Truncated Marginal Neural Ratio Estimation

Simulation-based Inference Assembly

for

Large Forward Models

ERC CoG UnDark,
Proj. No. 864035

Christoph Weniger, University of Amsterdam

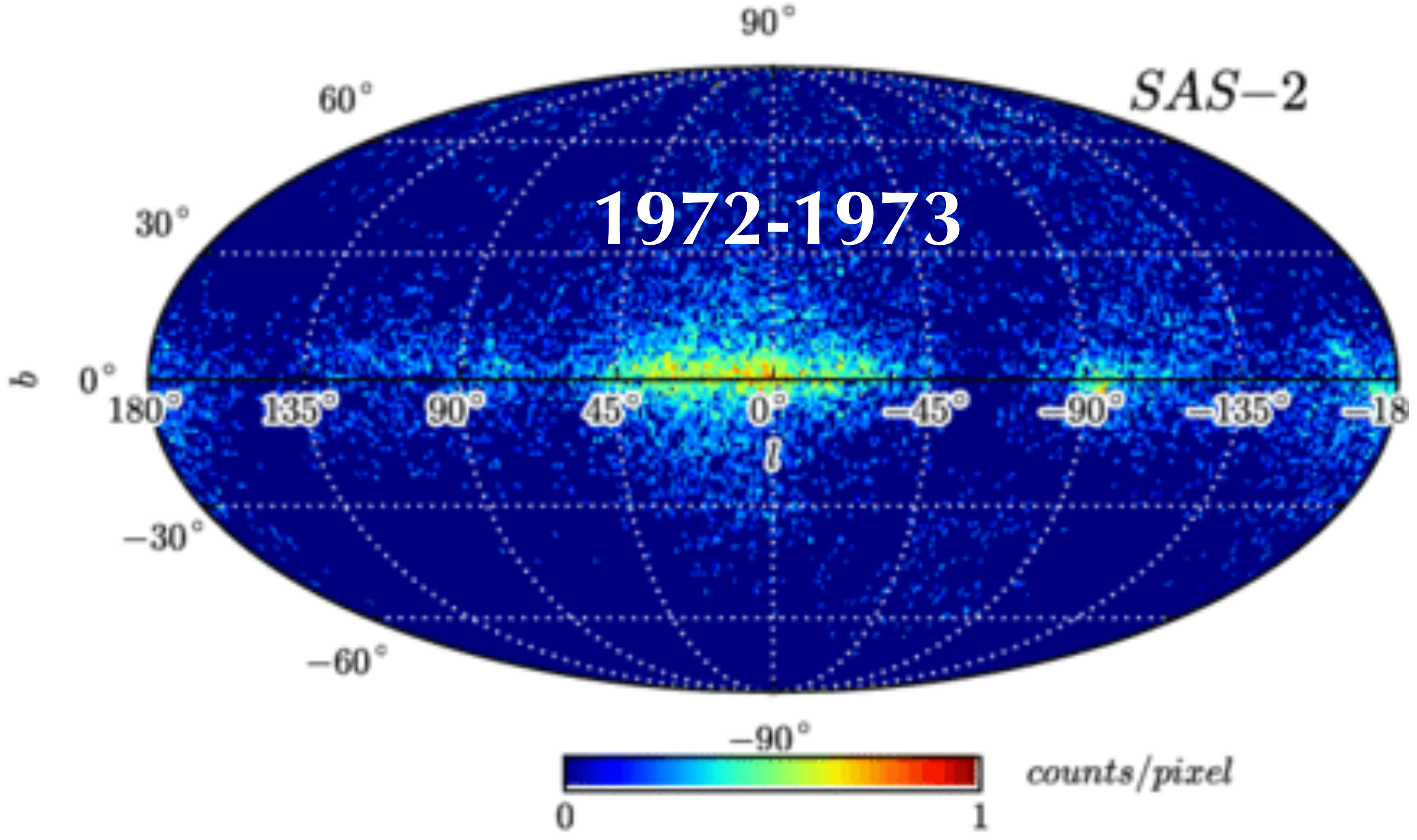
KICC Focus Meeting on Astrostatistics and Astro-Machine Learning

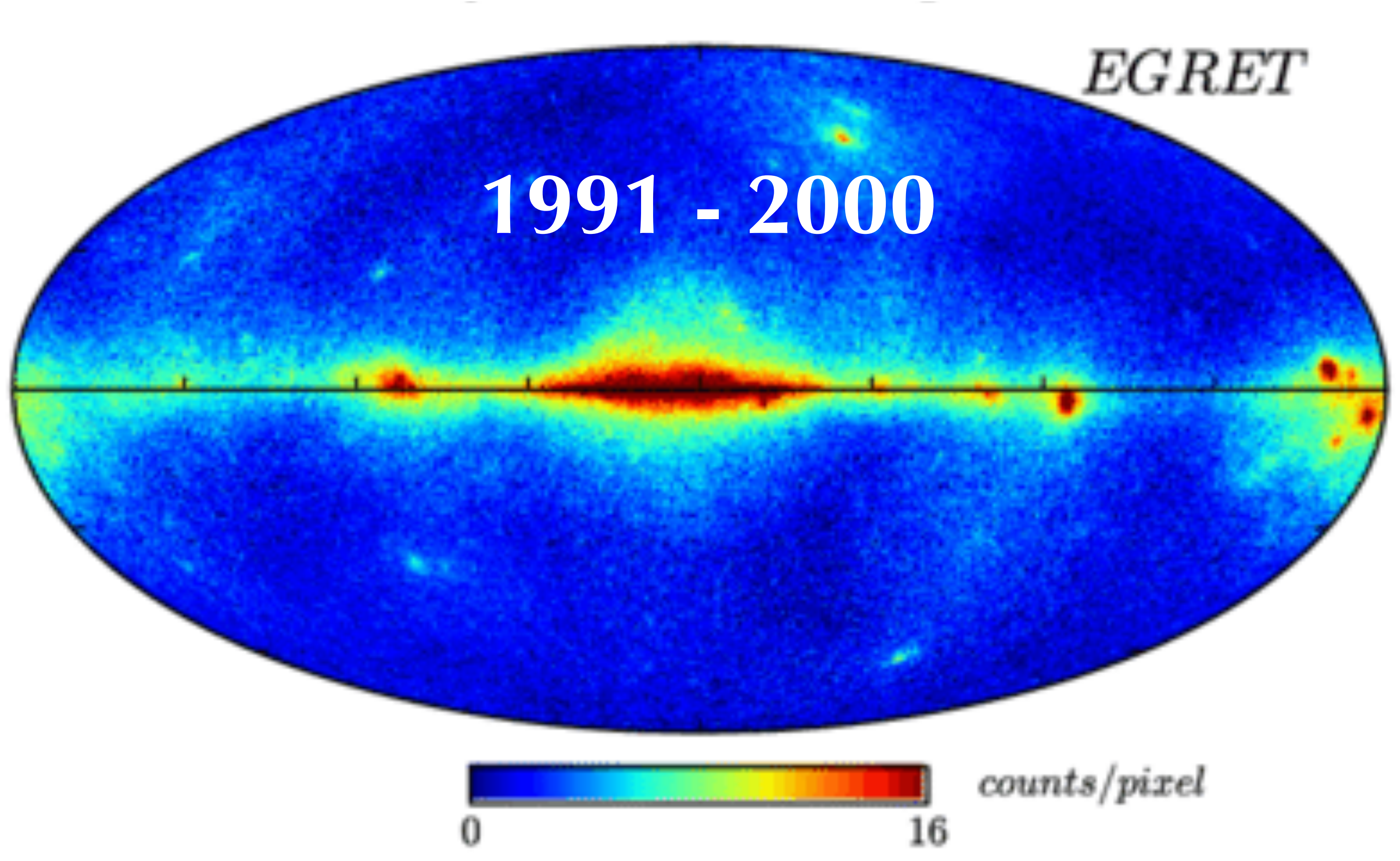
Remote Cambridge, 1 June 2023

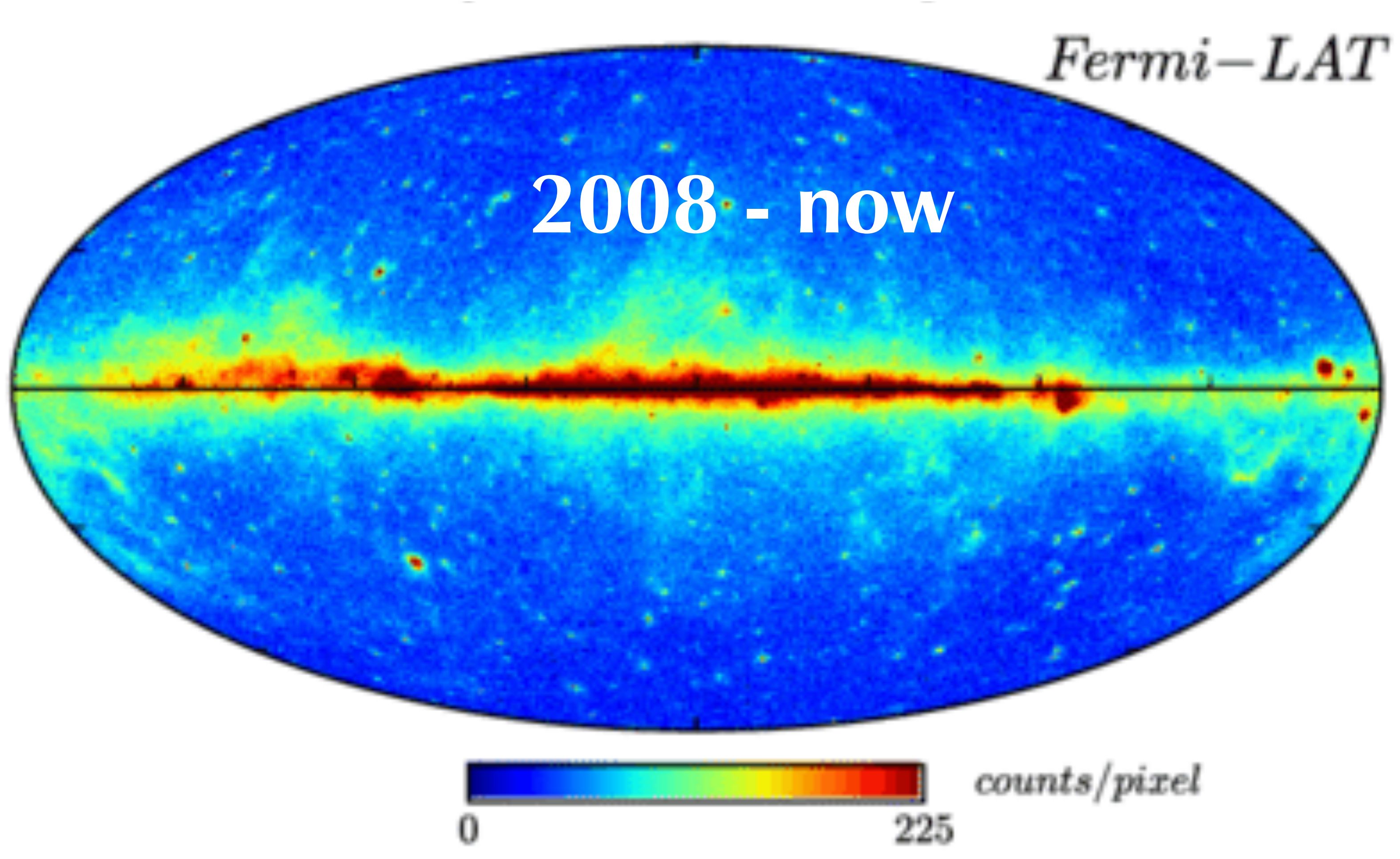
GRAPPA
GRavitation AstroParticle Physics Amsterdam

The challenges of better astrophysical data

More data leads to more objects.
More objects leads to more parameters.
More parameters leads to suffering.







- ~6 Gamma-ray sources
- Diffuse emission

- 188 Gamma-ray sources
- Extended point sources
- Resolved gas emission

- 6658 Gamma-ray sources
- Highly resolved gas
- Many extended sources
- Fermi bubbles
- GeV excess
- Dark gas
- …

Casandjian 0806.0113
Abdollahi+ 2201.11184

“Large forward model”

A forward model generates samples $\mathbf{x}, \mathbf{z} \sim p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$. What makes forward models *large*?

1. **Model Complexity**
   Number of parameters, variables and components. A “large” model in that sense includes many different factors and potentially complex interactions between them.

2. **Amount of Data**
   Different data sources, volume of data. A “large” model in that sense would be one that can make predictions based on vast amounts of observational data.

3. **Computational Resources**
   A “large” model requires significant computational power, memory, or time to run.

Observation: If a model is large enough, joined inference of $p(\mathbf{z} \mid \mathbf{x})$ becomes intractable.

Question: Can we develop practical alternatives to joined inference and still do science?

Truncated Marginal Neural Ratio Estimation

Neural Ratio Estimation (NRE)

NRE turns parameter inference into binary classification, and generates maximally informative* data summaries.

