sub-threshold sources).**

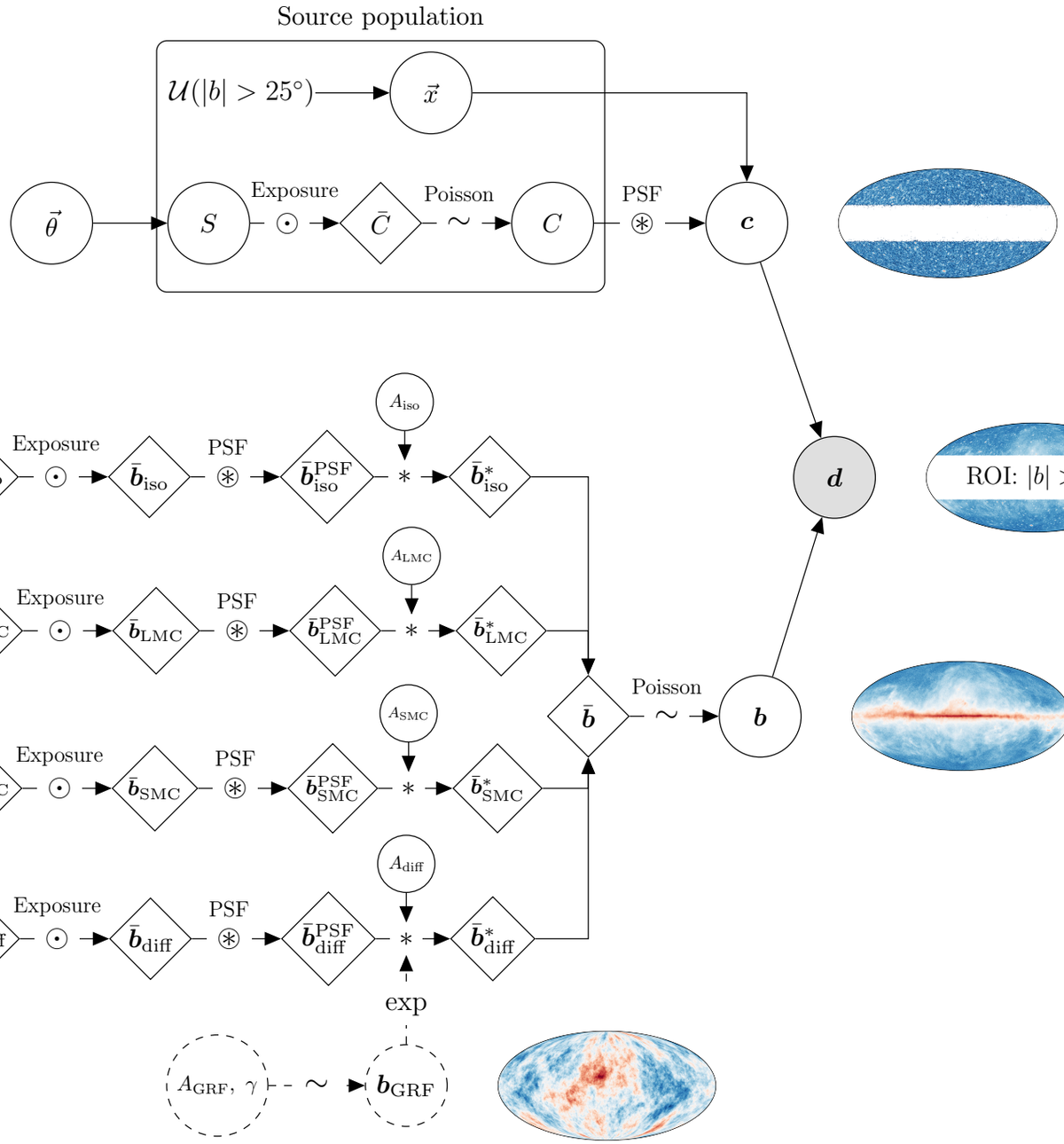
→ **Sub-threshold sources are a background for all science analyses! Is there a way to consistently merge detected and sub-threshold sources with SBI?**

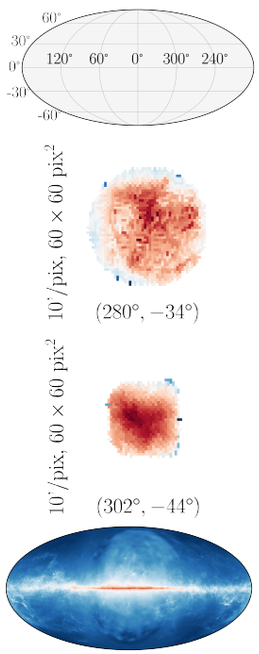
Detection in the high-latitude GeV gamma-ray sky

Objective: Localise the bright part of the gamma-ray sources at high latitudes (detection).
under realistic instrument response and background uncertainty.

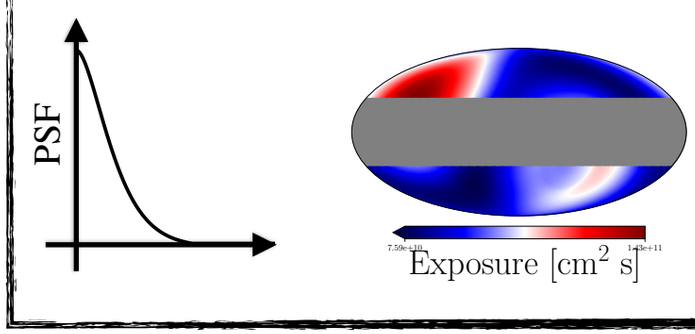
LAT data selection: 14 years, `SOURCEVETO FRONT` events from **1 to 10 GeV (single energy bin)**.







And the instrument response:



Neural Ratio Estimation as a Flavour of SBI

SBI might also use a different flavour to get a posterior distribution: neural ratio estimation.

→ Learn the posterior-to-prior ratio r with a binary classifier!

[B. Miller et al., J. Open Source Softw. 7 (2022) 75]

swyft — **T**runcated **M**arginal **N**eural **R**atio **E**stimation

$$r(x;\Theta) = \frac{p(\Theta\,|\,x)}{p(\Theta)} = \frac{p(x,\Theta)}{p(x)p(\Theta)}$$

joint sample

Class 1: Matching (data, parameter) pairs

(🐱, cat)
(🐕, dog)
(🏠, house)
(⭐, star)
(🦍, monkey)

$x, \Theta \sim P(x, \Theta)$

marginal sample

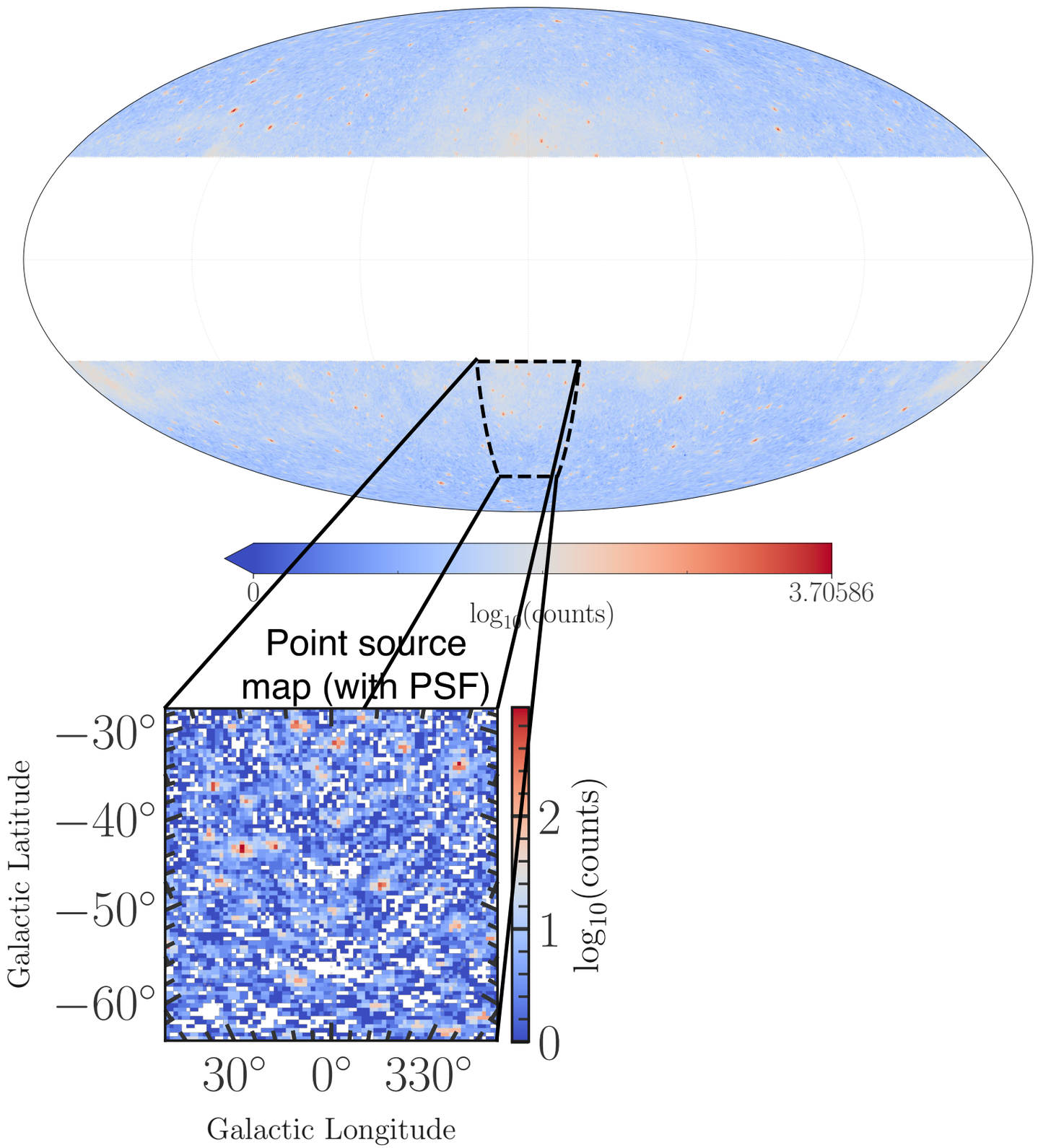
Class 0: Scrambled (data, parameter) pairs

(🐱, dog)
(🐕, cat)
(⭐, star)
(🏠, monkey)
(🦍, house)

$x, \Theta \sim P(x)P(\Theta)$

In a nutshell: We train a neural network as a binary classifier to tell us if in a pair (x, Θ) Θ generated x which **only requires a forward model of the physics involved**.

Source Detection using SBI — Method



[N. Anau Montel & C. Weniger, arXiv:2211.04291]

Source detection in SBI language:
Given the actual observed sky, what is the probability of observing a source at a certain position with emitted photons C exceeding a certain threshold C_{th}?