Generate training data

$\mathscr{D} \equiv \{(\mathbf{x}_i, \mathbf{z}_i) \mid i = 1,2,\ldots,N\}$, where $\mathbf{x}, \mathbf{z} \sim p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$

Initialise real-valued neural network, $f_\phi(\mathbf{x}, \mathbf{z})$

Train neural network using the mini-batch loss function

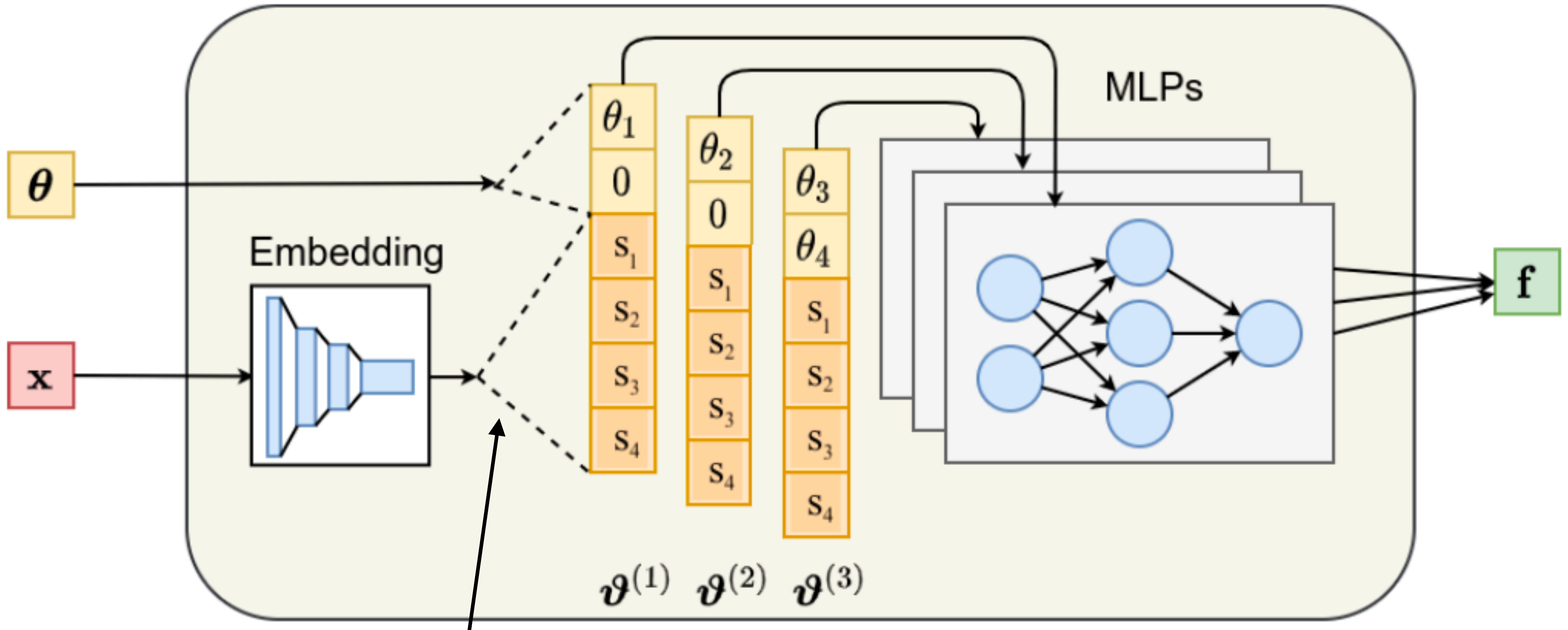
$$\mathscr{L}(\phi) = -\sum_{i \in B} \ln \sigma(f_\phi(\mathbf{x}_i, \mathbf{z}_i)) + \ln \sigma(-f_\phi(\mathbf{x}_i, \mathbf{z}_{P(i)}))$$

$Y = 1 : \mathbf{x}, \mathbf{z} \sim p(\mathbf{x}, \mathbf{z})$ $\qquad$ $Y = 0 : \mathbf{x}, \mathbf{z} \sim p(\mathbf{x})p(\mathbf{z})$

Here, B denotes a mini-batch, and P denotes random sample permutations.

After training, $f_\phi(\mathbf{x}, \mathbf{z}) \approx \ln r(\mathbf{x}; \mathbf{z}) = \ln \dfrac{p(\mathbf{z} \mid \mathbf{x})}{p(\mathbf{z})}$

Example network



Automatically learned data summaries

*Data summary maximises distance between $p(\mathbf{z} \mid \mathbf{s}(\mathbf{x}))$ and $p(\mathbf{z})$ in terms of JS divergence.

Hermans+ 1903.04057 (neural ratio estimation)

Implicit likelihood (SBI, simulation-based inference)

Training NN* in the context of SBI
=> **Automatic super-charged ABC**.

Obtain **optimal data summary**

$$t(\mathbf{x}) \equiv \mathrm{NN}^{(1)}_{\phi}(\mathbf{x})$$

such that

$$p(z\,|\,t(\mathbf{x})) \simeq p(z\,|\,\mathbf{x}) \ .$$

Estimate **posterior** ($\epsilon \to 0$),

$$r(\mathbf{x}, z) \equiv \mathrm{NN}^{(2)}_{\phi}(x, t(\mathbf{x})) \simeq \frac{p(z, t)}{p(z)p(t)}$$

such that

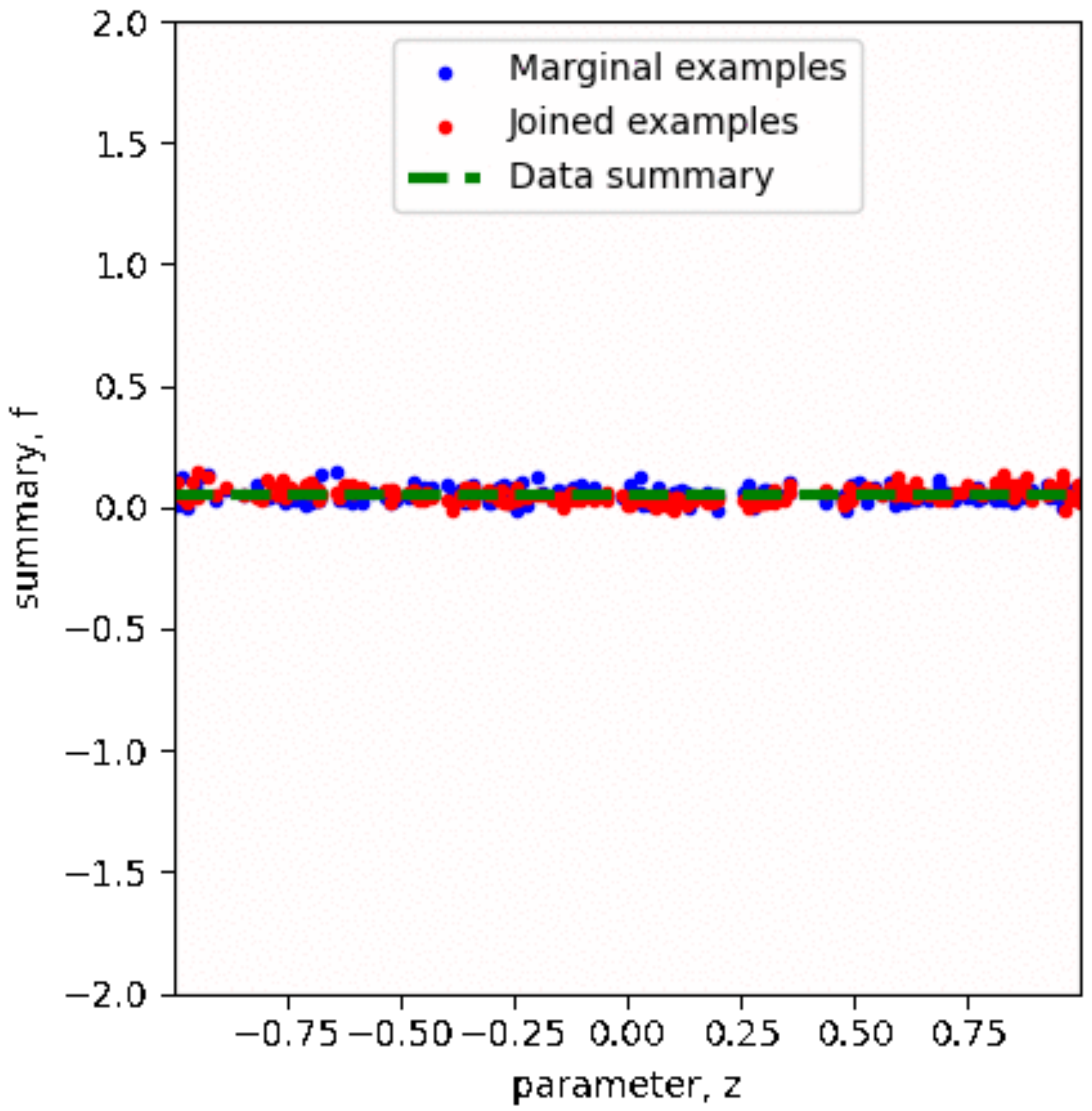
$$r(\mathbf{x}, z) \simeq \frac{p(z\,|\,\mathbf{x})}{p(z)} \ .$$

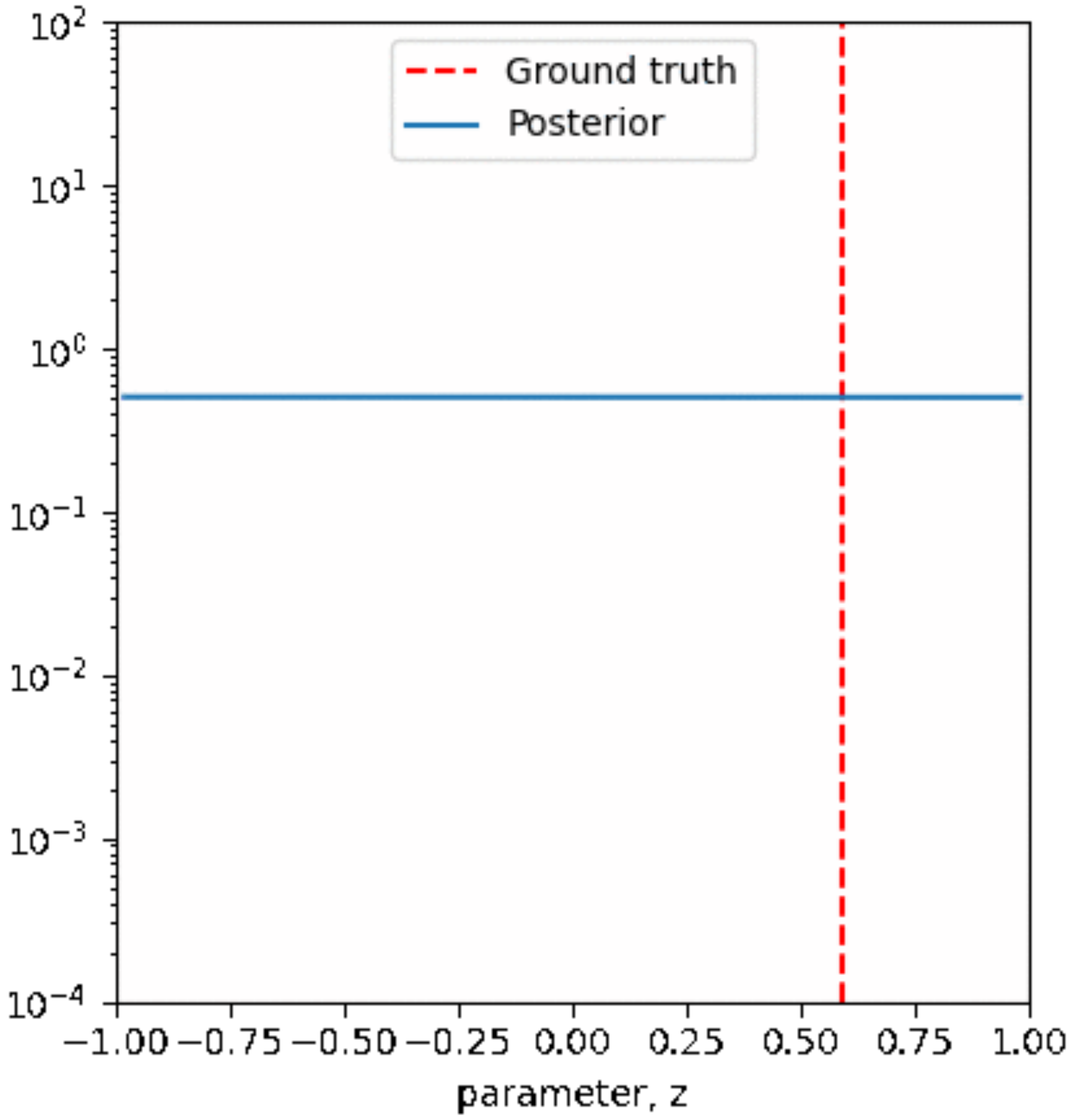
Points

Contour

Neural ratio estimation (NRE)
Train a neural network to discriminate

- Real sims: $z, \mathbf{x} \sim p(\mathbf{x}\,|\,z)p(z)$
- Scrambled sims: $z, \mathbf{x} \sim p(\mathbf{x})p(z)$





Toy model: $\mathbf{x} = \mathbf{v} \cdot z^2 + \boldsymbol{\epsilon}$

*I discuss here Likelihood-to-evidence neural ratio estimation. There are multiple approaches that have the same asymptotic behaviour. Training is here done using a simple binary cross entropy loss, there are multiple other approaches with the same goal, which work however for similar reasons; see e.g. Cranmer+ 1911.0142 for a review

Gutman&Hyvärinen 2010 (as NCE), Mnih&Teh 2012 (self-normalizing),
Hermans+ 1903.04057 (neural ratio estimation)
Miller+ 2107.01214

Marginal NRE (MNRE)

Often we are **interested in marginal posteriors.** Estimating the joined is only a intermediate step.

Statistical model $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$

Joined inference

Marginal SBI

$\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{x})$

Projection

$h \sim p(h \mid \mathbf{x})$

$h = h(\mathbf{z})$

Result for parameter summary of interest, h

- 1-dim marginals for **parameter inference**
  $p(z_3 \mid \mathbf{x})$
- 1-dim conditional marginals for correlation studies
  $p(z_4 \mid \mathbf{x}, z_3, z_5, z_9)$
- 2-dim marginals for studying **parameter correlations**
  $p(z_4, z_7 \mid \mathbf{x})$
- Classification probabilities for **model comparison**
  $p(M_1 \mid \mathbf{x}), p(M_2 \mid \mathbf{x})$
- Probabilities for **object detection**
  $p(n_i \geq 1 \mid \mathbf{x})$
- Probabilities for **source detection**
  $p(F_i > F_{th} \mid \mathbf{x}, n_i \geq 1)$