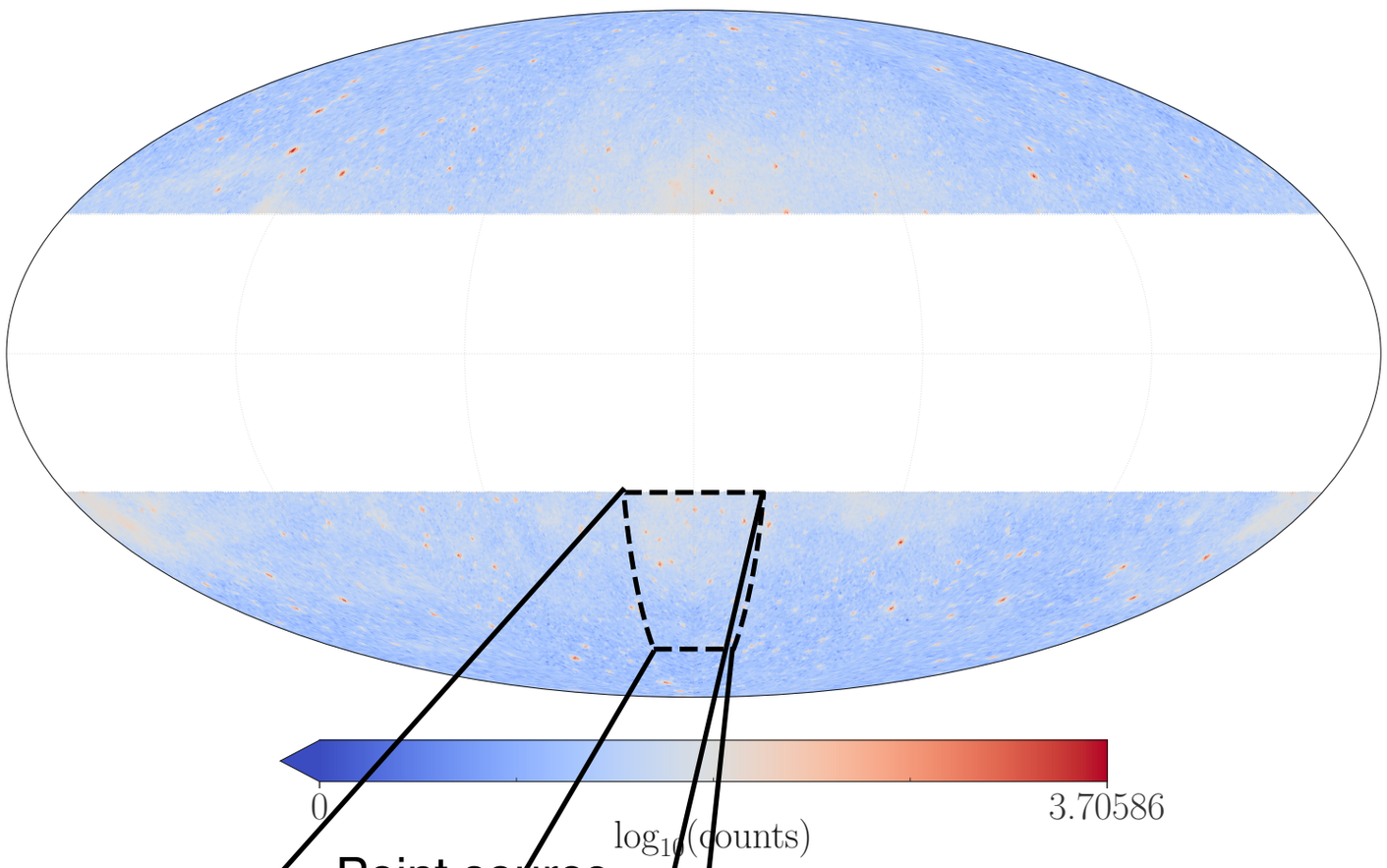
$$r(\Omega, C_{\mathrm{th}}; x) = \frac{p(\mathbb{I}_x(C \geq C_{\mathrm{th}}) = 1, \Omega \,|\, x)}{p(\mathbb{I}_x(C \geq C_{\mathrm{th}}) = 1, \Omega)}$$

Source Detection using SBI — Method

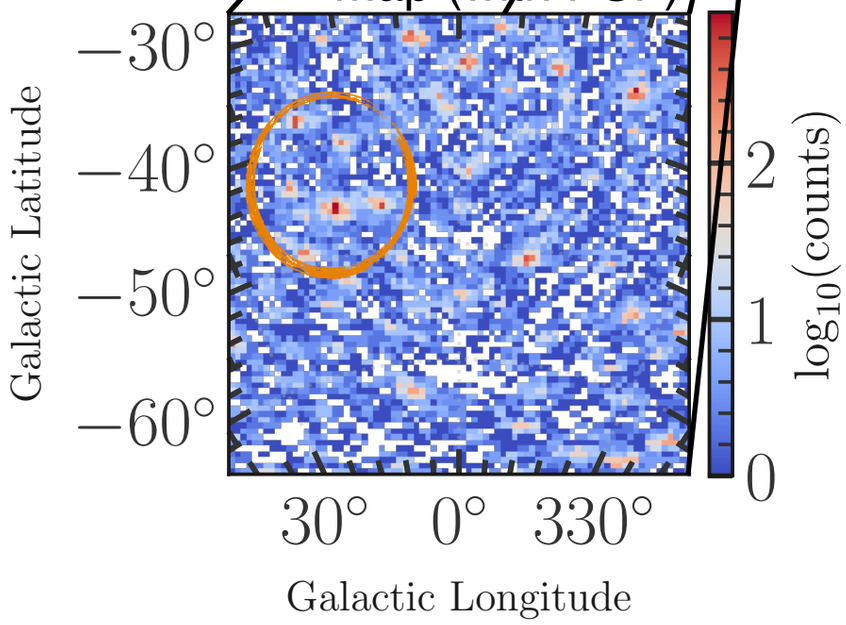
[N. Anau Montel & C. Weniger, arXiv:2211.04291]

Source detection in SBI language:
Given the actual observed sky, what is the probability of observing a source at a certain position with emitted photons C exceeding a certain threshold C_{th}?

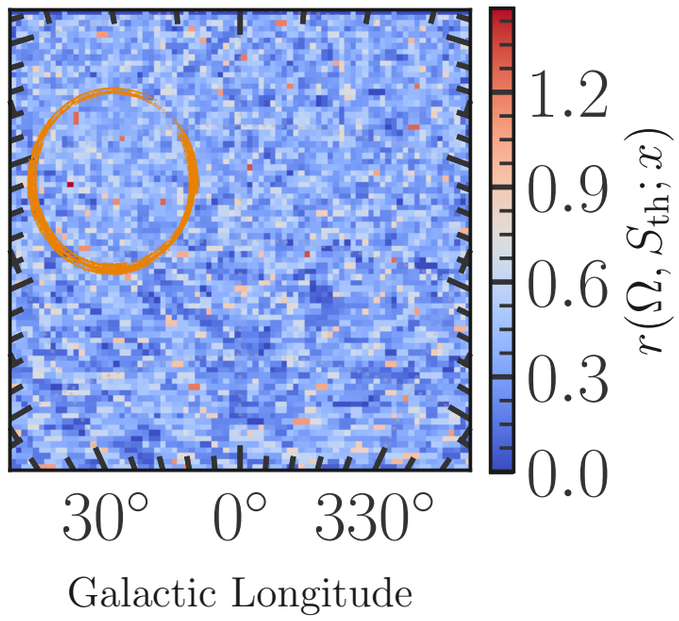
$$r(\Omega, C_{\mathrm{th}}; x) = \frac{p(\mathbb{I}_x(C \geq C_{\mathrm{th}}) = 1, \Omega \,|\, x)}{p(\mathbb{I}_x(C \geq C_{\mathrm{th}}) = 1, \Omega)}$$



Point source map (with PSF)



Under the hood:
→ classifier trained on full-sky data maps
→ labels are full-sky maps firing at pixels with point sources
→ UNET architecture as summary statistic on spherical image data (“DeepSphere”)
[N. Perraudin et al., Astron.Comput. 27 (2019)]

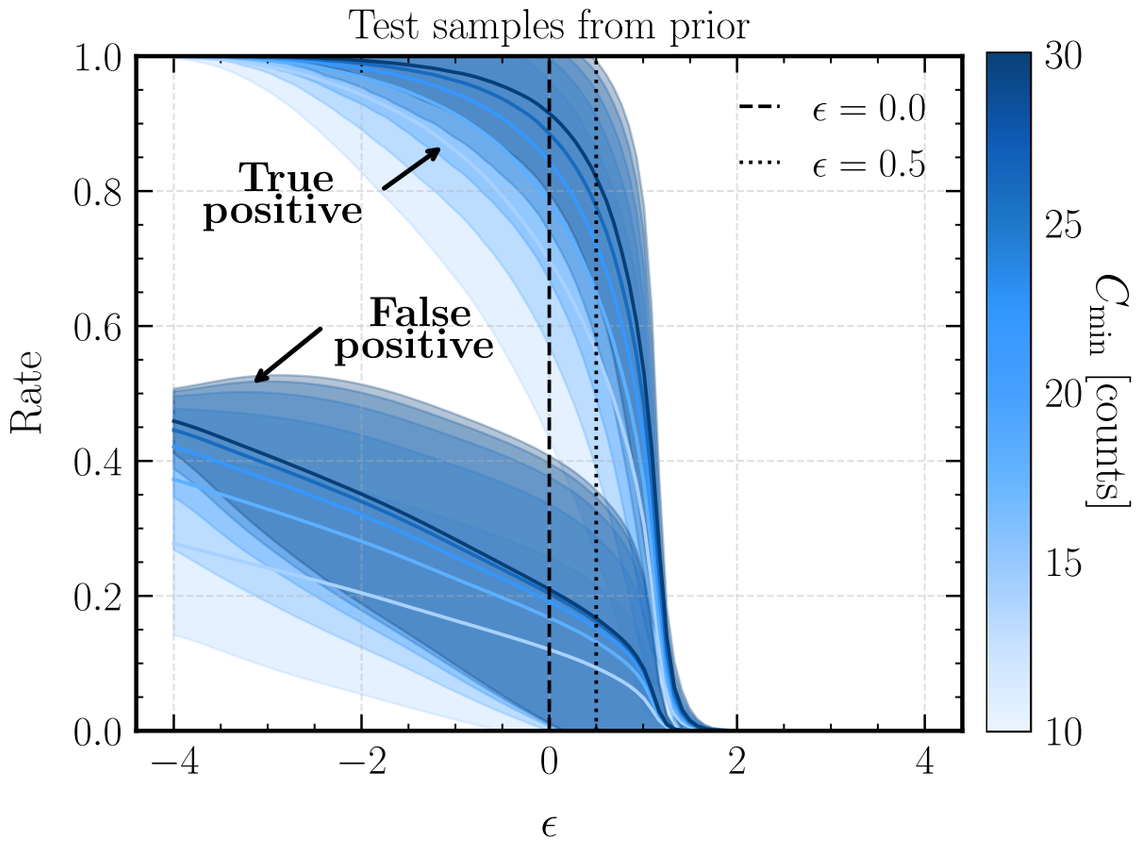


Source Detection using SBI — Results

When do we declare a source detection?