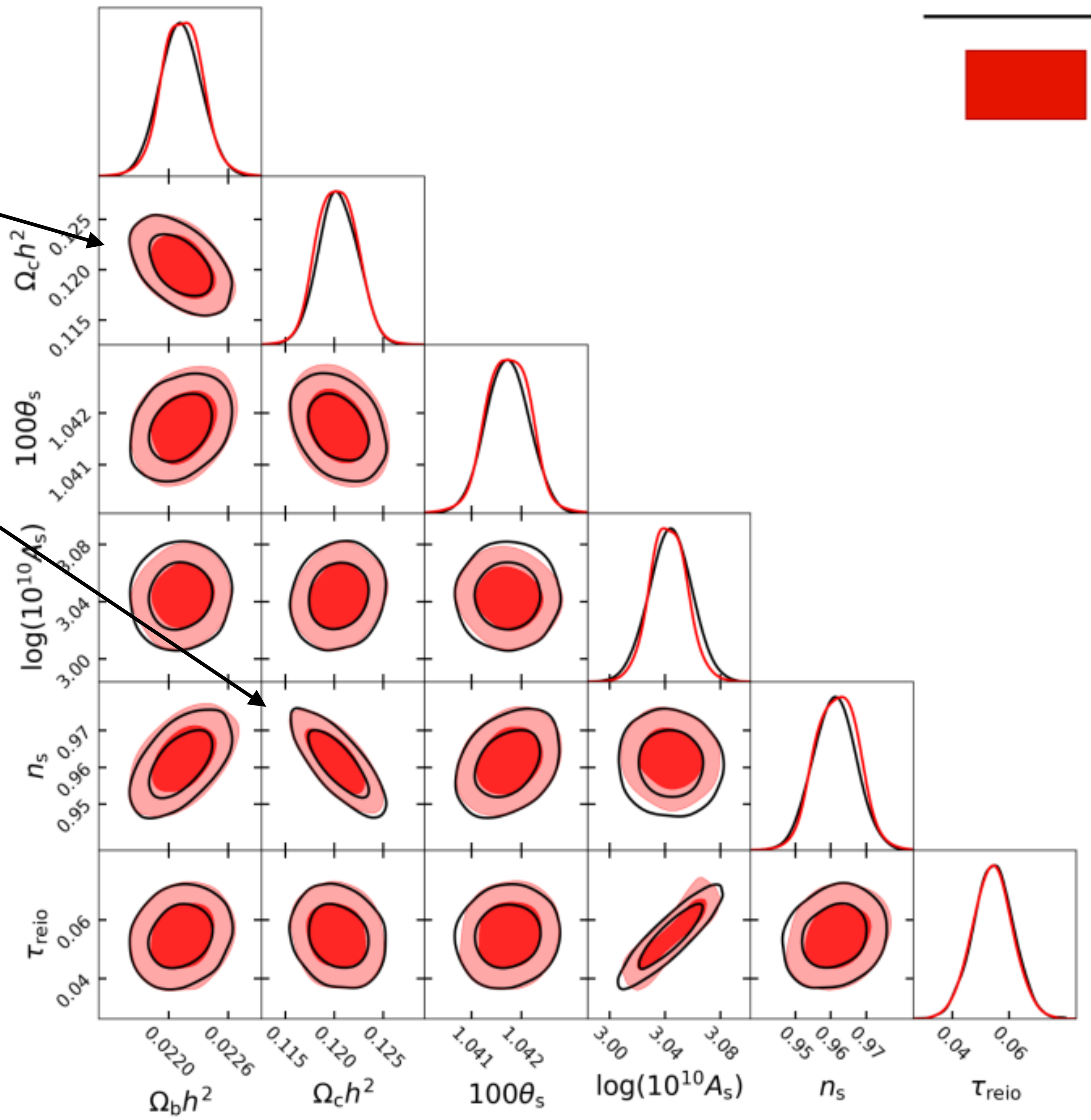
Open questions: Can we do anything with marginal inference that we could do with the joined?
(goodness of fit tests, posterior predictive distribution, …)

Example MNRE: Planck cosmology

MNRE (and SBI in general) recovers **correct marginal posteriors** in the presence of correlations.

$f_\phi(\mathbf{x}; \Omega_c, \Omega_b)$

$f_\phi(\mathbf{x}; n_s, \Omega_c)$



swyft	parameters	MCMC
3×10^3 simulations		5×10^6 simulations
Seen by swyft	ω_b, ω_{cdm}, $100\theta_s$, $\ln(10^{10}A_s)$, n_s, τ	Sampled by MCMC
Ignored by swyft	A_{radio}, A_{dusty}, A_{dustTT}, A_{SZ}, A_{CIB}, A_{kSZ}, $A_{SZ\times CIB}$, A_{Planck}, c_{100A}, c_{100B}, c_{143B}, c_{217A}, c_{217B}	

Based on
Planck HiLLiPoP likelihood

Cole+ 2111.08030

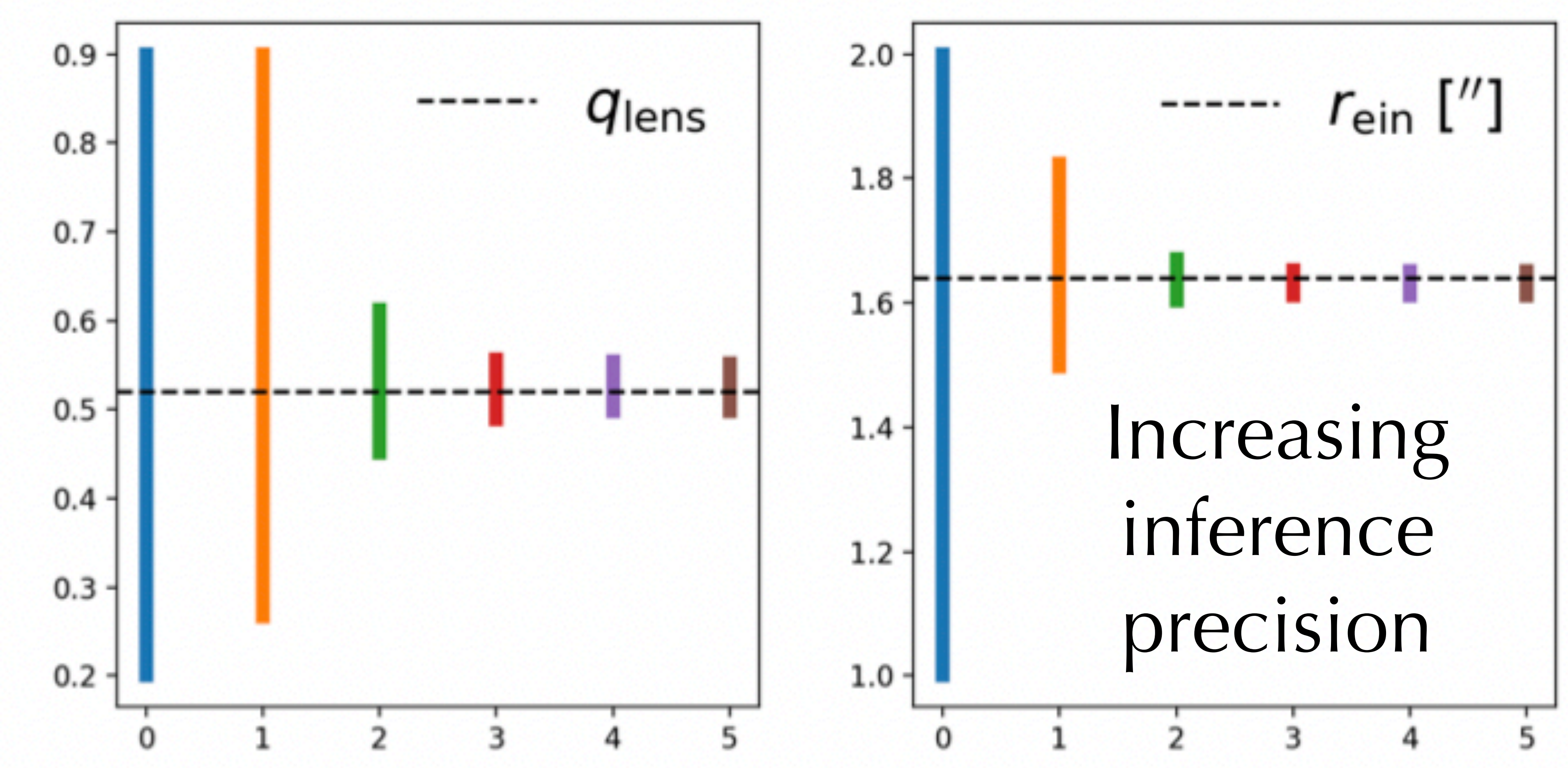
Sequential NRE (SNRE)

Hermans+ 2020, Durkan+ 2020, Delaunoy+ 2022, Miller+ 2022

Sequential inference aims at increasing inference precision by employing **proposal distributions $\tilde{p}(\mathbf{z})$ that focuses training data** on relevant parts of the parameter space.

Example

- Strong lensing images
- Model with ~25 parameters, each truncated separately.
- We use 6 rounds, the final training data looks pretty much like the observation.



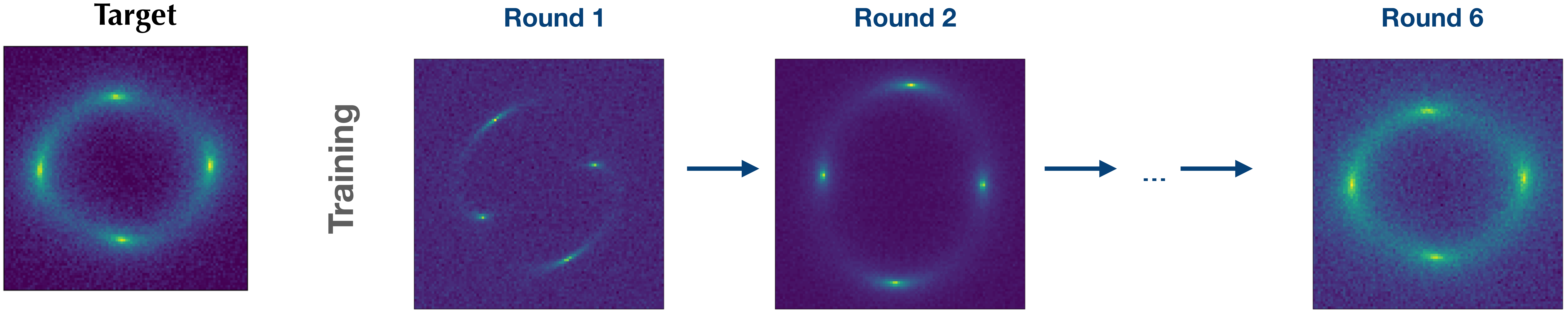


Image credit: Noemi Anau Montel

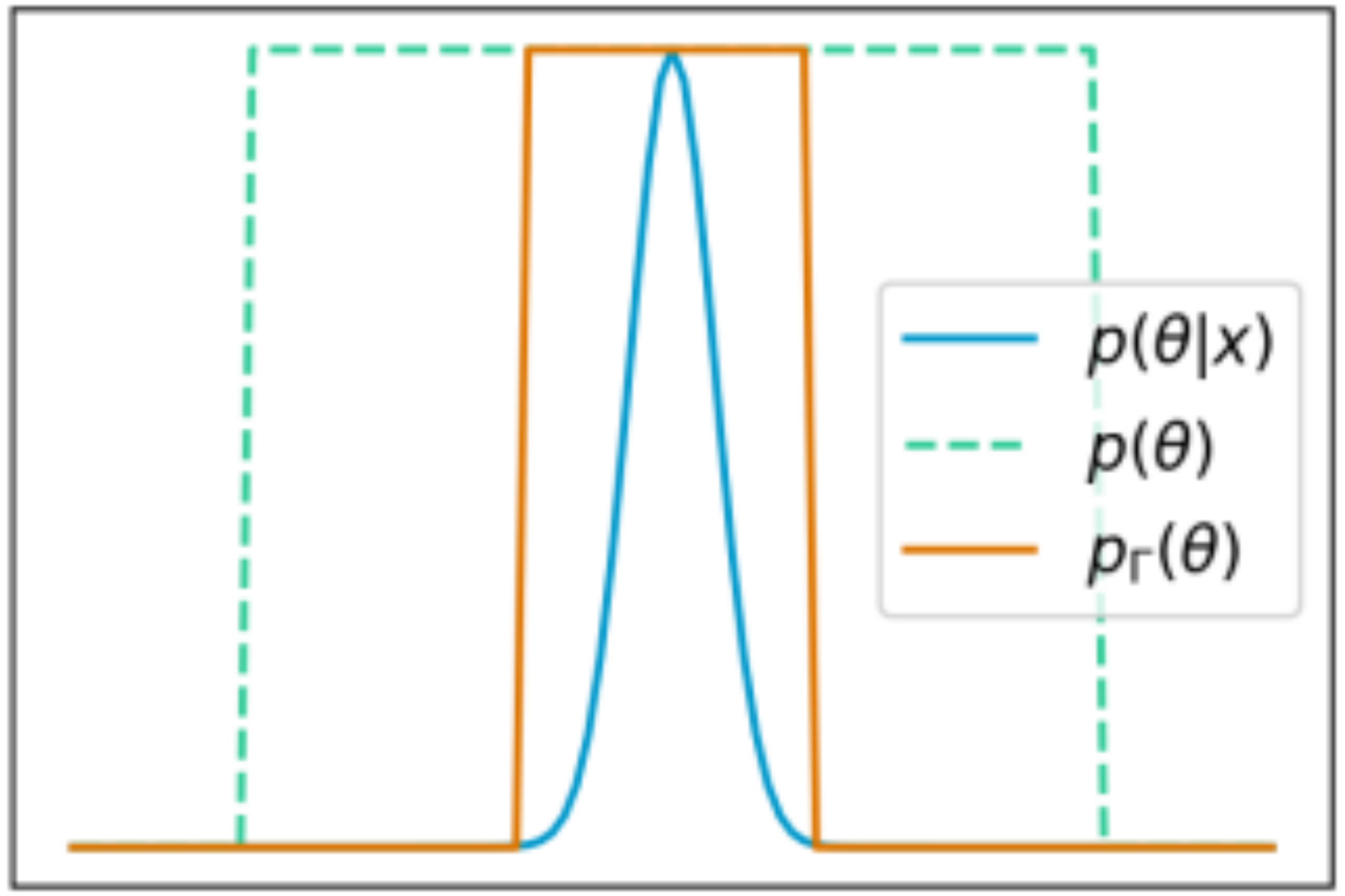
SNRE & MNRE ⇒ Truncated Marginal Neural Ratio Estimation (TMNRE)

Combination of marginal and sequential inference requires truncated priors as proposals.

$$\tilde{p}(\mathbf{z}) = \frac{1}{Z}\mathbb{I}(\mathbf{z} \in \Gamma)p(\mathbf{z})$$

$\mathbb{I}(\mathbf{z} \in \Gamma)$: **Indicator function,** which only allows values within the truncation region Γ

$$\Gamma = \{\mathbf{z} \in \mathbb{R}^N : \tilde{r}(\mathbf{x};\mathbf{z}) > \epsilon\} \qquad \tilde{r}(\mathbf{x};\mathbf{z}) \approx \frac{p(\mathbf{x}\,|\,\mathbf{z})}{p(\mathbf{x})}$$