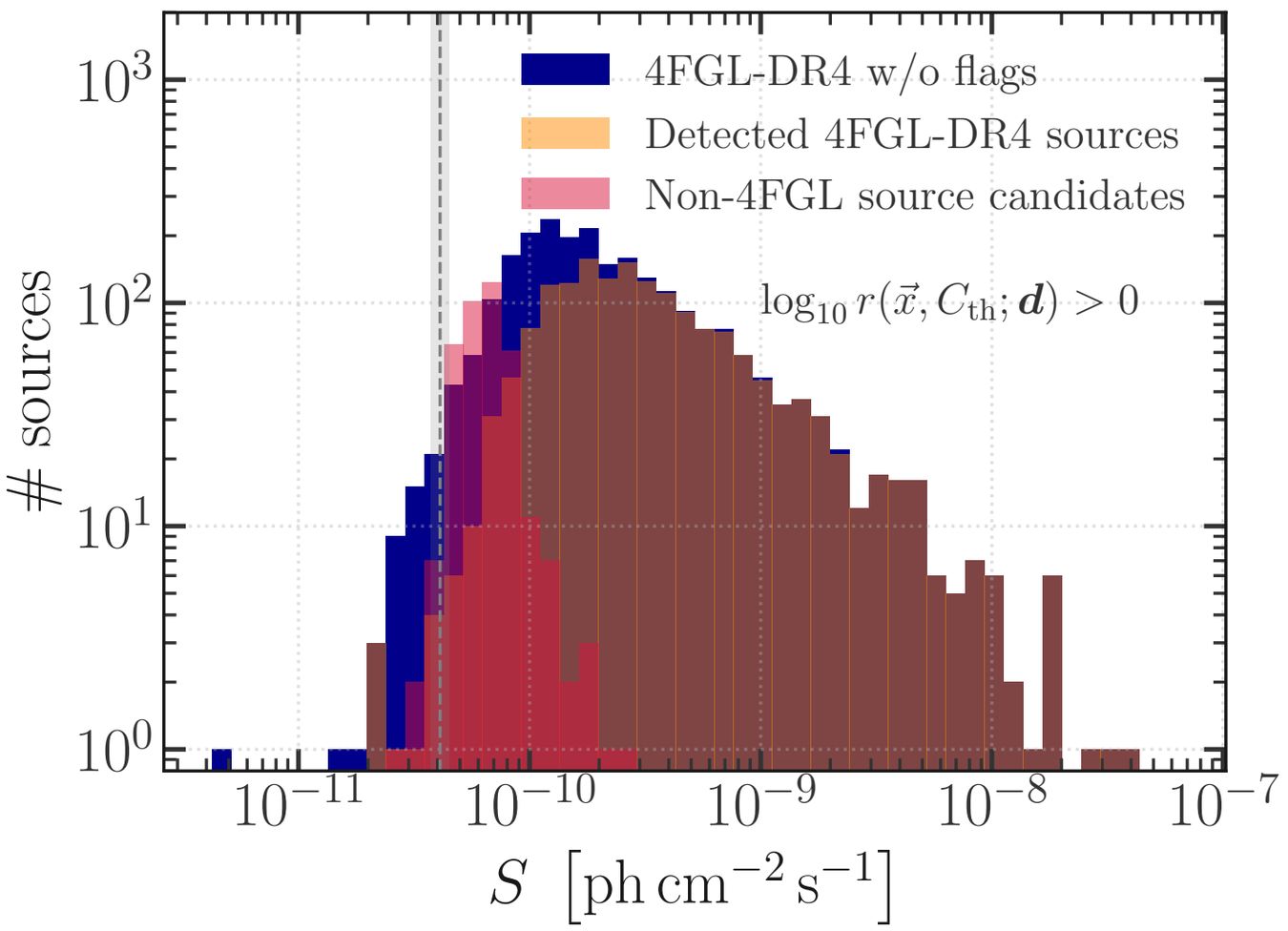
→ Explore on simulated data what threshold ϵ on r is a good balance between true positives and false negatives!

$\log_{10} r(\Omega, C_{\mathrm{th}}; x) > \epsilon$?



[Eckner, Anau Montel, List et al.; *PRD* 112 (2025) 10, 103022]

using 4FGL-DR4 (w/o flags) as ground truth:



98% of all 4FGL-DR4 above $S \sim 3 \times 10^{-10}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ are recovered. We find 387 source candidates not listed in 4FGL-DR4 (may be false positives or genuine sources!)

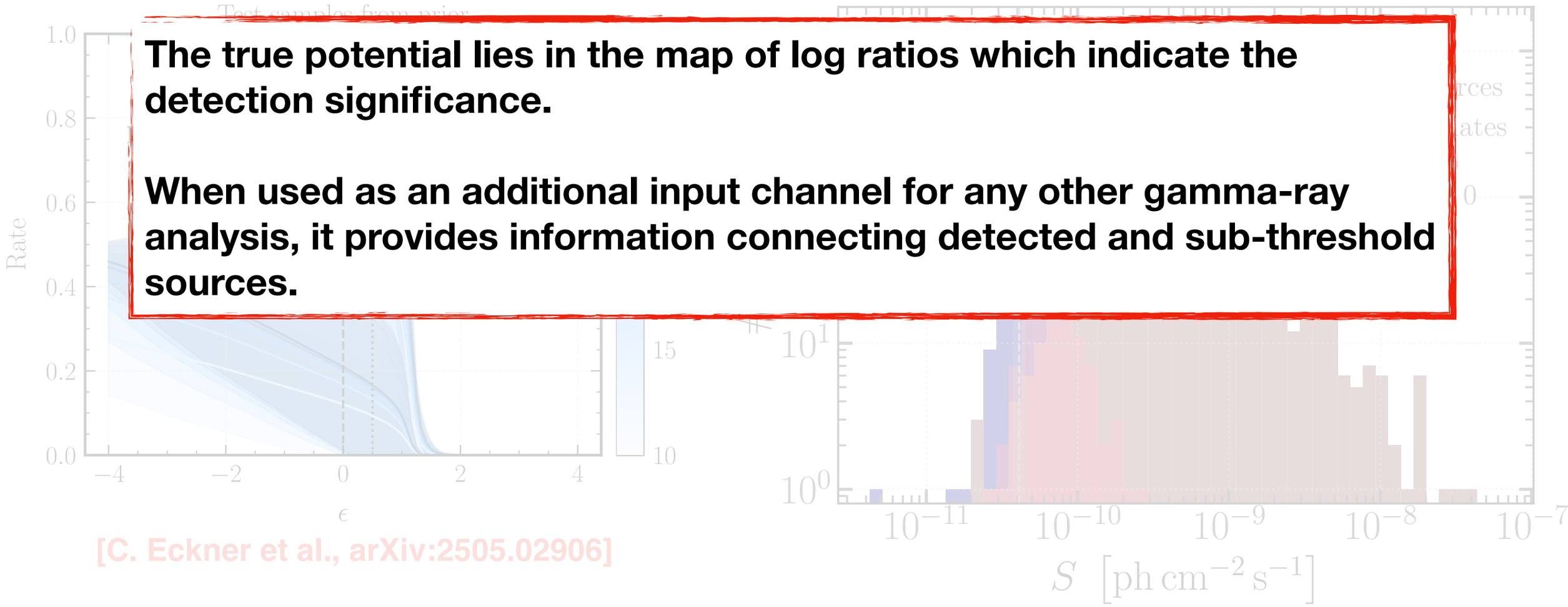
Source Detection using SBI — Results

When do we declare a source detection?

→ Explore on simulated data what threshold ϵ on r is a good balance between true positives and false negatives!

$\log_{10} r(\Omega, C_{\mathrm{th}}; x) > \epsilon$?

using 4FGL-DR4 (w/o flags) as ground truth:

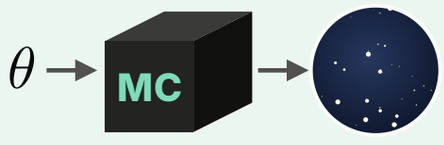


The true potential lies in the map of log ratios which indicate the detection significance.

When used as an additional input channel for any other gamma-ray analysis, it provides information connecting detected and sub-threshold sources.

[C. Eckner et al., arXiv:2505.02906]

98% of all 4FGL-DR4 above $S \sim 3 \times 10^{-10}\,\mathrm{cm^{-2}\,s^{-1}}$ are recovered. We find 387 source candidates not listed in 4FGL-DR4 (may be false positives or genuine sources!)



4 Black-box forward model

applies anywhere on the map ↑

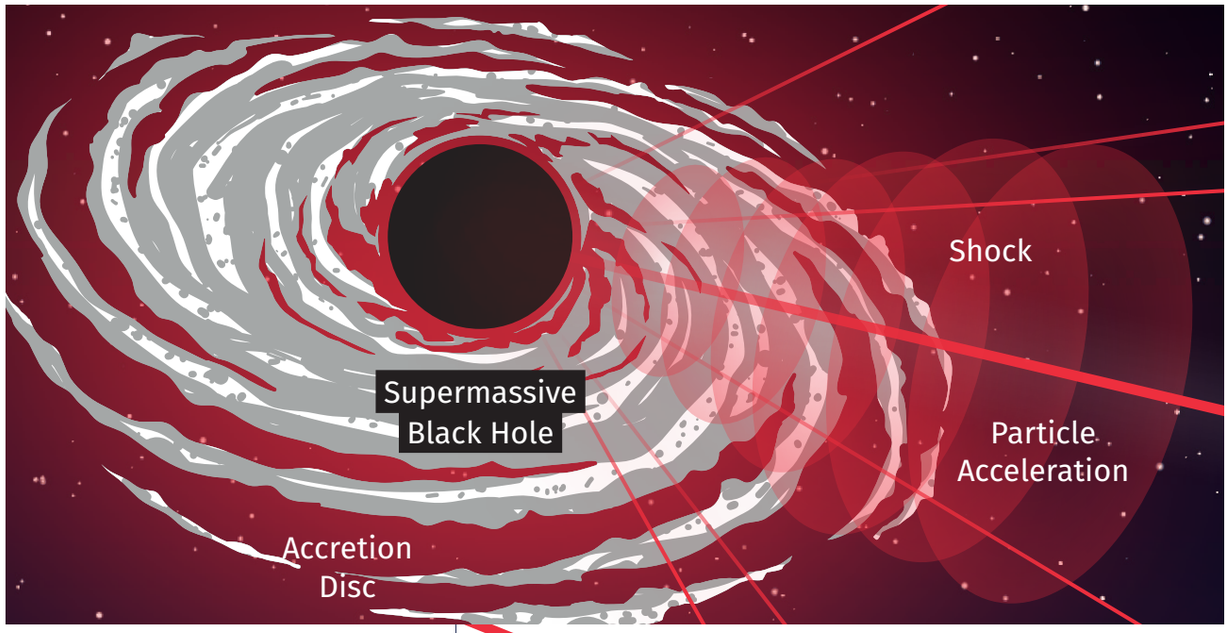
signal or instrument exists only as a simulator: no closed-form $\mathcal{L}$, and Wilks' theorem often fails