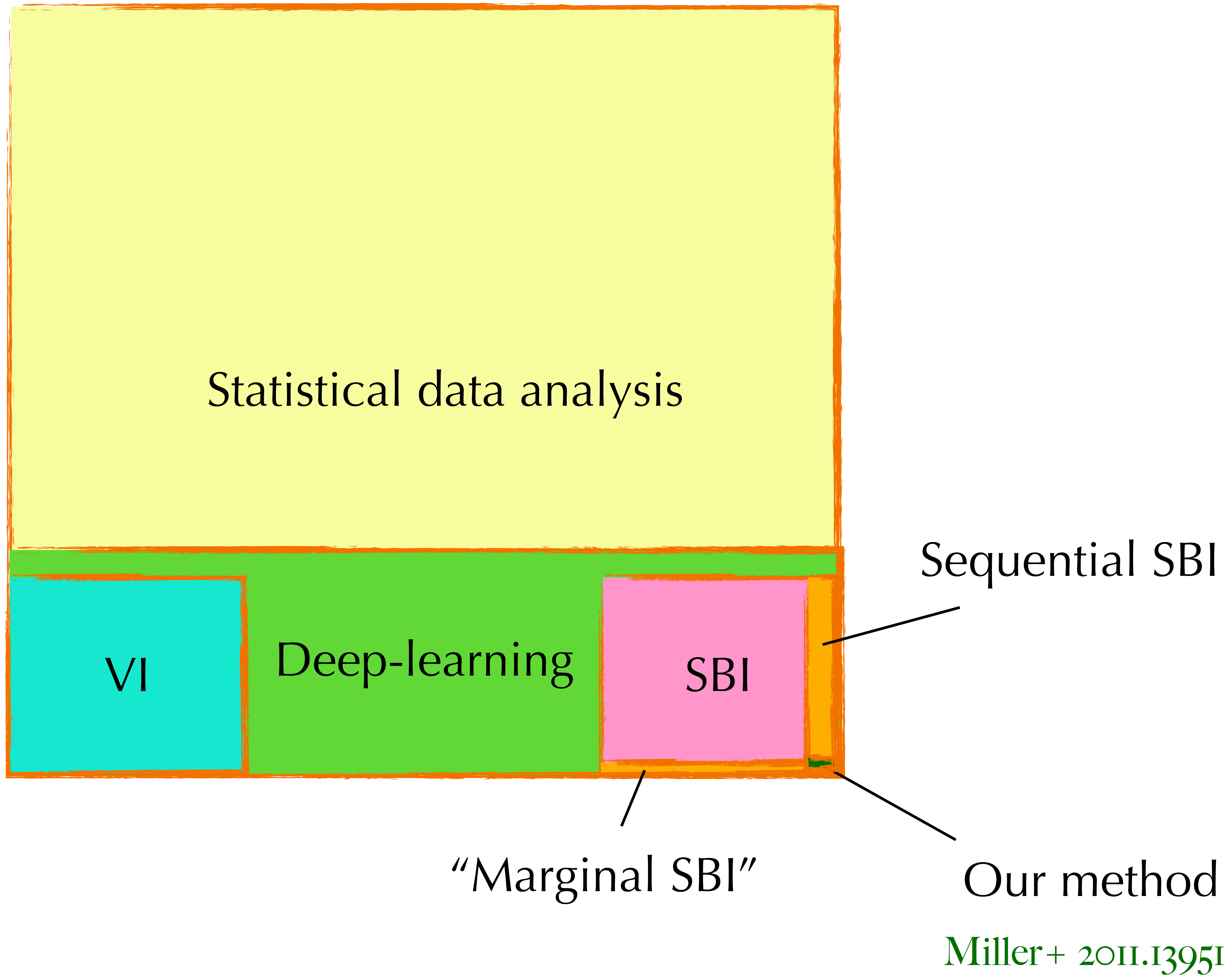


We define the truncation region as the region with high estimated likelihood-to-evidence ratio $\tilde{r}(\mathbf{x};\mathbf{z})$, and exclude regions with very low likelihood.

Open questions: How to best truncate under different circumstances? Parameter-wise?

[Miller+ 2011.13951, 2107.01214 - swyft & TMNRE]

Pretty niche (but growing exponentially?)



Rate of papers using TMNRE is growing exponentially

2021

1. "Fast and Credible Likelihood-Free **Cosmology** with Truncated Marginal Neural Ratio Estimation" Cole+ 2111.08030

2022

2. "Estimating the warm dark matter mass from **strong lensing** images with truncated marginal neural ratio estimation" Anau Montel+, 2205.09126
3. "SICRET: **Supernova Ia Cosmology** with truncated marginal neural Ratio EsTimation" Karchev+2209.06733 **Talk: Kosio Karchev**
4. "One never walks alone: the effect of the perturber population on subhalo measurements in **strong gravitational lenses**" Coogan+ 2209.09918
5. "Detection is truncation: studying **source populations** with truncated marginal neural ratio estimation" Anau Montel+ 2211.04291

2023 (first ~3 months)

6. "Debiasing **Standard Siren Inference** of the Hubble Constant with Marginal Neural Ratio Estimation" Gagnon-Hartman+ 2301.05241
7. "Constraining the X-ray heating and reionization using **21-cm power spectra** with Marginal Neural Ratio Estimation" Saxena+ 2303.07339
8. "Peregrine: Sequential simulation-based inference for **gravitational wave signals**", Bhardwaj+ 2304.02035 **Talk: James Alvey**
9. "Albatross: A scalable simulation-based inference pipeline for analysing **stellar streams** in the Milky Way", Alvey+ 2304.02032

Applications

1) Sequential SBI for gravitational wave

TRUNCATED MARGINAL NEURAL RATIO ESTIMATION (TMNRE) WITH PEREGRINE

Gravitational Wave Parameter Inference with Neural Ratio Estimation

STEP 1: (RE-) SIMULATE

- Sample parameters $\boldsymbol{\theta}_{\mathrm{GW}}$ from (truncated) prior $p(\boldsymbol{\theta}_{\mathrm{GW}})$
- Simulate data $x \sim p(x \mid \boldsymbol{\theta}_{\mathrm{GW}})$

$$\boldsymbol{\theta}_{\mathrm{GW}} \equiv (q, \ldots) \to x(\boldsymbol{\theta}_{\mathrm{GW}}) = h(\boldsymbol{\theta}_{\mathrm{GW}}) + n_{\mathrm{IFO}}$$

STEP 2: RATIO ESTIMATION

- Train ratio estimators $r(x; \vartheta)$ on simulated data to approximate the posterior-to-prior ratio $r(x; \vartheta) \sim p(\vartheta \mid x)/p(\vartheta)$ for each set of parameters of interest $\vartheta = (\theta_i, \theta_j, \ldots)$

STEP 3: INFERENCE

- Obtain a prior sample from $p(\vartheta)$
- Target a specific observation x_0 and compute the ratios $r(x_0; \vartheta)$ across the prior sample
- Weight the samples according to this ratio

STEP 4: TRUNCATION

- Use ratios $r(x_0; \vartheta)$ to remove regions with vanishingly low posterior density
- If this leads to a reduction in prior volume, re-simulate from **Step 1** with truncated prior
- Repeat until converged across all parameters, then obtain resulting posteriors

Bhardwaj+ 2304.02035

Related work: Dax+ 2106.12594

Sequential inference in multiple rounds

The first two truncation rounds

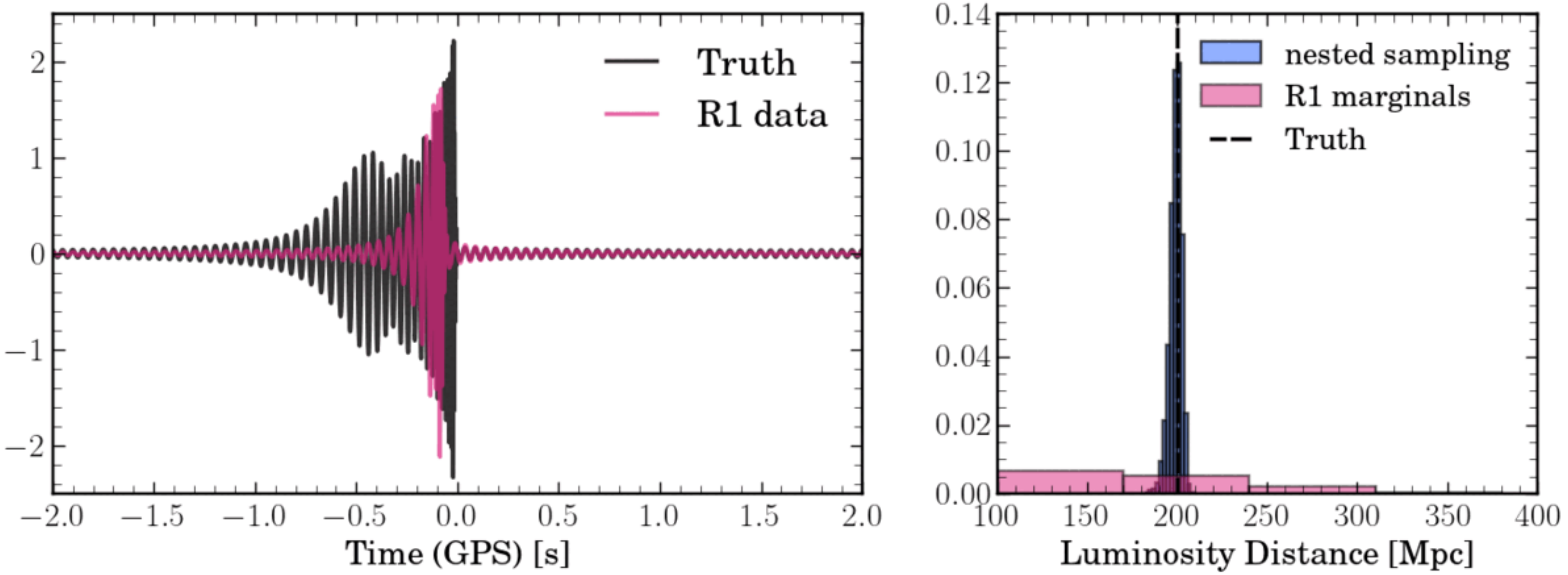
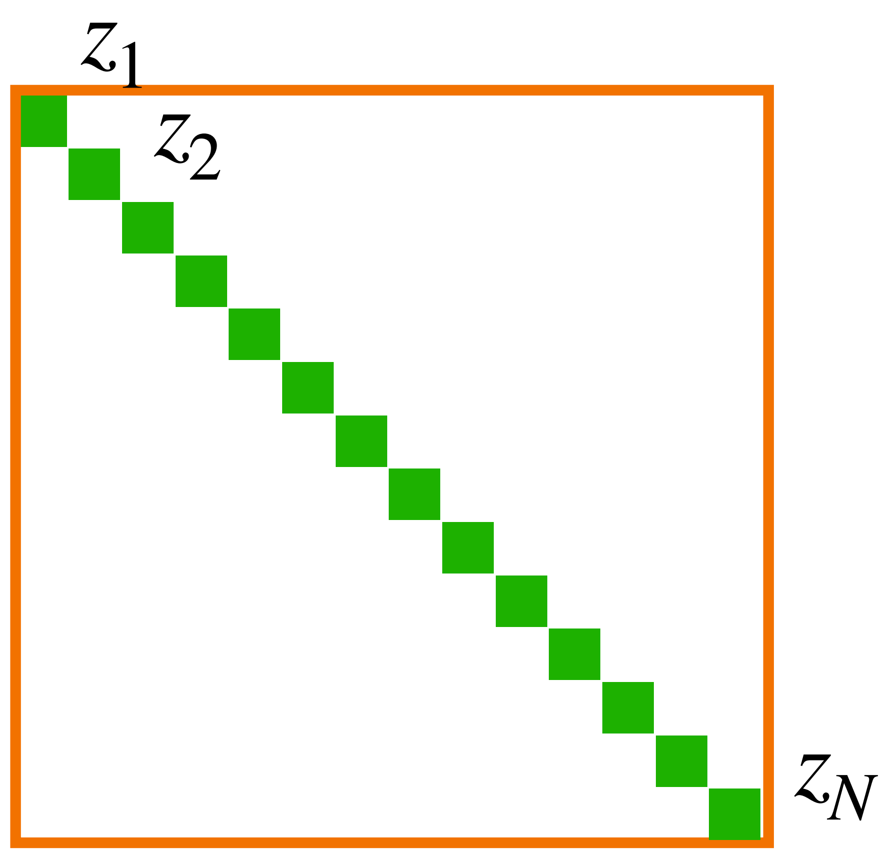
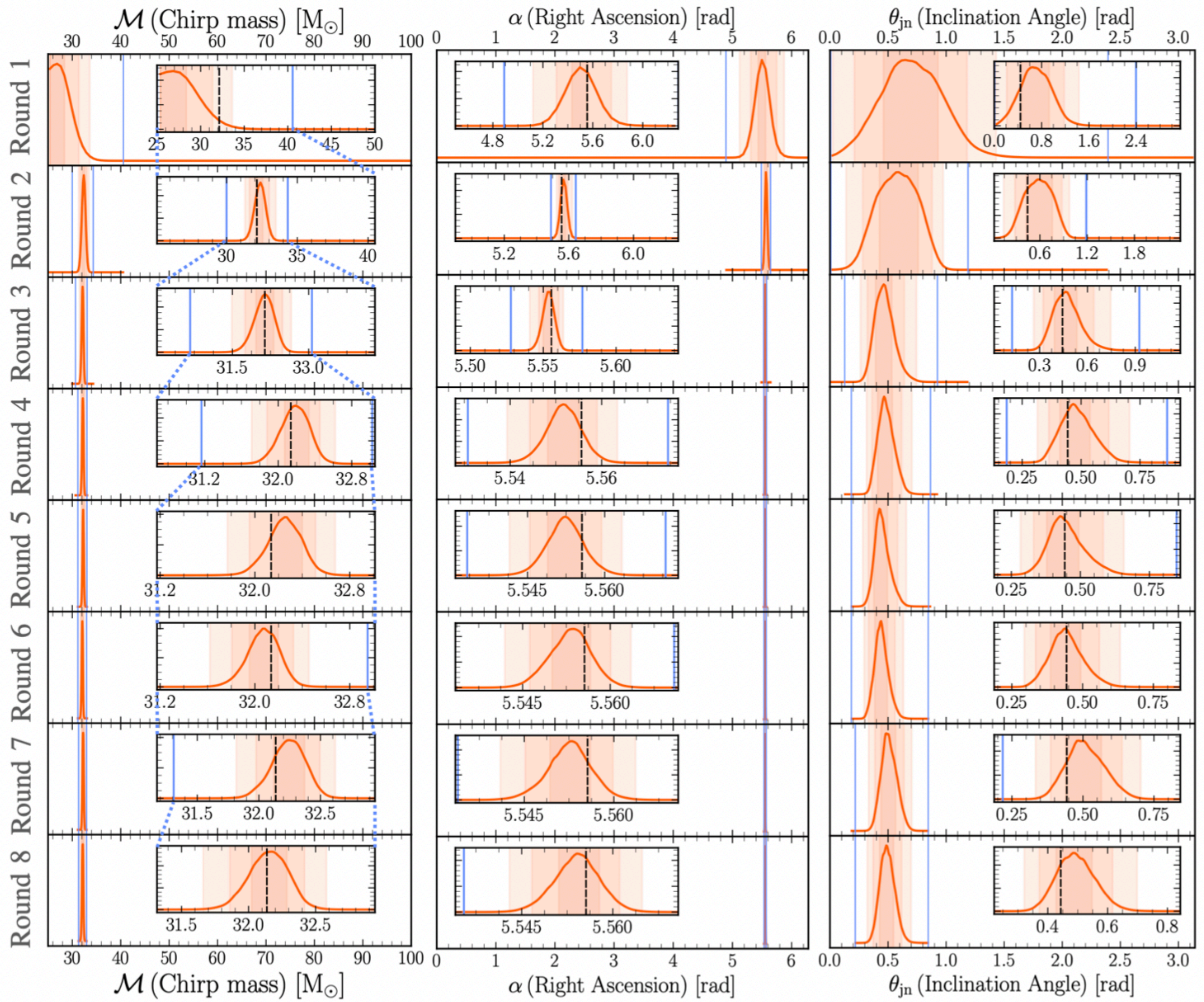