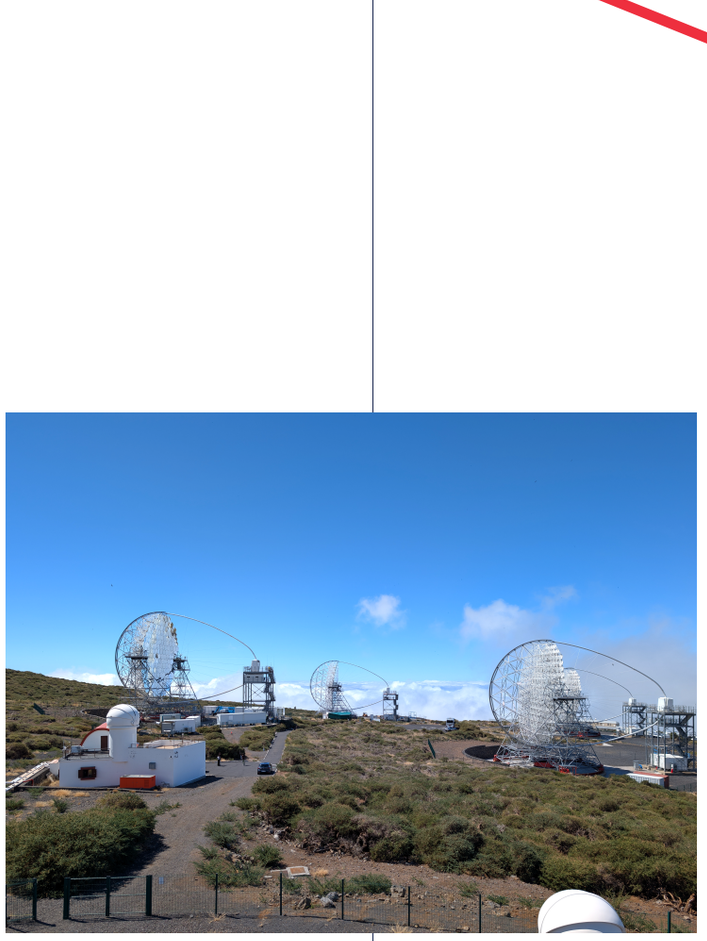
ALP–γ mixing in turbulent B-fields | IGMF pair cascades | IACT air-shower + detector MC

Analysing gamma-ray data of ground-based Cherenkov telescopes with SBI

ongoing project with: Miguel Molero, Jorge Camalich, Sami Caroff, Francesca Calore

Gamma-ray Detection on Ground

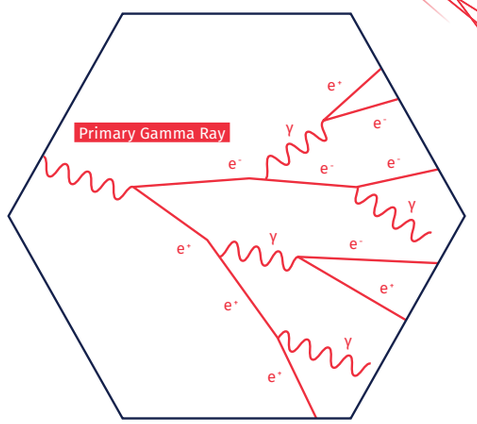




Gamma Ray

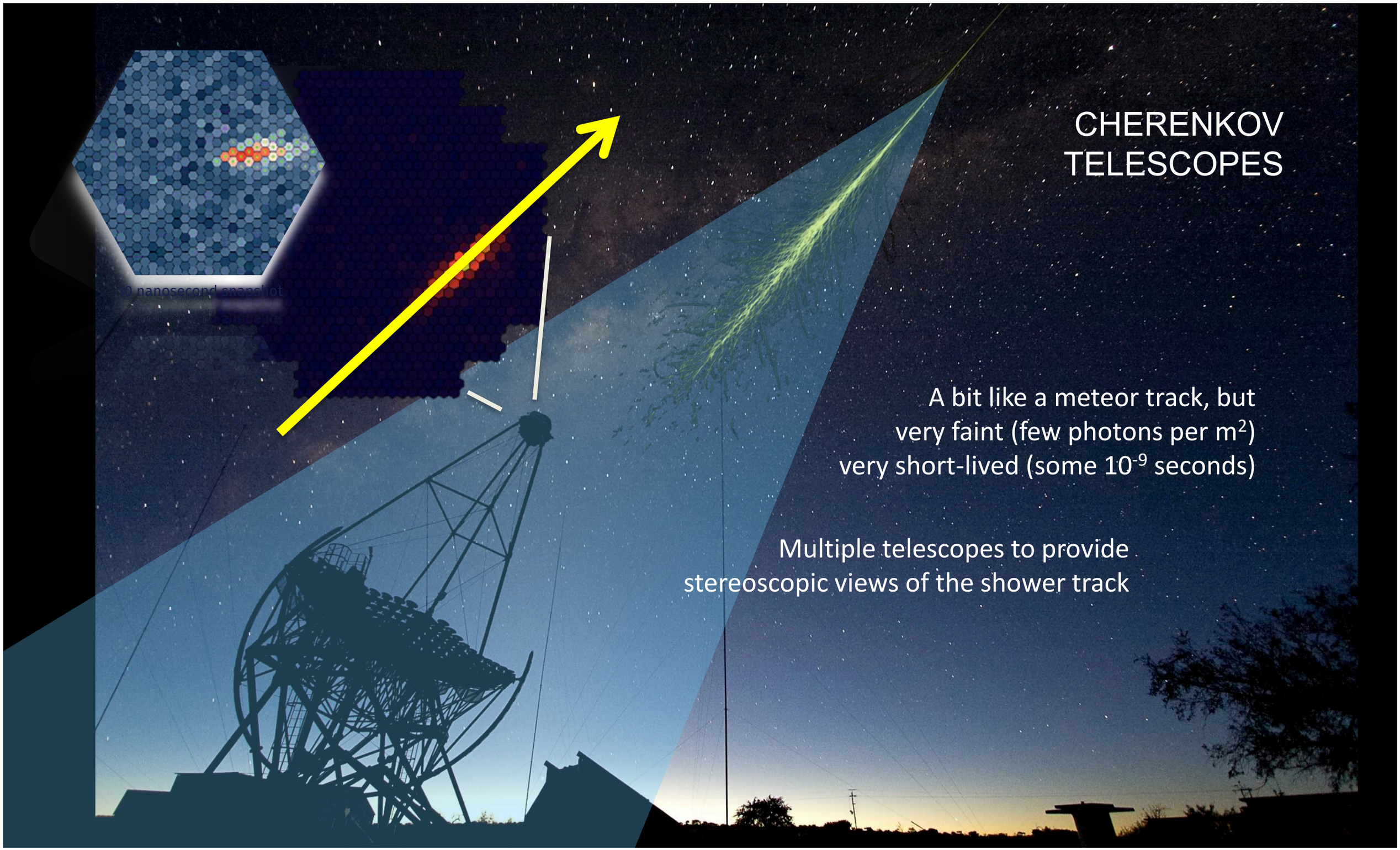
Atmosphere

Air Shower



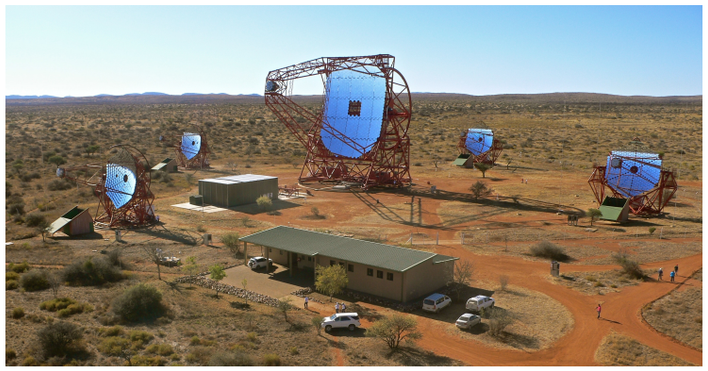
gamma ray interacts with the atmosphere

Gamma-ray Detection on Ground

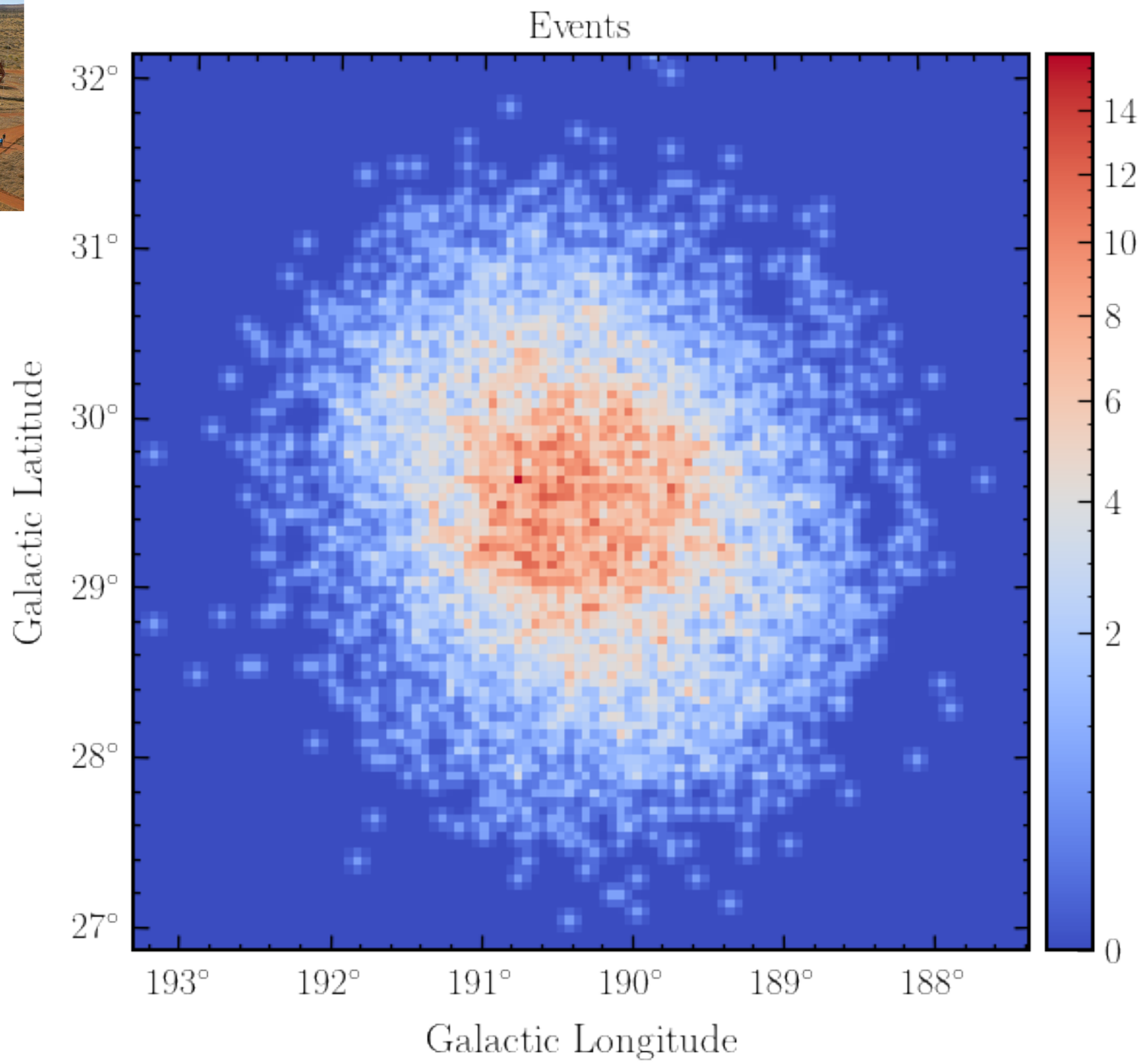


Credit: Werner Hofmann

What are you actually seeing?



H.E.S.S. telescope array in Namibia



IACTs are dominated by background: high-energy proton showers misclassified as gamma-ray events.→ Modelling the background is essential for any analysis.

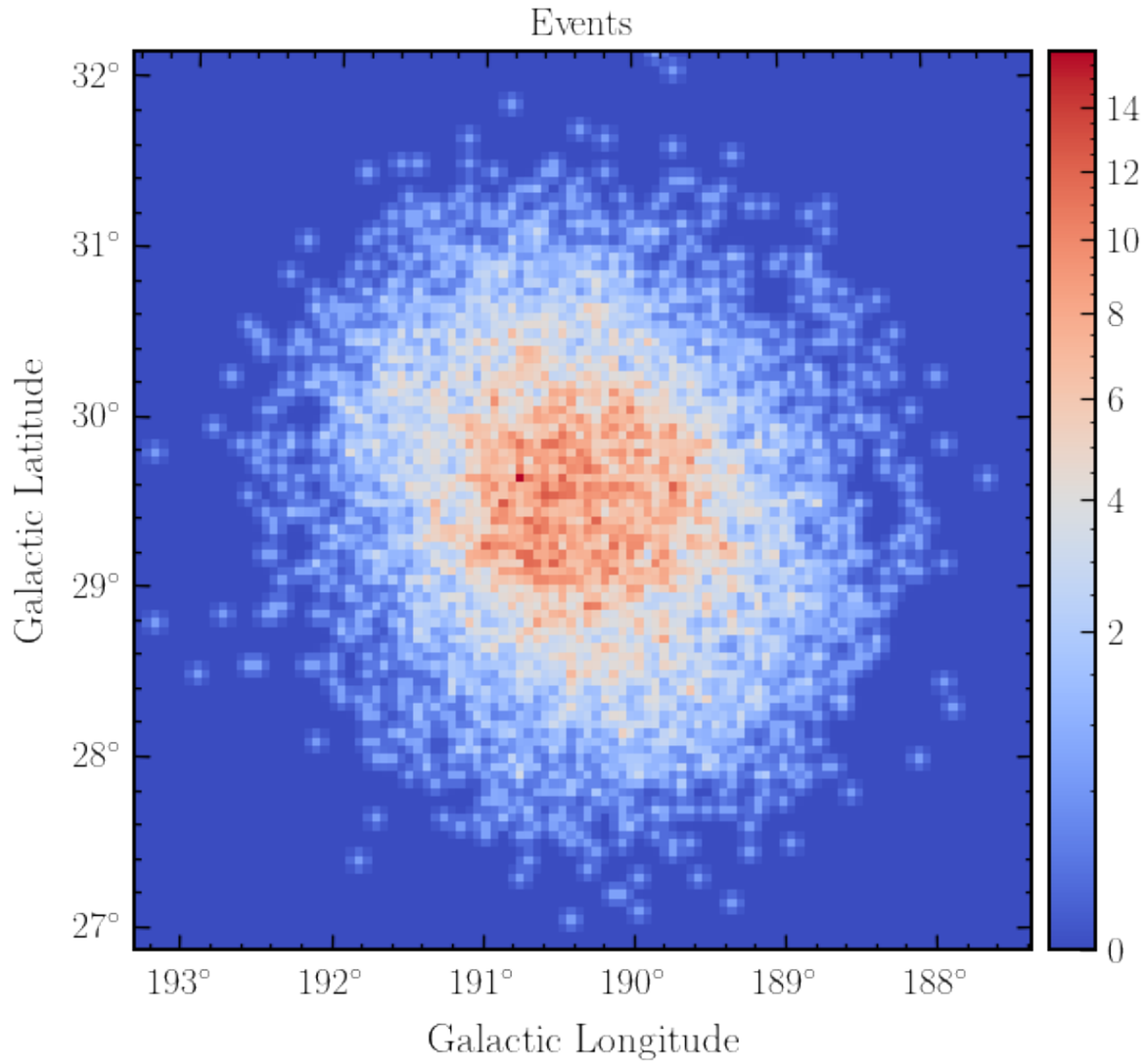
Definition of the model

In the field-of-view frame, the cosmic-ray background is nothing but an inhomogeneous Poisson point process with position and energy-dependent intensity λ.

→ However, everything needs to be conditioned on the observation properties.

$$\lambda(u \mid c) = \underbrace{\rho(c)}_{\text{rate}} \times \underbrace{p(u \mid c)}_{\text{acceptance}}$$

- $u = (x, y, \log_{10} E_{\text{reco}}) \rightarrow$ 3D info
- c : observation conditions
- ρ : one total rate per run
- p : normalised acceptance shape



Observation conditions for H.E.S.S. dataset:
$c = (\cos\theta_{\text{zen}},\ \sin\varphi_{\text{az}},\ \cos\varphi_{\text{az}},\ \varepsilon_{\text{opt}})$, **muon efficiency** ε_{opt}: related to camera throughput; **zenith angle** θ_{zen}: impacts threshold and rate via traversed atmospheric column; **azimuth** φ_{az}: its angle to the shower broadens images due to the geomagnetic field

Definition of the model

The acceptance shape p will be modelled via a normalising flow and conditioned on the observation conditions. The rate ρ is its own feed-forward network.

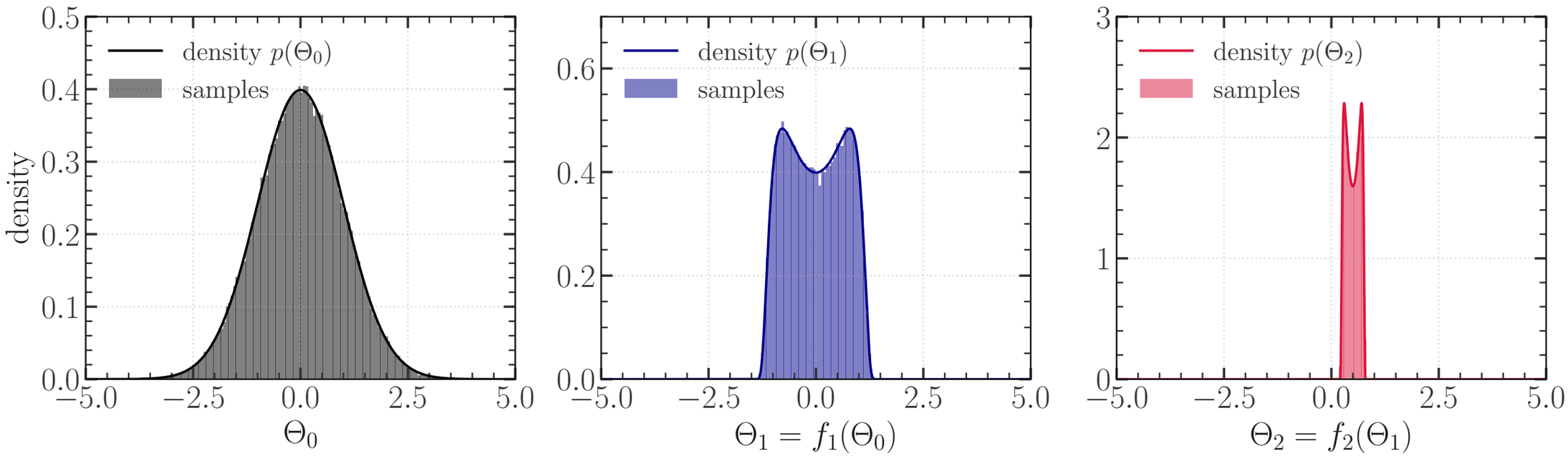
→ Density distribution answering the question what is the probability to observe an event with this energy in this field-of-view position under these conditions?

What is a normalising flow?

A neural network architecture that transforms a simple base distribution into a complex one via a chain of invertible, differentiable maps. Given $\Theta_0 \sim p(\Theta_0)$ (e.g. standard normal), define:

$$x = \Theta_K = f_K \circ \cdots \circ f_1(\Theta_0)$$

$$\log p(x) = \log p(\Theta_0) - \sum_{k=1}^{K} \log \left| \det J_{f_k}(\Theta_{k-1}) \right|$$



Training and calibration of the model

What do we want the generative model to do?

→ The Poisson mean μ at any infinitesimal volume δ in the FoV coincides with the observed Poisson scattered events in this volume element.

$$\mu = \int_{\delta} \lambda(u \,|\, c)\, \mathrm{d}u \quad \xrightarrow[\text{with } N_k \text{ events:}]{\text{per run } k} \quad \underbrace{\ln L = \ln \mu_k - \sum_{i}^{N_k} \ln\left[\rho(c_k) p(u_i \,|\, c_k)\right]}_{\text{minimise}}$$

Implementation of systematic uncertainty via calibration:

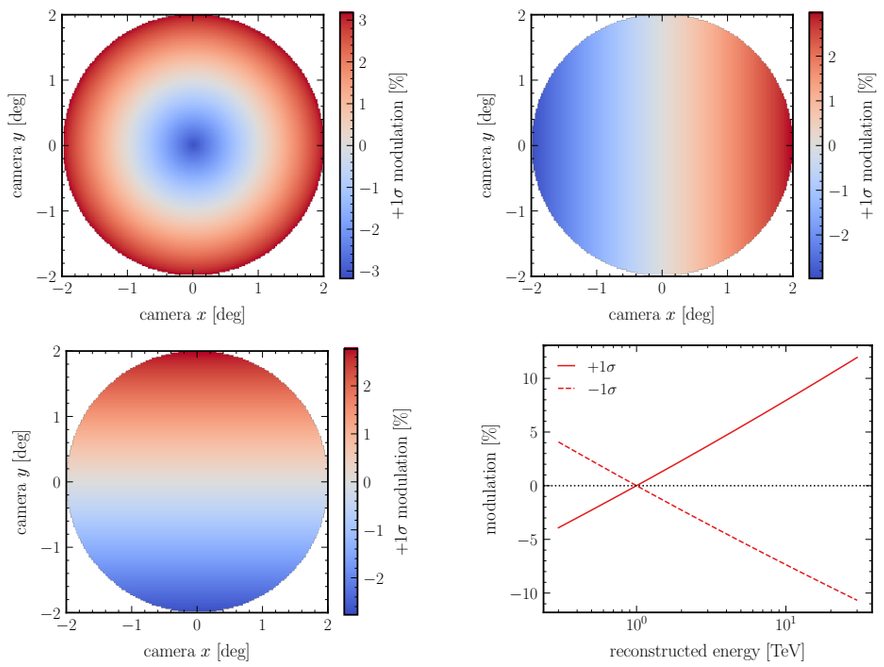
→ Allow for additional scatter of the of the trained intensity λ to accommodate systematics.

→ This it what makes it a probabilistic generative model!

$$\lambda_z = \lambda \, \exp\left(B(u) \cdot z_k\right), \quad z_k \sim \mathcal{N}(0, \Sigma_k),$$

$$B = \left[\, 1,\ \frac{r}{R_0} - 1,\ \frac{x}{R_0},\ \frac{y}{R_0},\ \frac{E - \bar{E}}{\sigma_E} \,\right]$$

Estimate the residual scatter from runs that were not part of the training dataset!