Image credit: Udditap Bhardwaj

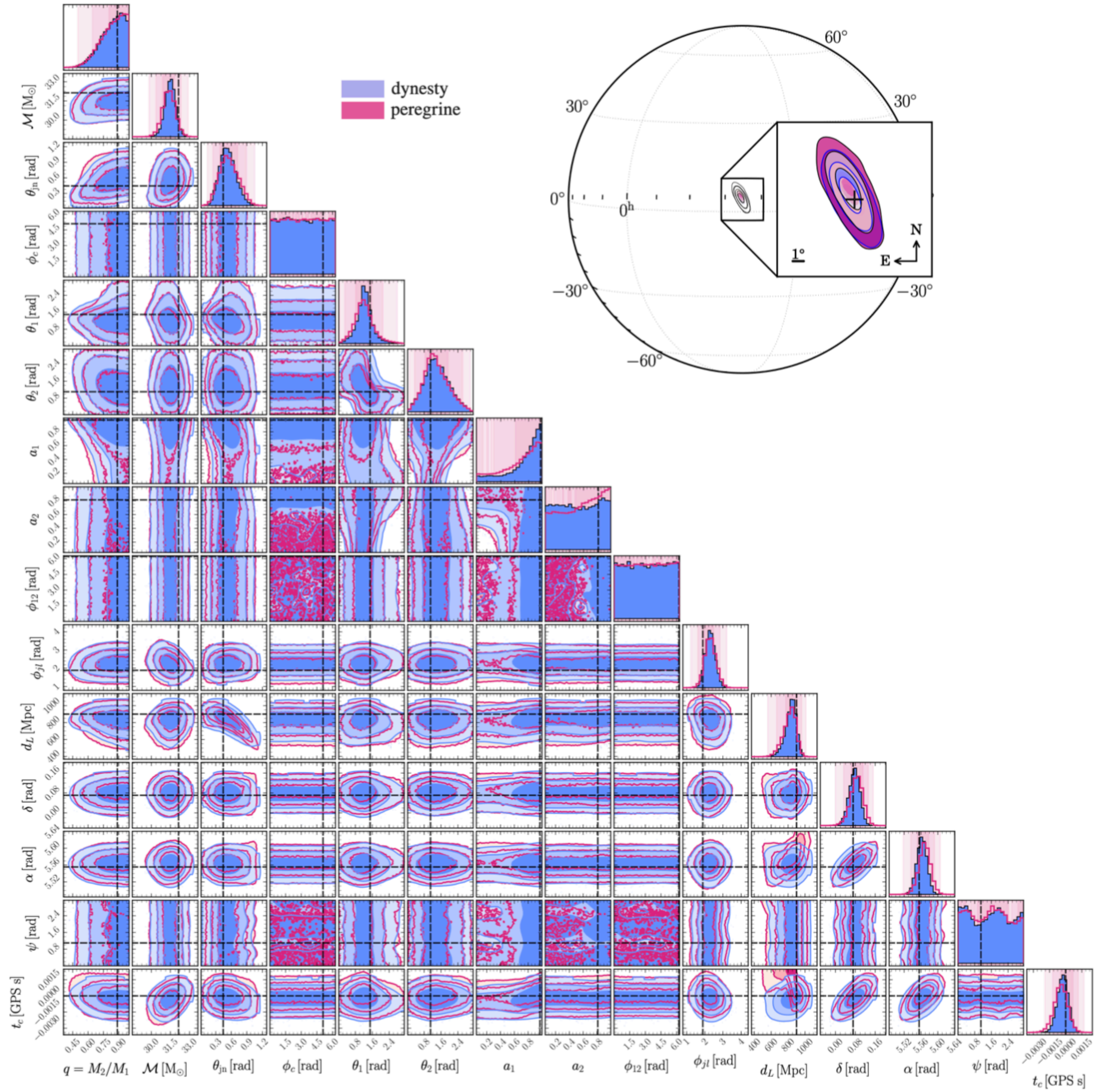
Truncation scheme: Neglect correlations





Bhardwaj+ 2304.02035

Inference includes correlations



Results

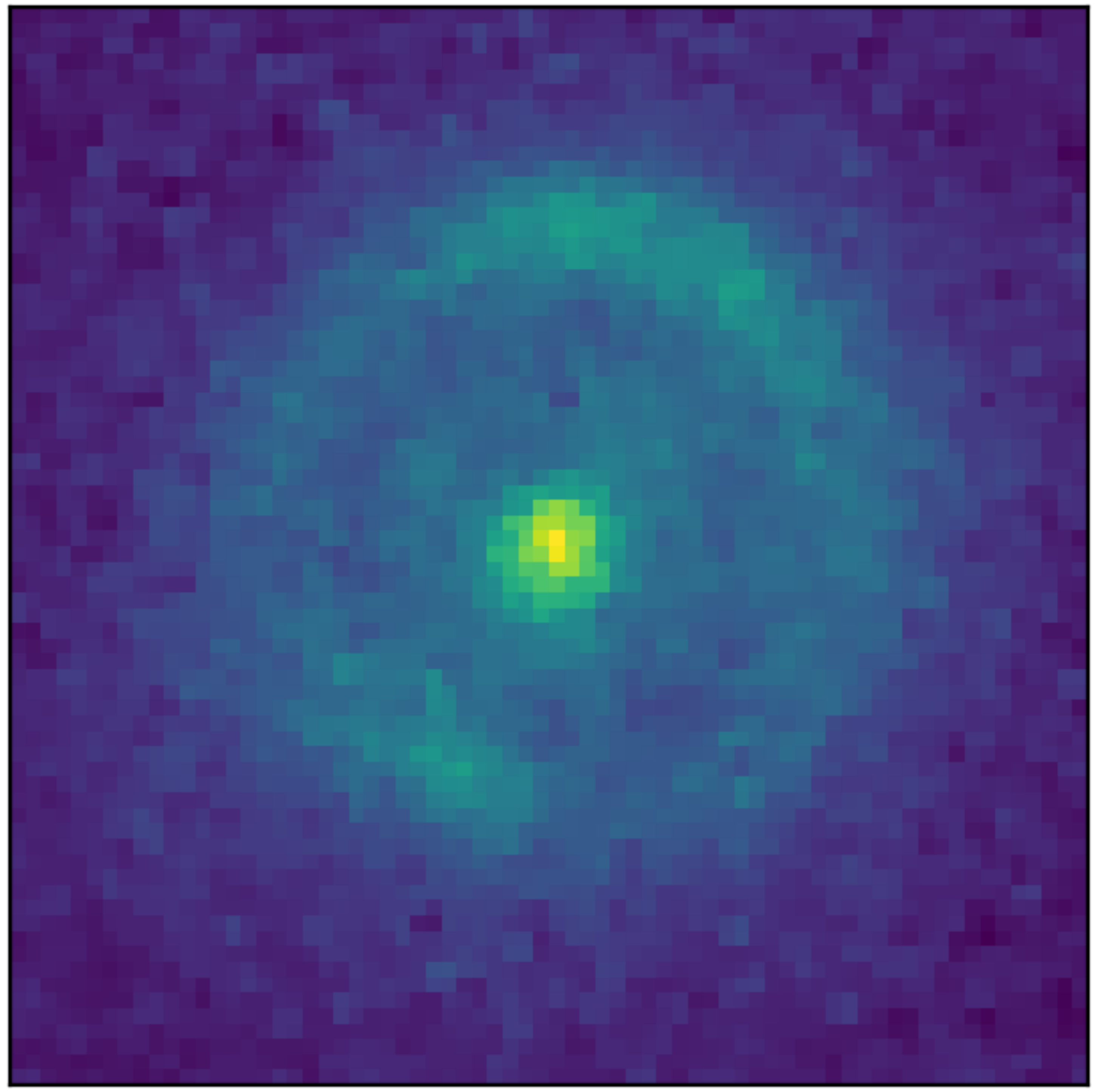
- Our results agree well with nested sampling, at 100x less simulation costs
- This enables usage of the best (but also slowest) waveform models for the analysis
- Anticipated use case: Precision analysis of complex signal situation (lensed, overlapping with glitches, multiple signals, etc)
- Our goal is *not* to upfront amortise all possible signals, but provide a framework for the fast precision analysis of individual signals

Bhardwaj+ 2304.02035

Related work: Dax+ 2106.12594

2) Strong lensing analysis of HST data

JVAS B1938+666



Summary

- First analysis that simultaneously fits source and lens light (usually lens light is subtracted)
- We do truncation using slice-sampling in 10+ dimensions to reduce training data variance
- First SBI application to real strong lensing data actually sensitive to subhalos

Anau Montel+ 23xx.yyyyy

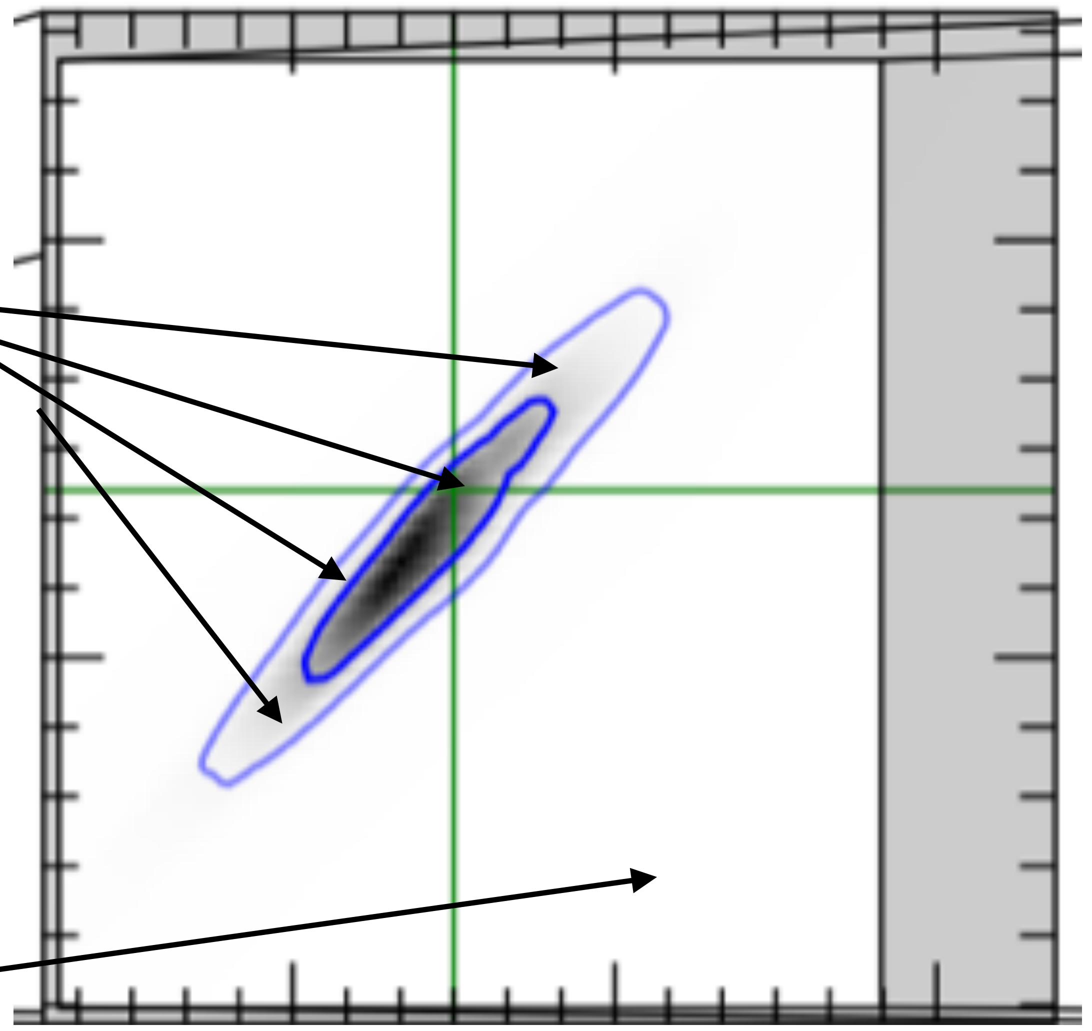
Vegetti et al. (2012) - subhalo detection claim

Şengül et al. (2021) - detection reanalysed

Truncation is connected to Nested Sampling

Neglecting correlations during truncation in general **decreases precision, not accuracy**.

Consistent with data



Everywhere else: inconsistent with data

Rectangular bounding regions are inefficient in focusing the simulation with strongly correlated parameters onto the target observation.

This does *not* lead to wrong results, but simulation efficiency can be potentially improved with complex correlated bounds.

Talk: Kilian Scheutwinkel

Connection/Synergy: Nested sampling (MultiNest, PolyChord, Dynesty, …) algorithms are good at this!