Generative model from the H.E.S.S. dataset

Training pool:
45 dark patches + 10 runs for PKS 2105-344 in flaring mode
→ PKS masked at 0.5° in camera FoV; position followed over time

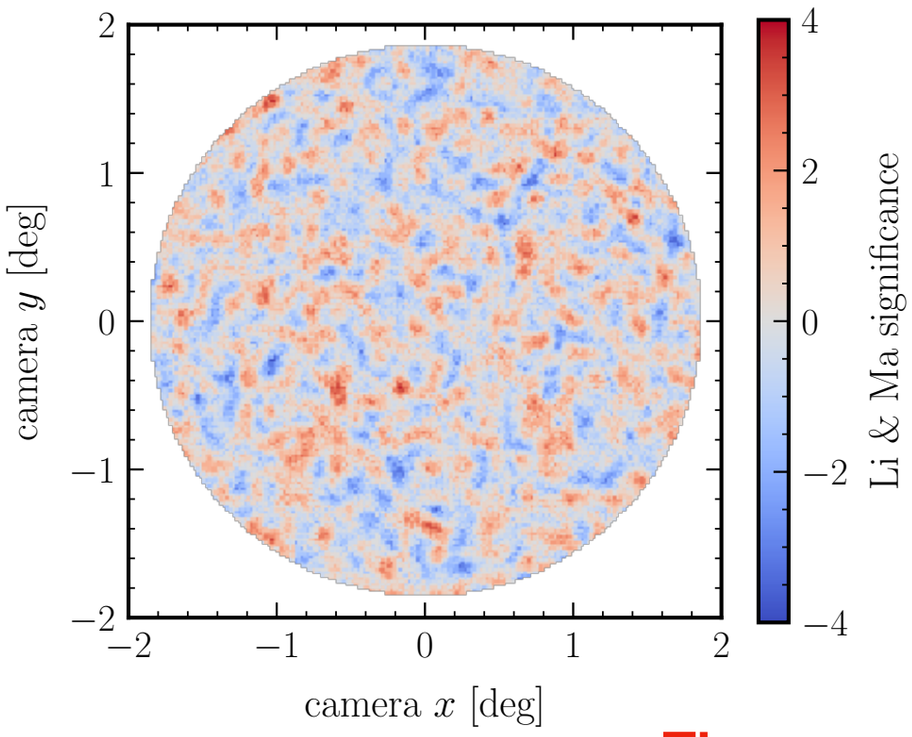
Calibration pool:
5 runs for PKS 2105-344 in flaring mode
→ PKS masked at 0.5° in camera FoV; position followed over time

Calibration:

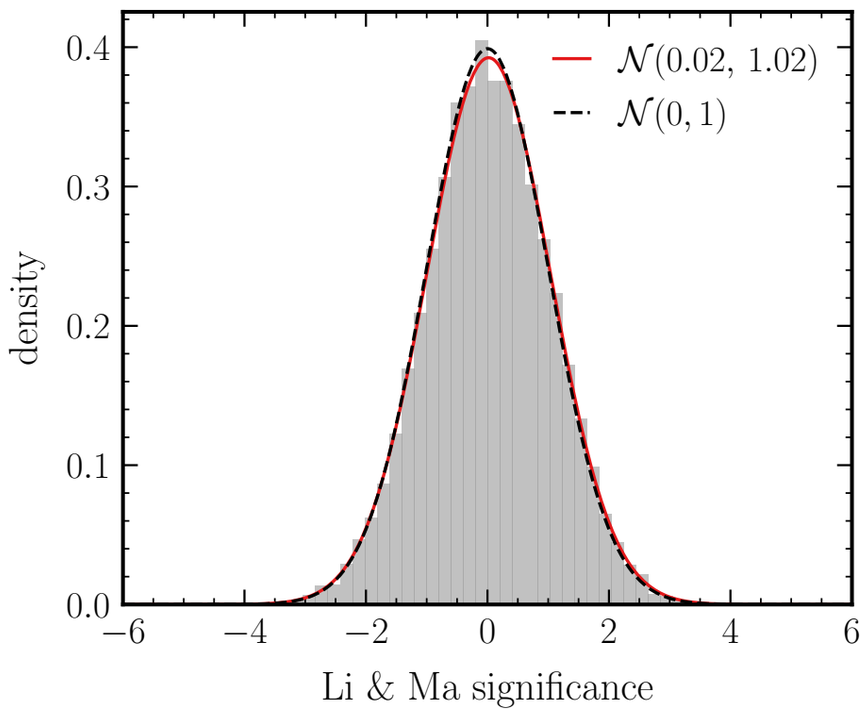
mode	1σ
normalisation	**15.18 %**
	95 %: 11.96–17.95 %
radial	3.13 %
x / *y* gradient	1.48 / 1.38 %
spectral tilt	3.82 %

⇒

Given its conditions, a run's rate is still uncertain by
$\sim 15\,\%$
and redrawn per run in the simulator.

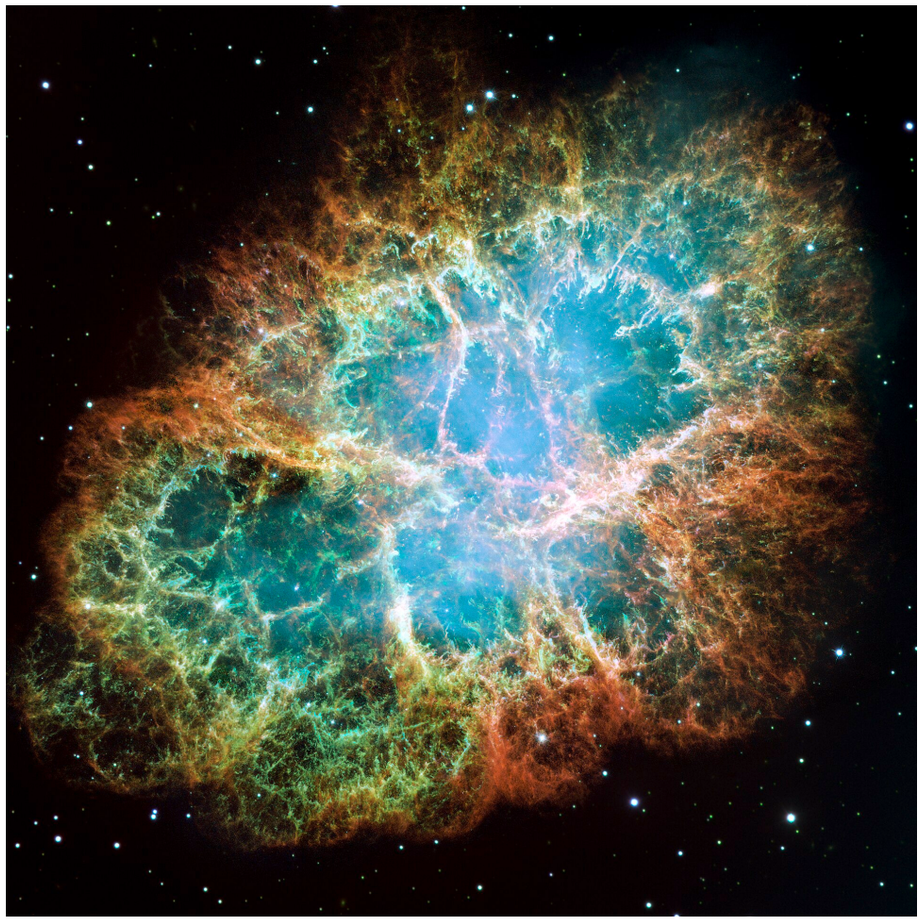


background residuals on stacked unseen test runs



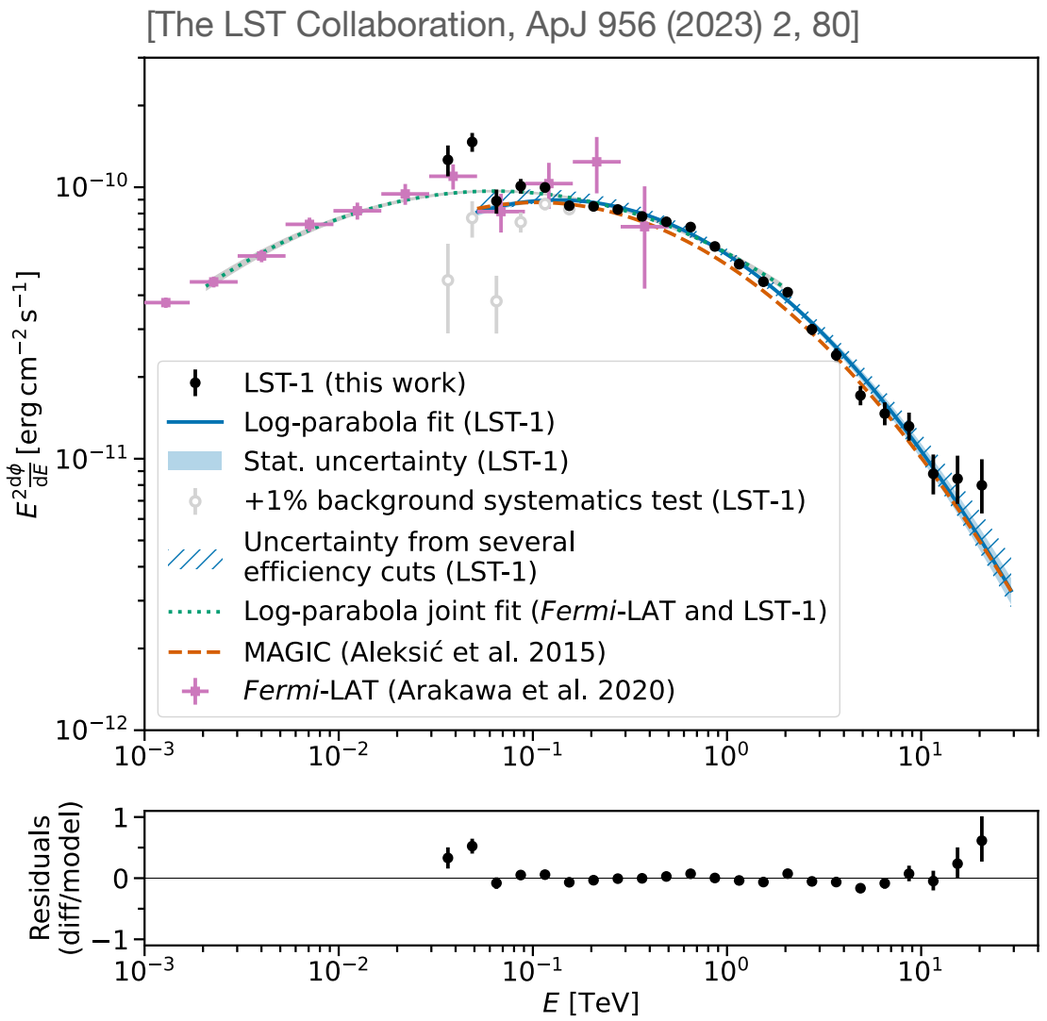
The generative model performs as it should.