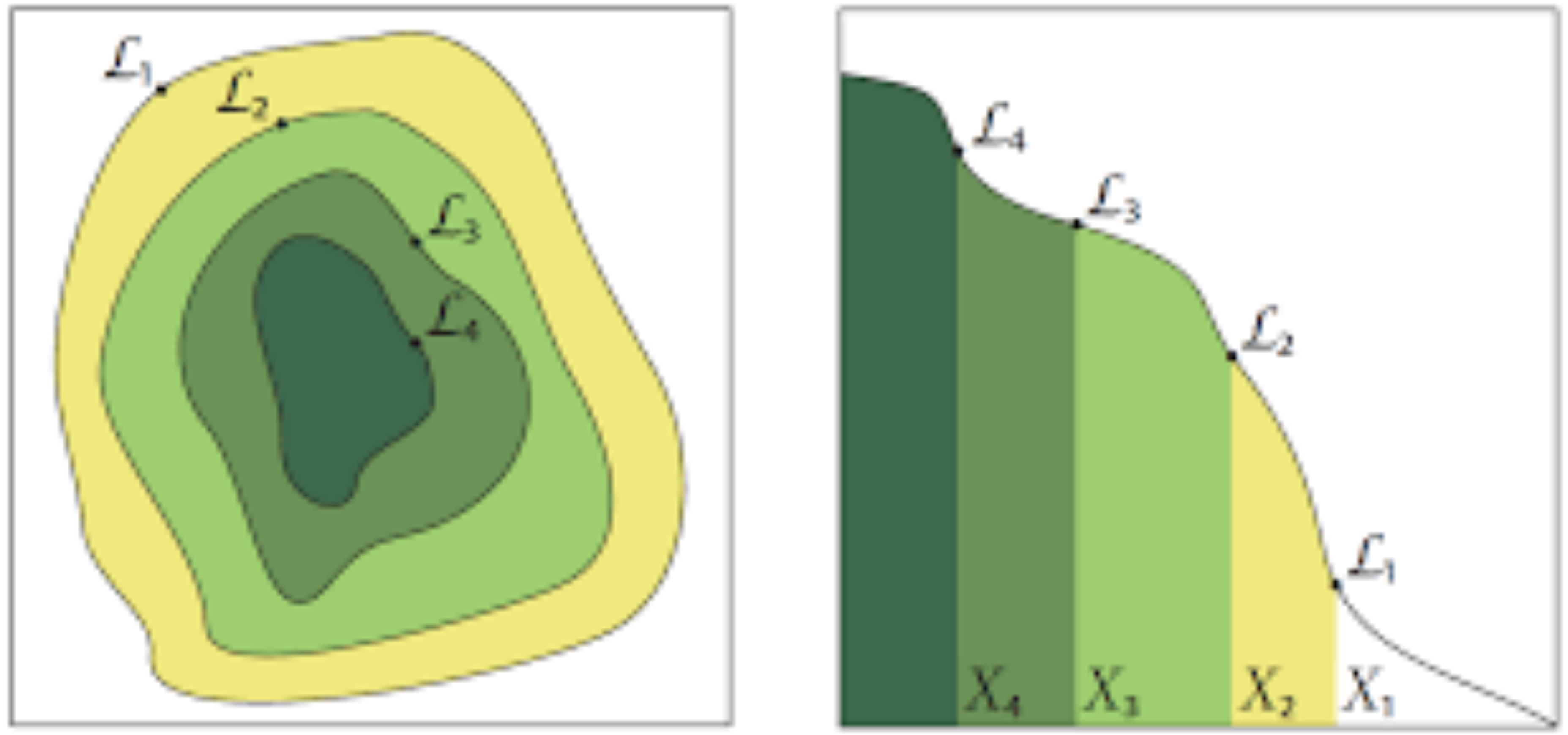


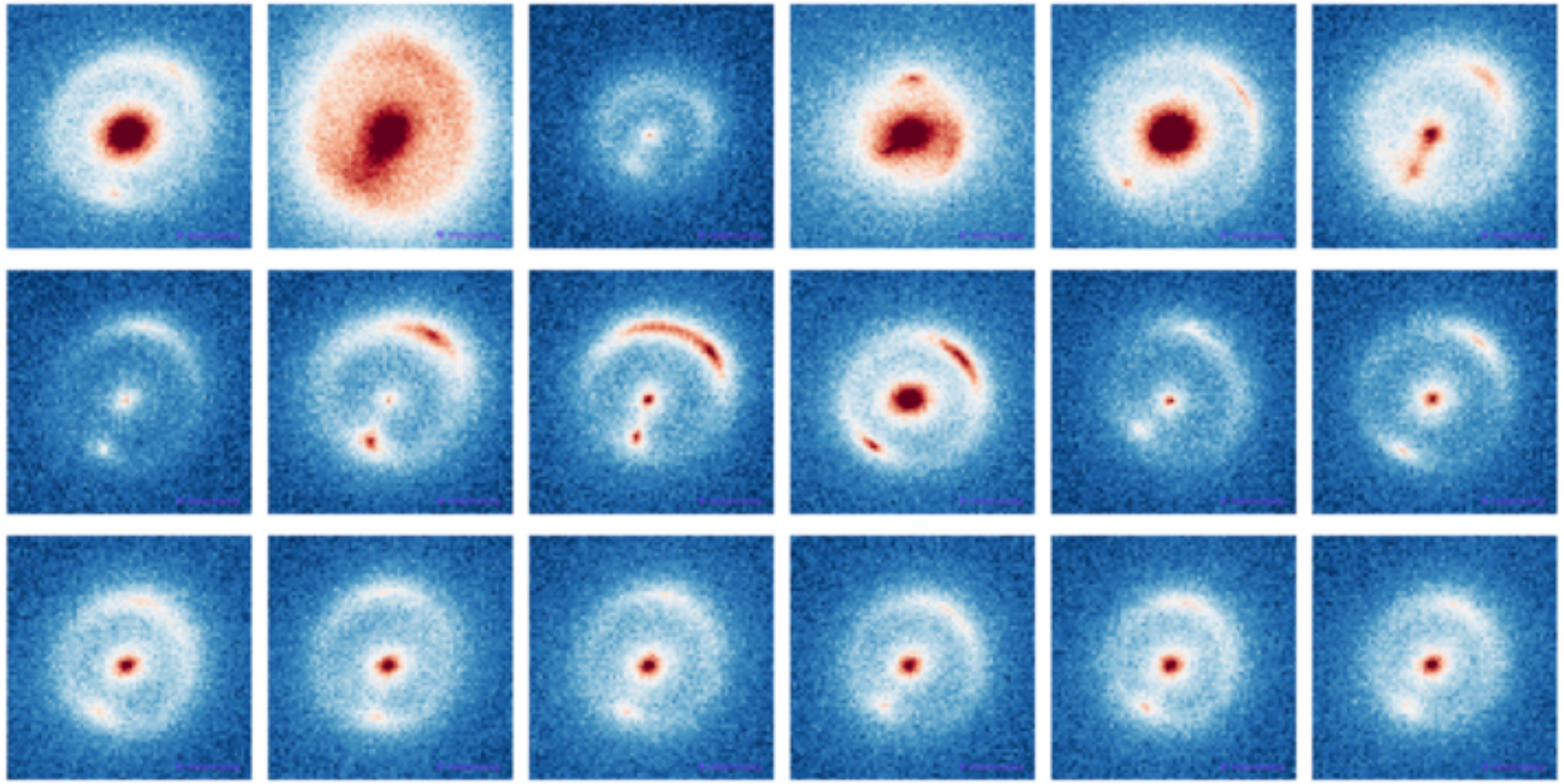
Image credit: Kosio Karchev, Will Handley

Truncation scheme: Include correlations

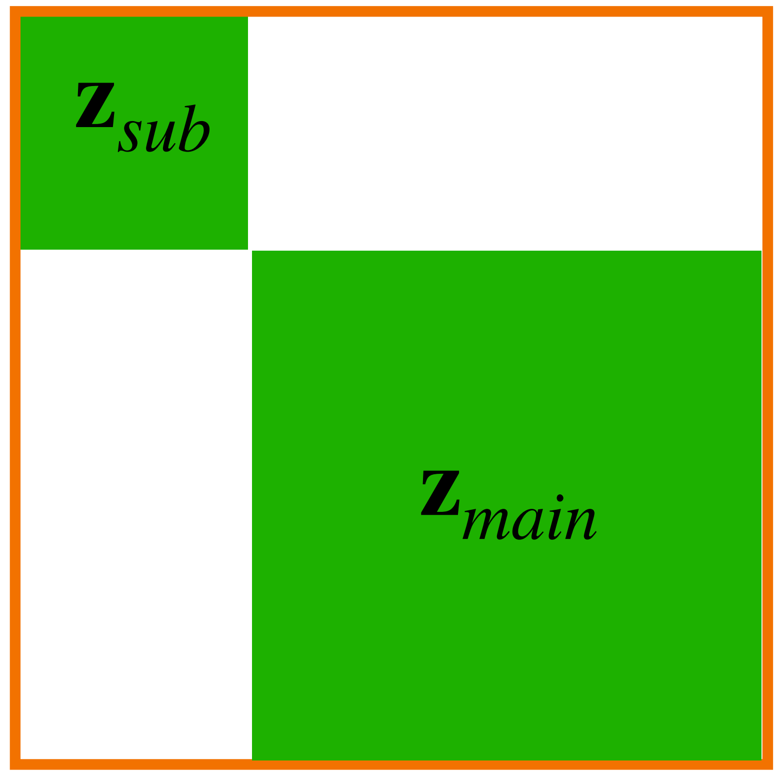
Round 1
Untruncated

Round 2: Box truncation
(Based on 14 1-dim posteriors)

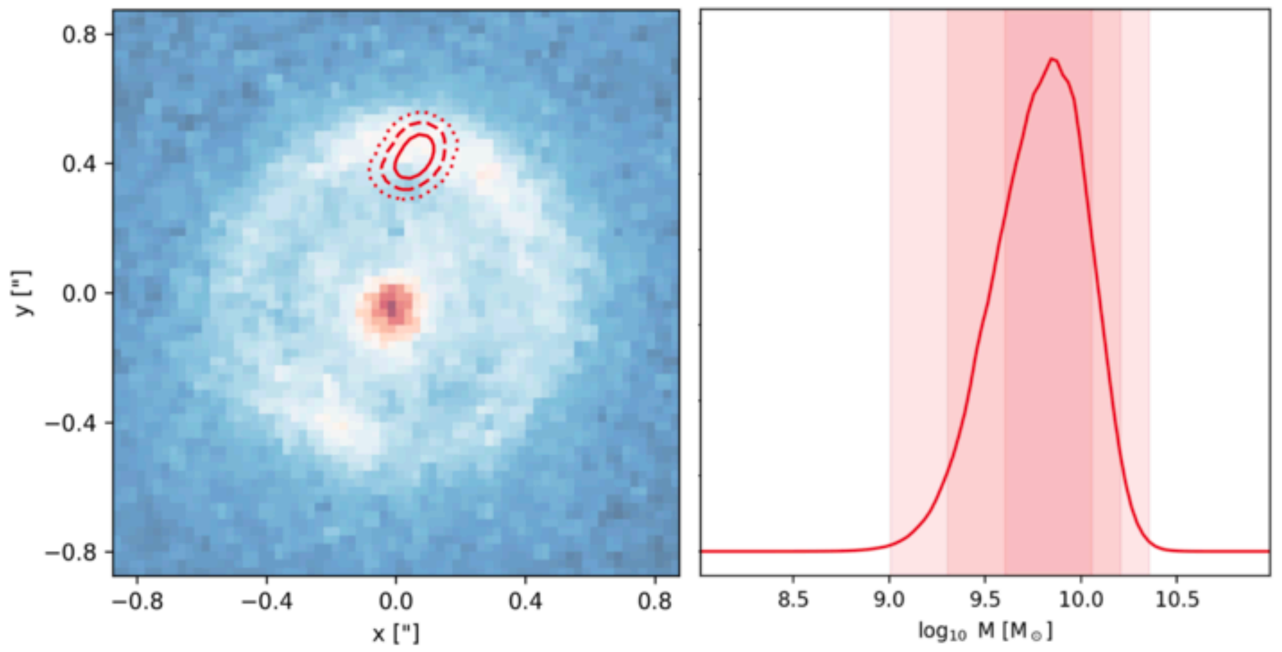
Round 3: “Nested” truncation
(14 dimensional joined posterior, explored with slice sampling)



Truncation scheme



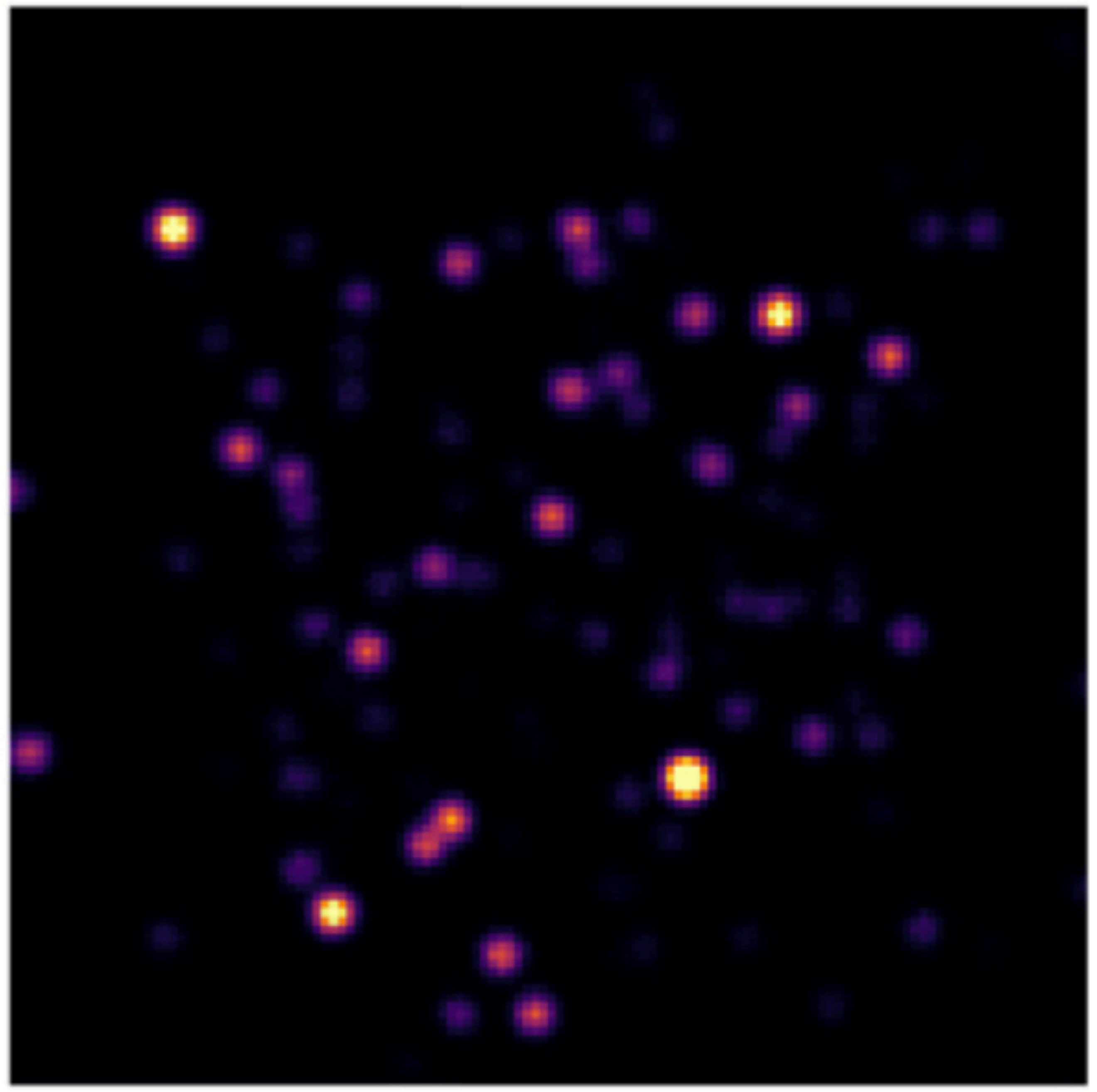
$$\mathbf{z}_{sub} = (x, y, M)$$

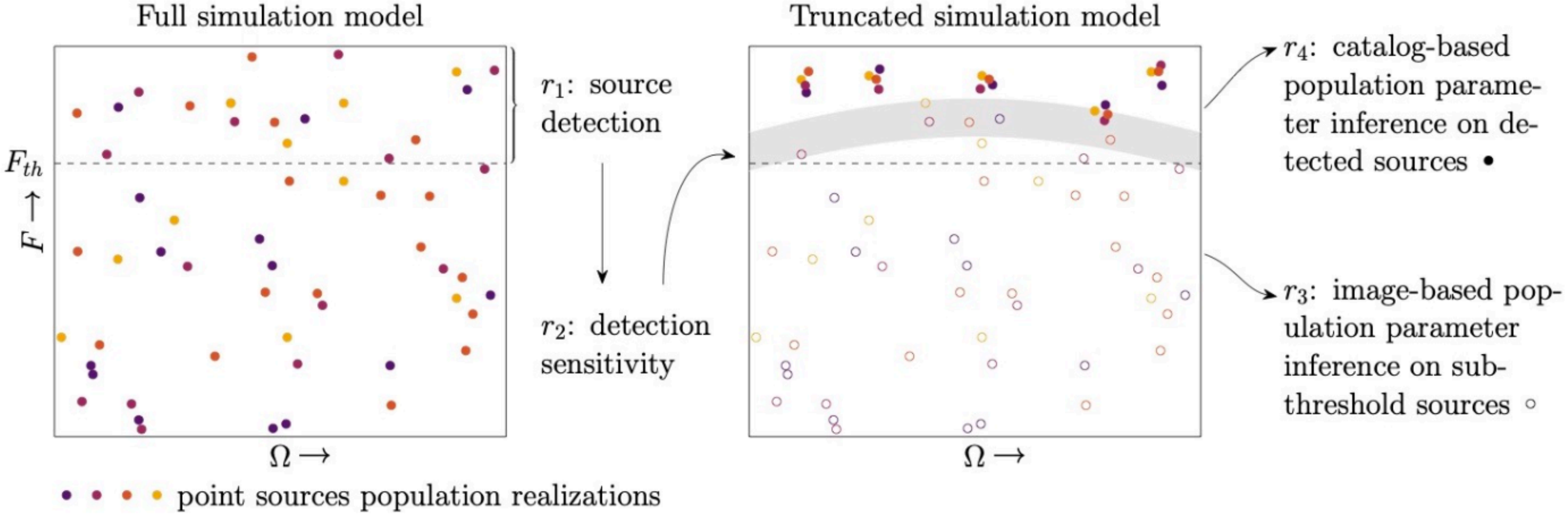


PRELIMINARY

Anau Montel+ 23xx.yyyyy

3) Source detection & population analaysis



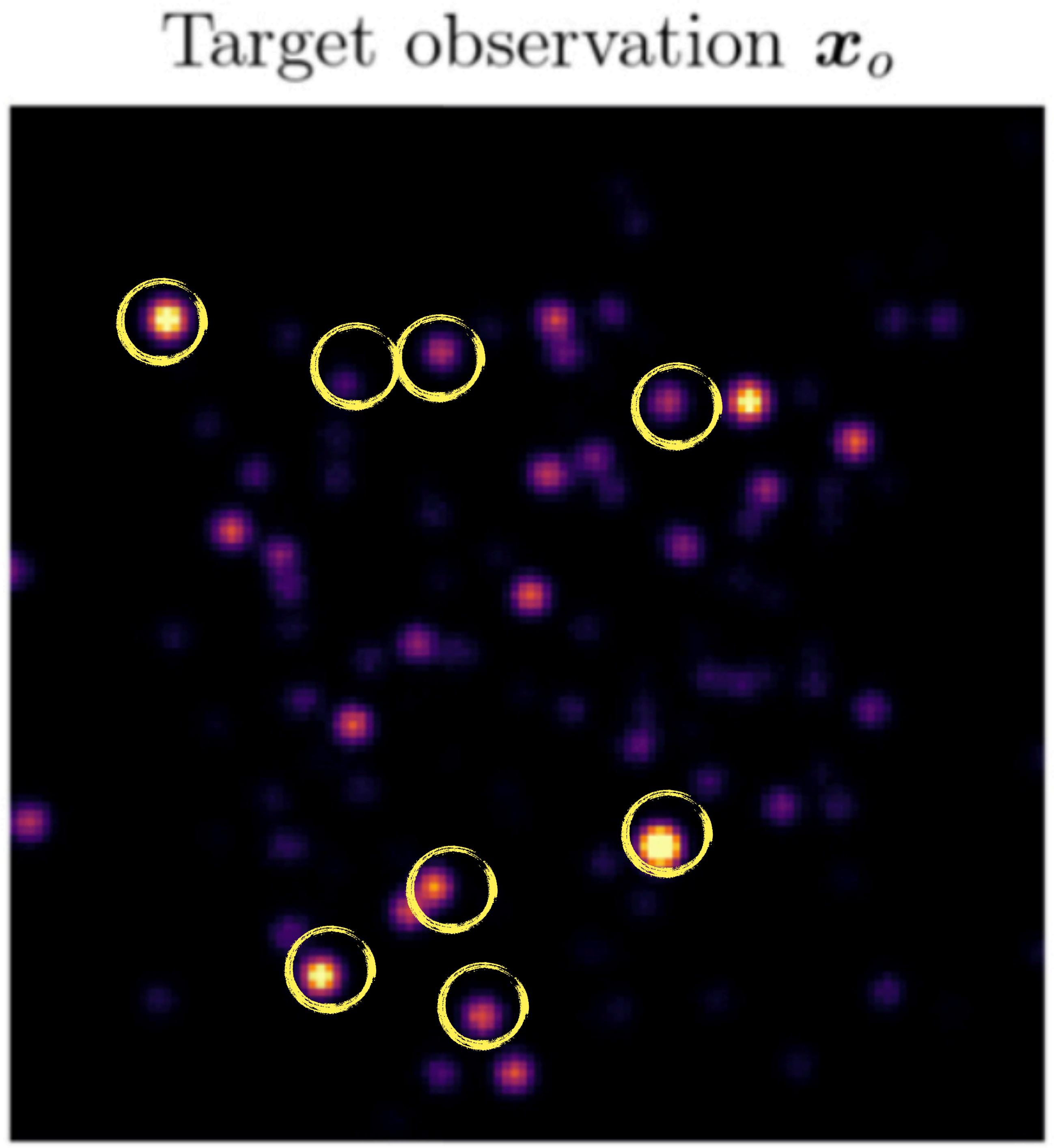


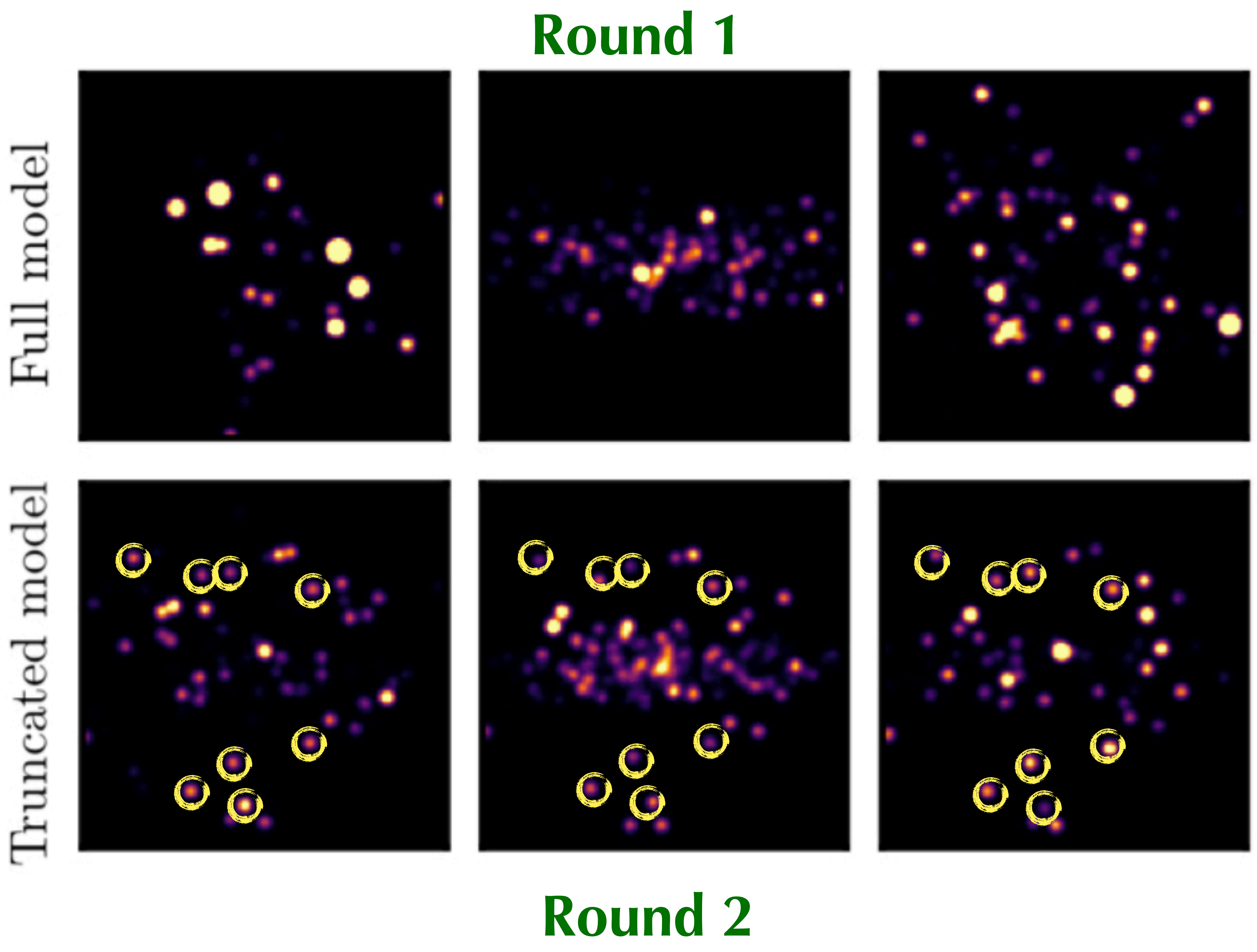
<u>Four</u> networks

- Source detection network (U-Net)
- Detection sensitivity network (MLP)
- Population parameters from detected sources (MLP)
- Population parameters from sub-threshold sources (object count network)