Application to an actual astrophysical source: Crab Nebula



[credit: NASA]

[The LST Collaboration, ApJ 956 (2023) 2, 80]



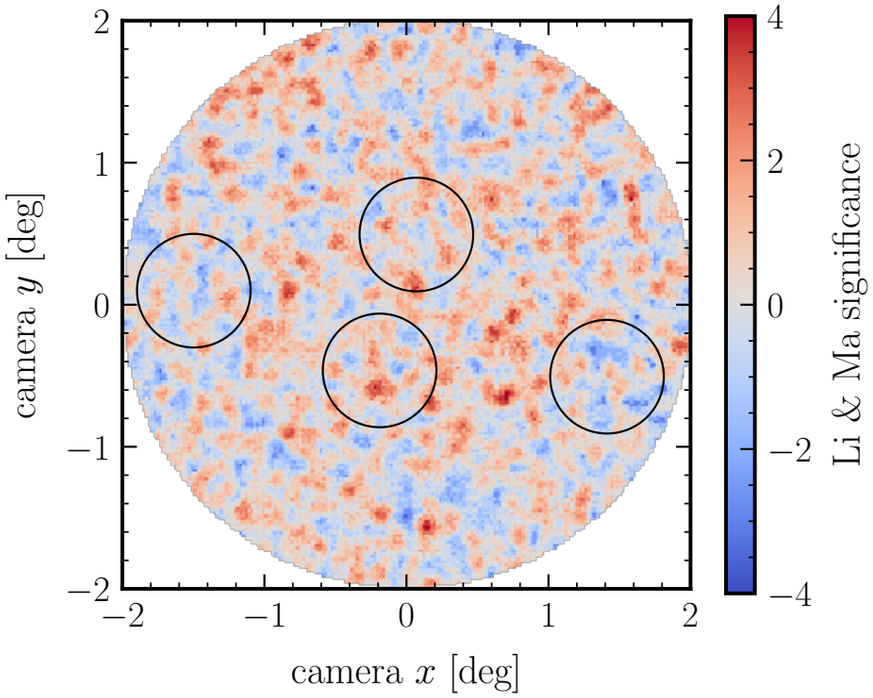
The case of the Crab Nebula in the H.E.S.S. data

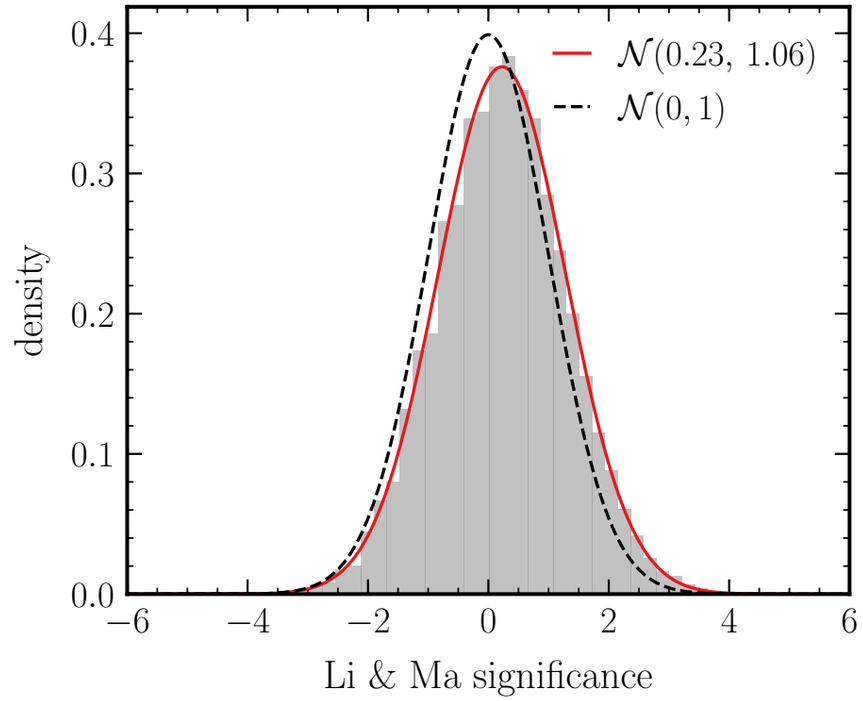
run	zen. [°]	$\varepsilon_{\mathrm{opt}}$	E_{safe}	offset
23523	48.6	0.907	0.87	0.5°
23526	45.8	0.914	0.71	0.5°
23559	45.4	0.920	0.66	1.5°
23592	48.4	0.912	0.87	1.5°

4–8 Dec 2004, 1.75 h live; E_{safe} in TeV

- Not in the training pool
- Mask 0.4° (worst-PSF need: 0.35°)
- Common energy band for > 0.871 TeV
- Inside the condition range but only 15% of training runs at or above its zenith
  → too few runs to train on the field

Background residuals on stacked Crab runs:

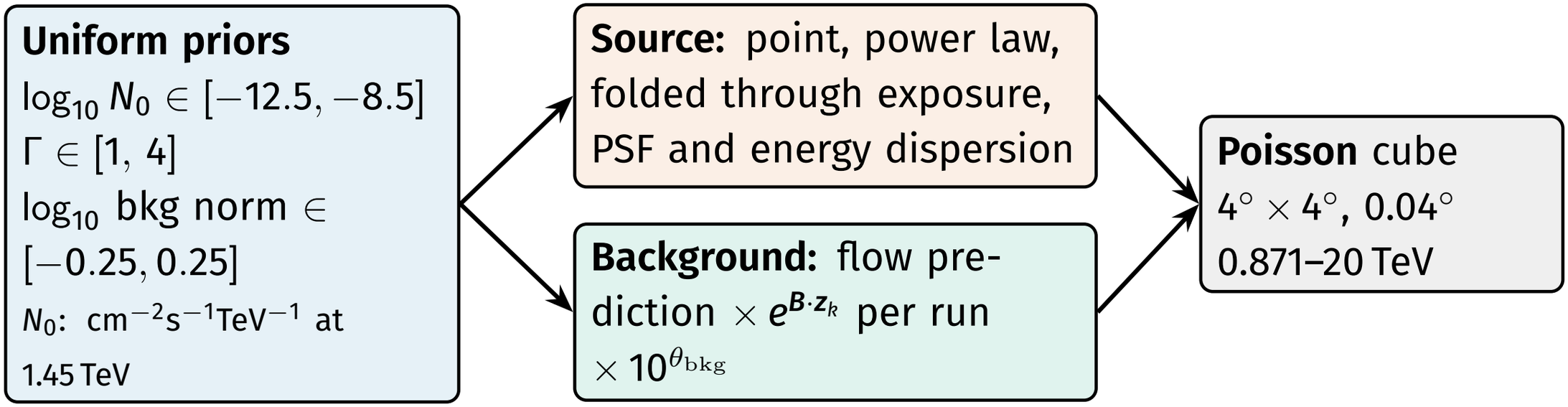




The total rate in all four runs lies above the prediction: We need a free floating background norm in the analysis, i.e., $\rho(c)$ with the available data is not sufficient for total rate prediction.

Simulation-based inference approach to the Crab

The approach needs first a physics simulator (generative CR model + gammapy simulation of Crab from spectral and spatial model) with parameter priors:



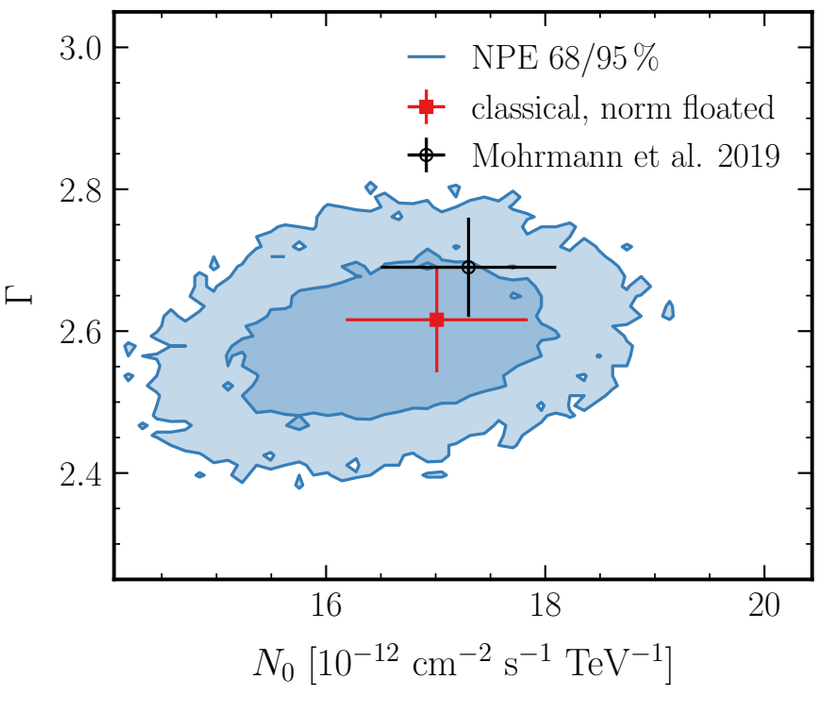
The inverse problem/parameter inference comes via a trainable normalising flow representing the posteriors: neural posterior estimation.

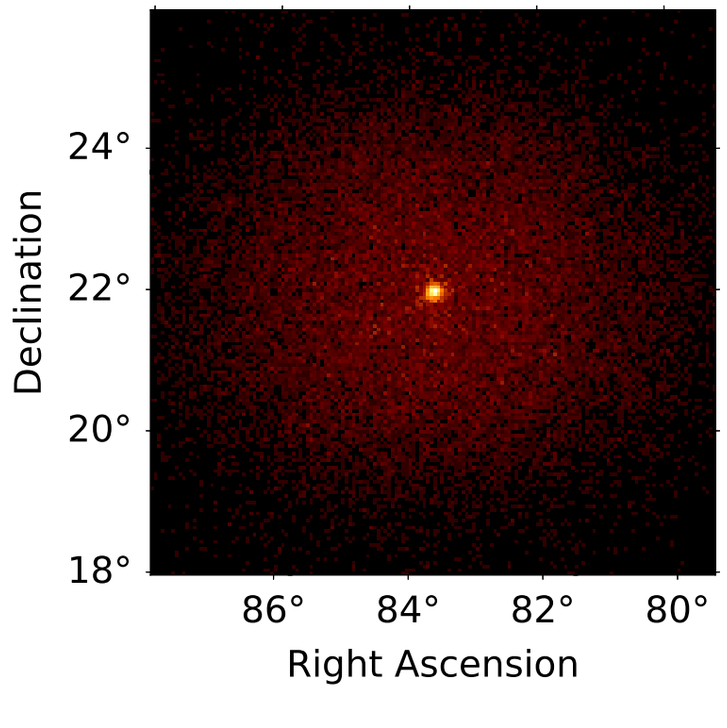
Train a conditional density $q(\theta \mid x)$ on 10,000 simulated pairs; applied to the observed cube it returns the posterior, with z_k and Poisson noise marginalised because they vary across the training set.

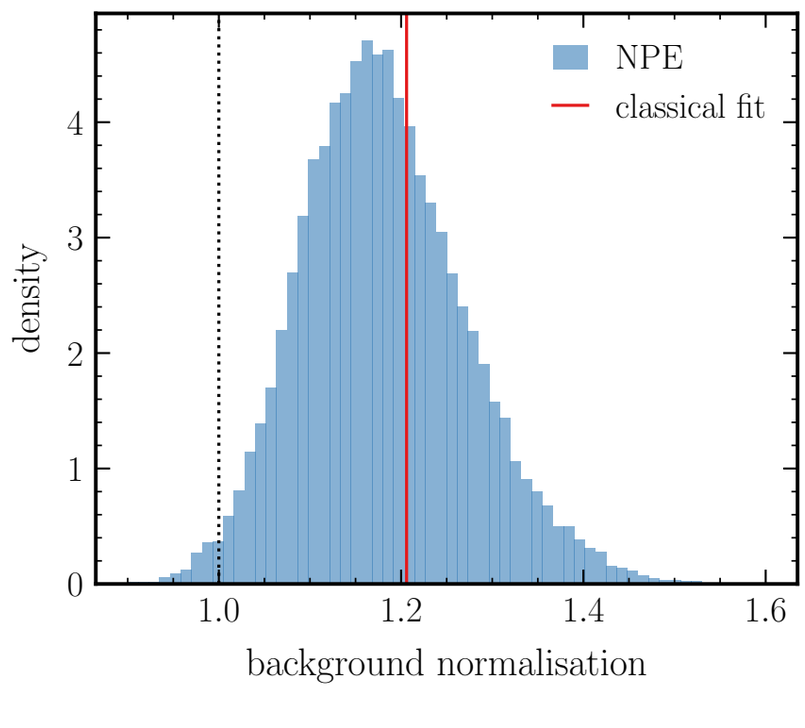
→ *Requires an embedding network that generates a summary statistic from the input cubes (trained alongside normalising flow); here a CNN is sufficient.*

Simulation-based inference approach to the Crab

Results in terms of Bayesian posteriors:







gammapy

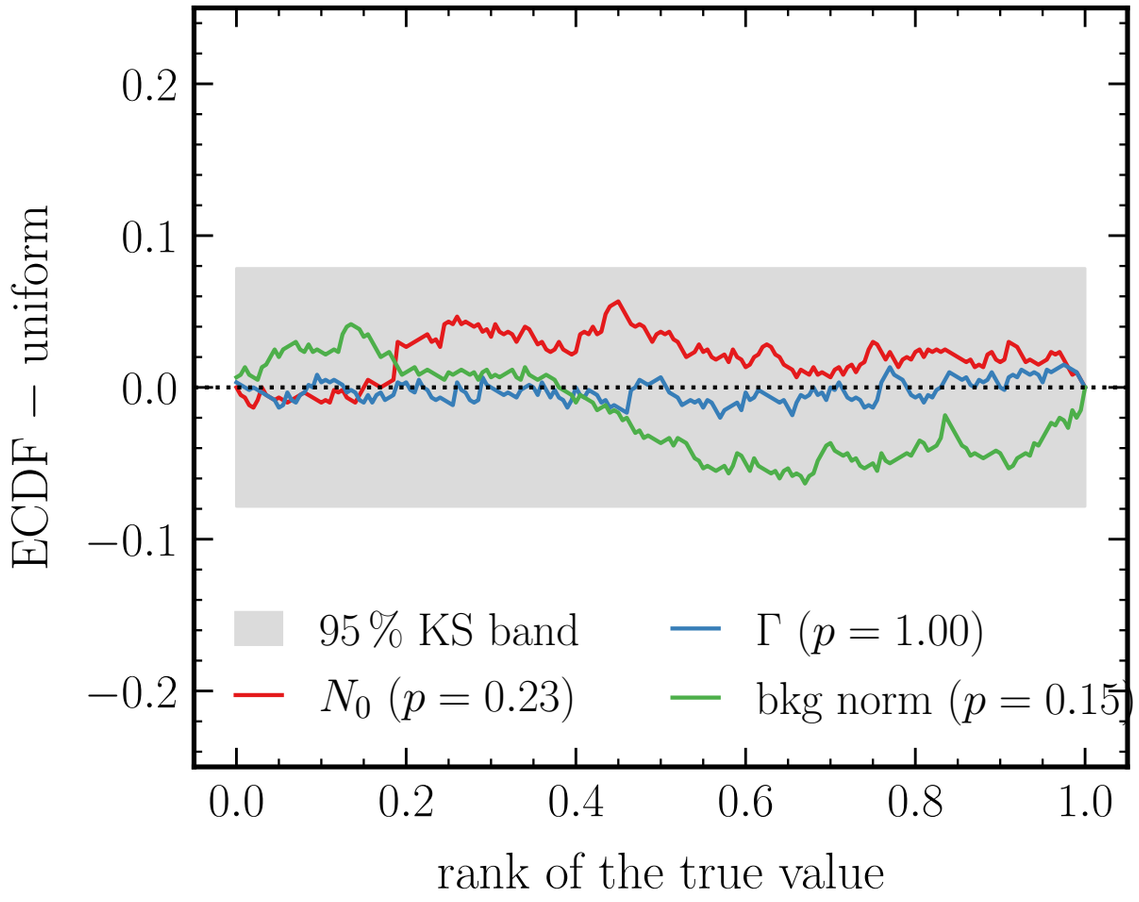
NPE background norm	**Classical 3D fit, norm floated**	**Mohrmann et al. 2019**
1.17 [1.09, 1.27] (median [68 %])	$\Gamma = 2.616 \pm 0.074$, $N_0 = 17.01 \pm 0.83$, norm 1.206	$\Gamma = 2.69 \pm 0.07$, $N_0 = 17.3 \pm 0.8$

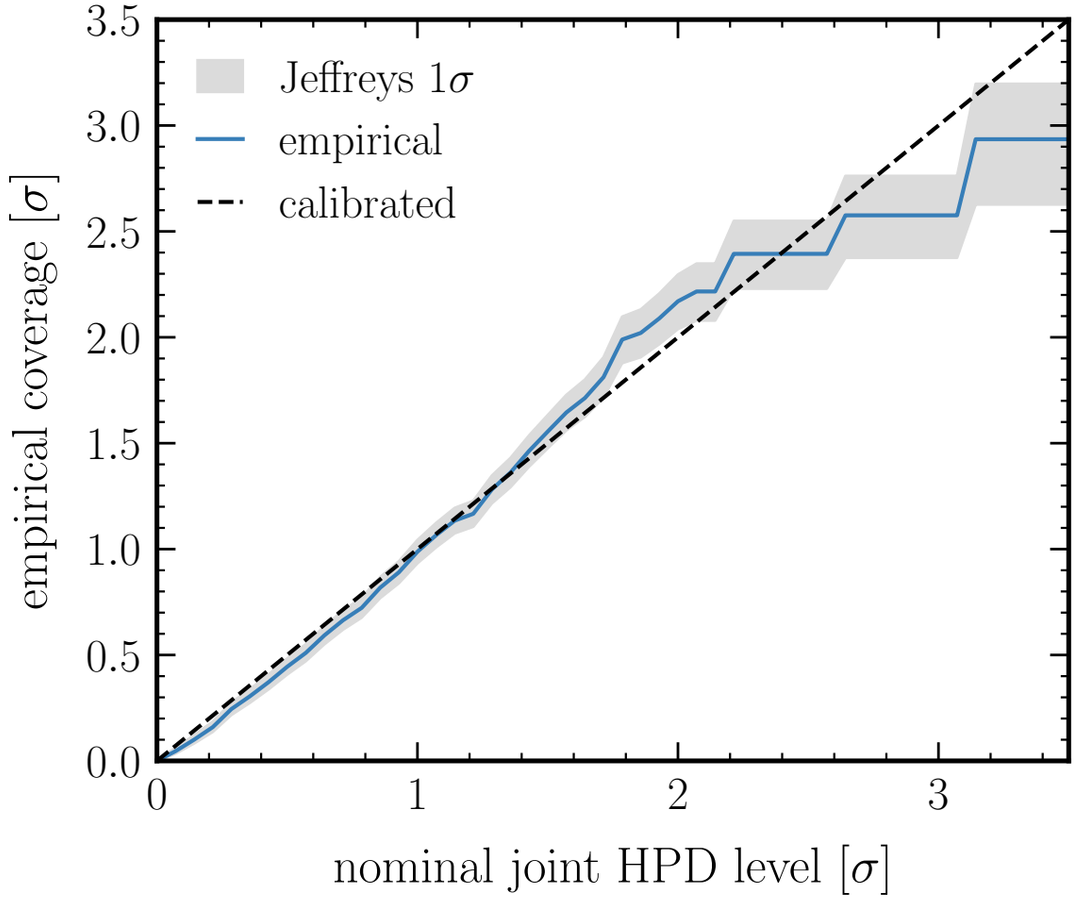
The retrieved posterior encompasses previous literature results and the standard gammapy fit (using the IRF background model). The CR background normalisation is exactly at the level the generative model failed to predict.

Simulation-based inference approach to the Crab

There are ways to validate the calibration of the obtained posteriors via SBI-specific methods (on simulations !): simulation-based calibration and coverage tests.

Basic idea: Prepare a large number of calibration simulations with parameters and datasets disjoint from the training sample. For each, get the posterior and draw a sufficiently large number of samples. Check that the desired statistical properties hold.





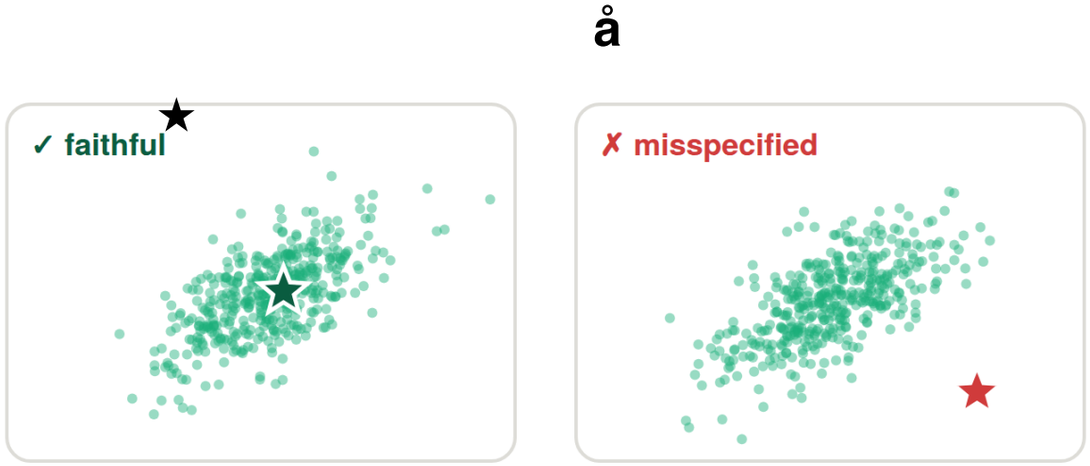
A calibrated posterior does not imply the absence of a reality gap between the simulator and the real data!

The “No free lunch” theorem of SBI

Everything I said in this talk comes with a bill. Here it is.

Theorem. An SBI posterior is only as good as its simulator. Its calibration is **not guaranteed**, and if reality is not in the simulator, the network *extrapolates, confidently and silently.*

Proof. Garbage in → garbage out. □



1. **Calibration:** coverage tests on simulations, always. Cheap once amortised.
2. **Mis-specification:** test the real data against the simulator (anomaly detection, posterior-predictive checks, alternative physics). Needs domain knowledge of where the physics goes wrong.

[N. Anau Montel et al., PRD 111 (2025)] [J. Alvey et al., Mach.Learn.Sci.Tech. 7 (2026) 1]
[L. Tomaselli et al., arXiv:2508.02404] [A. Wehenkel et al., arXiv:2405.08719]

SBI BISTRO

table γ · workshop menu

Intractable likelihood	0.00
Nuisance marginalisation	0.00
Stacking N targets	0.00
Stochastic backgrounds	0.00
Fancy simulator physics	0.00
Calibration checks	DUE
Faithful simulator	DUE
Domain knowledge	priceless
TOTAL	your physics

THANK YOU · COME AGAIN

SBI removes the likelihood, not the physics.

Backup slides