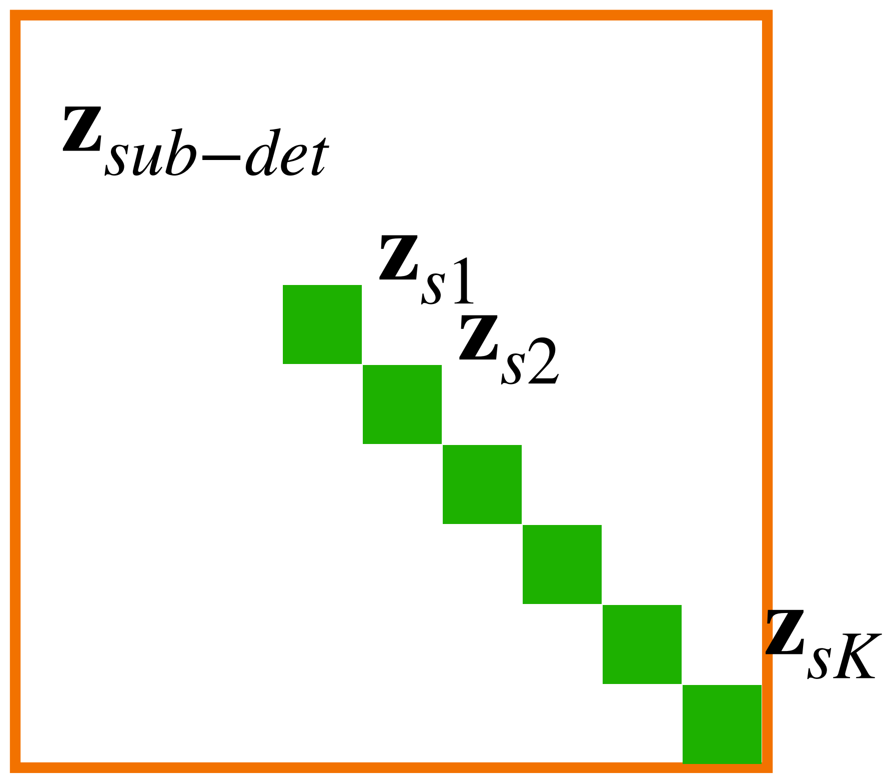
Anau Montel, CW 2211.04291

Truncation scheme: Source detection



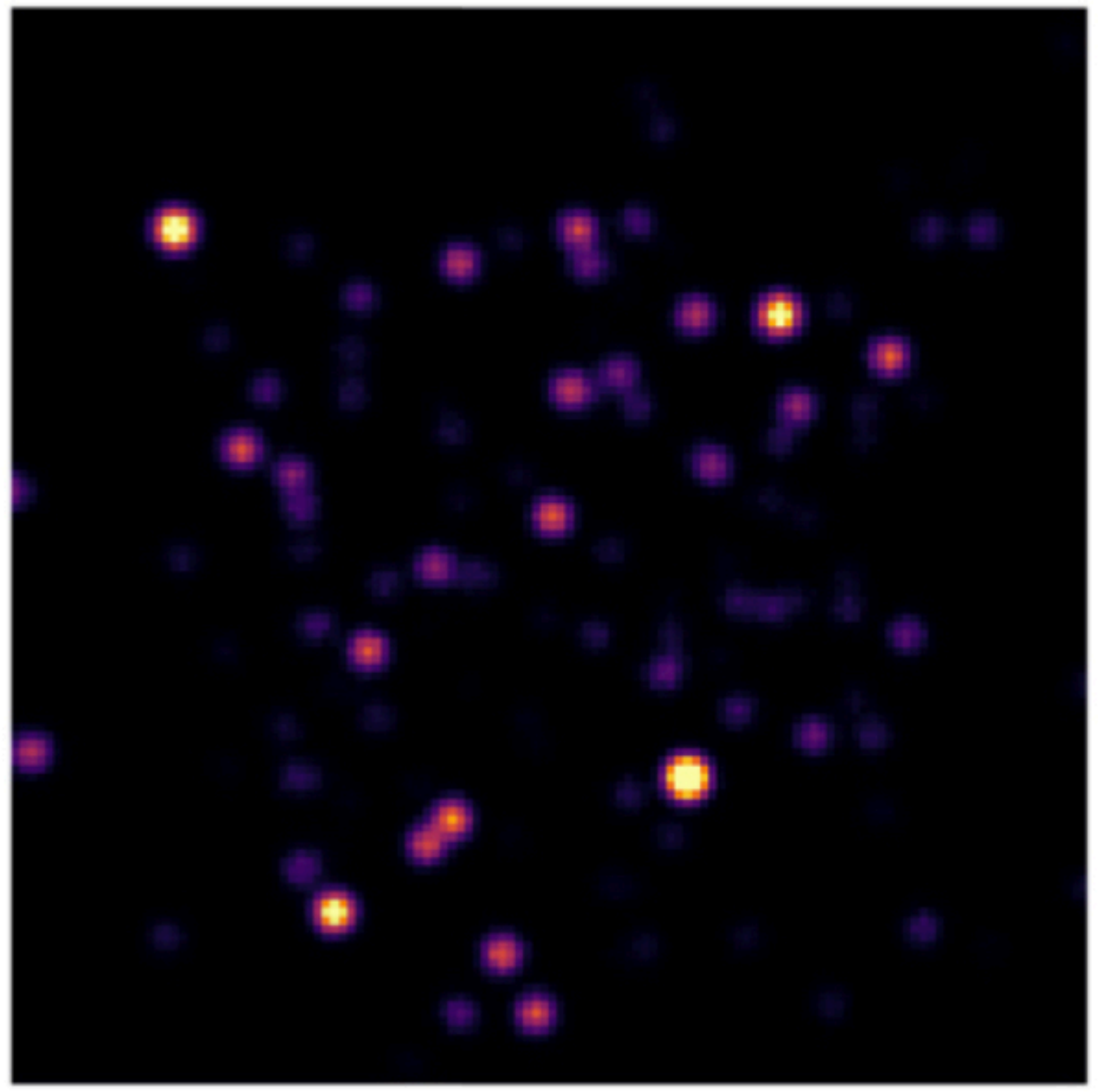


Truncation scheme

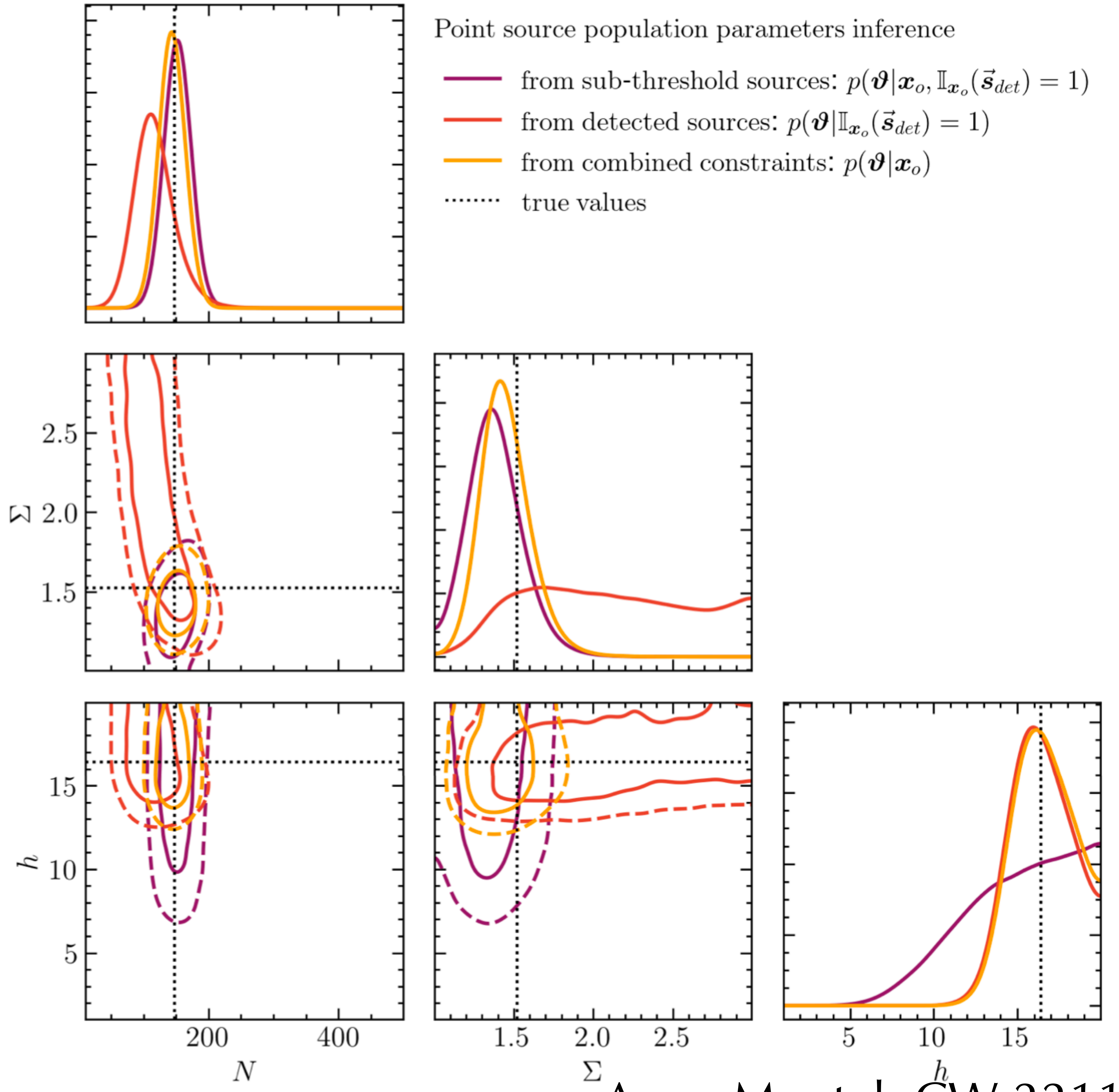


Anau Montel, CW 2211.04291

Integrating detected and sub threshold sources



TMNRE



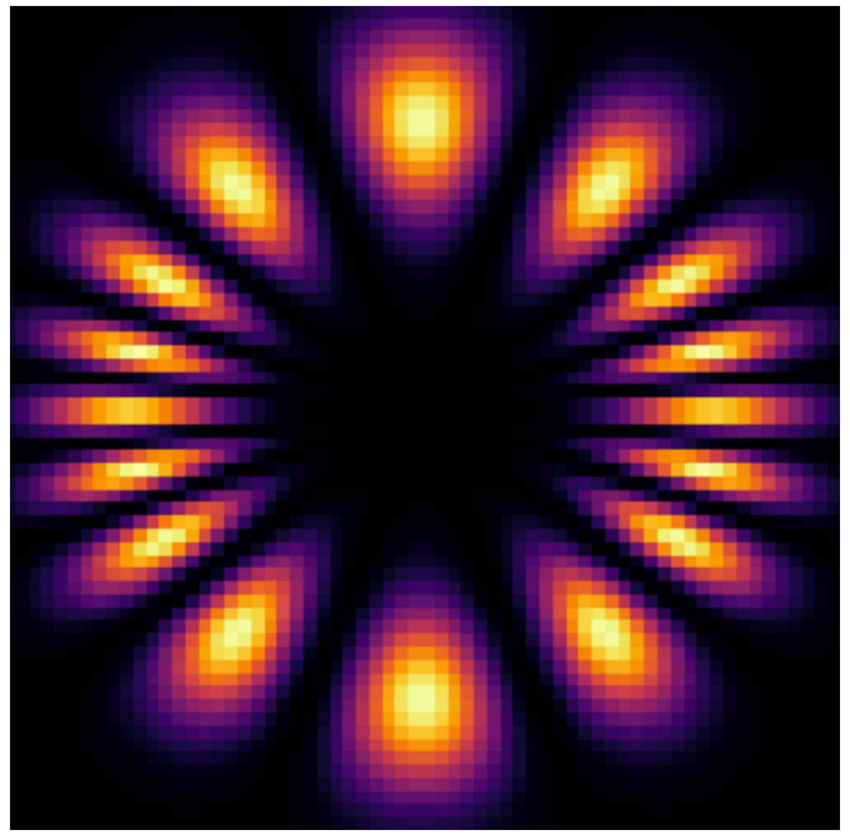
AFAIK this is the *only method* that can yield self-consistently results for

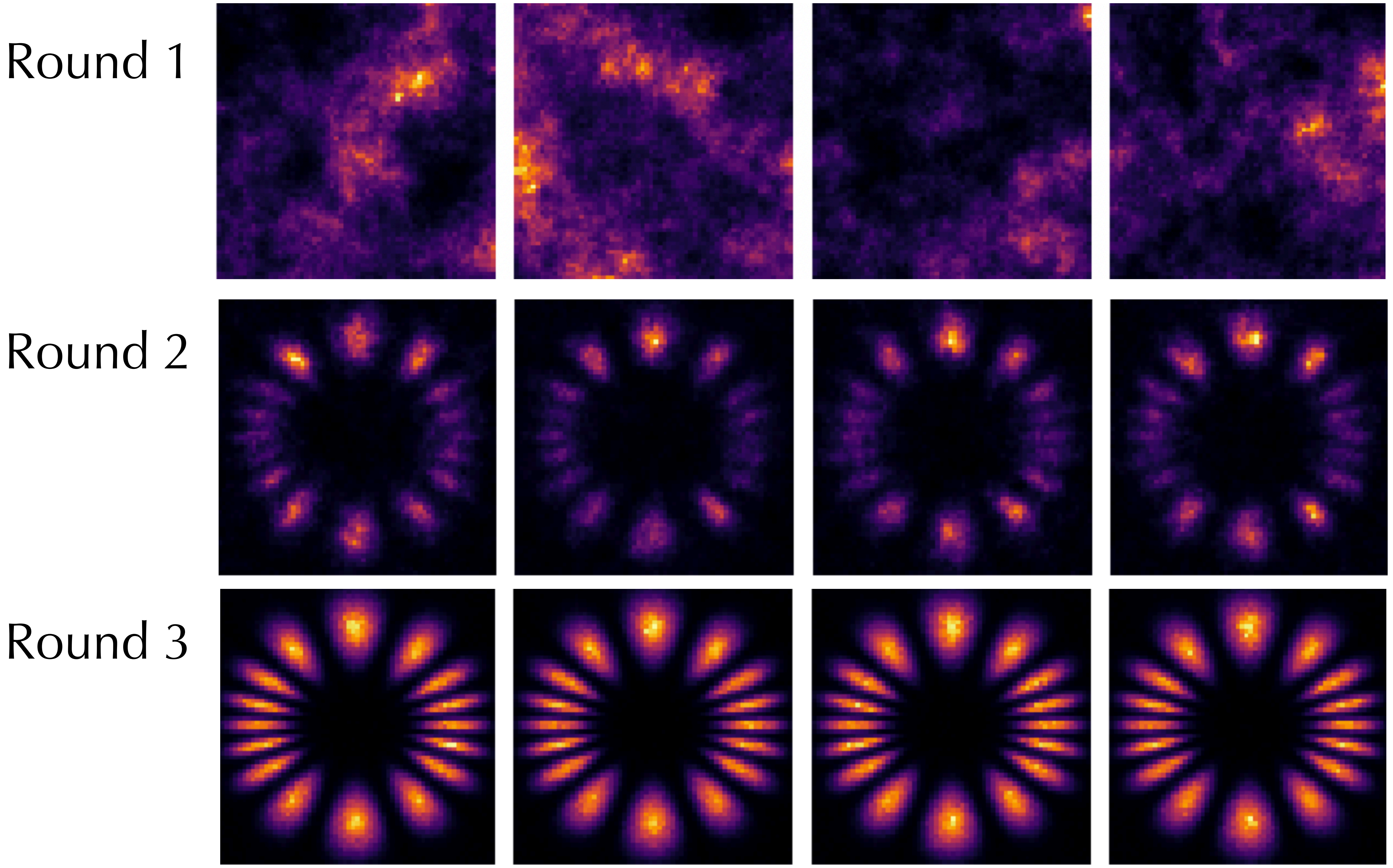
- Catalogue of detected sources
- Sensitivity estimates
- Population parameter constraints from detected sources
- Population parameter constraints from undetected sources

Anau Montel, CW 2211.04291

4) Image analysis I

Towards image analysis with SBI: Sequential inference is also possible for high-dimensional image analysis problems





- Toy model: Exponentiated Gaussian random field

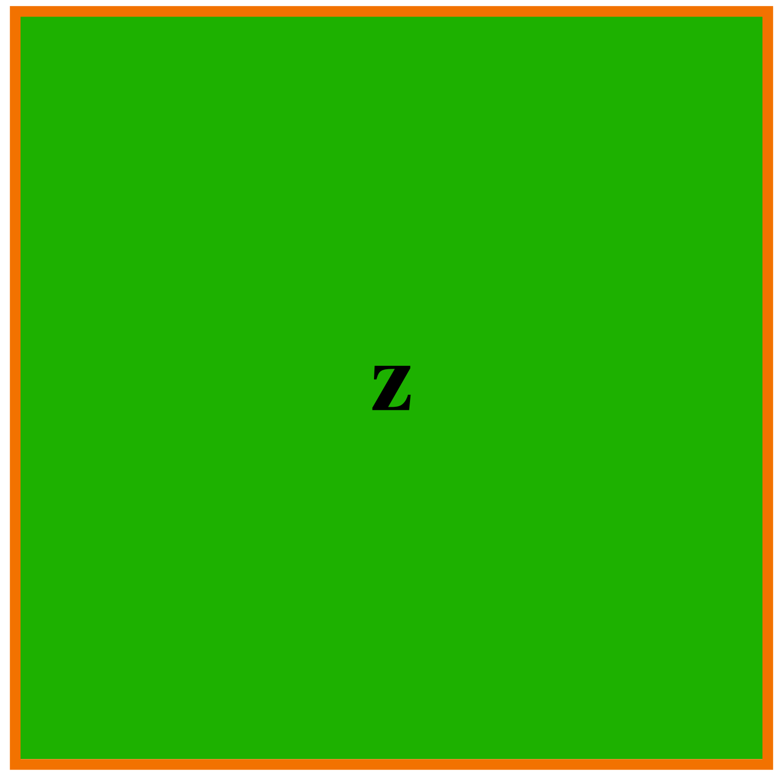
$$x_i = e^{z_i} + \epsilon \ , \quad \mathbf{z} \sim \mathcal{GP}$$

- To this end, we train the joined likelihood

$$\frac{p(\mathbf{x}\,|\,\mathbf{z})}{p(\mathbf{x})}$$

(Gaussian approx)

Truncation scheme



=> Posterior draws

Ongoing work: CW, Anau Montel, List

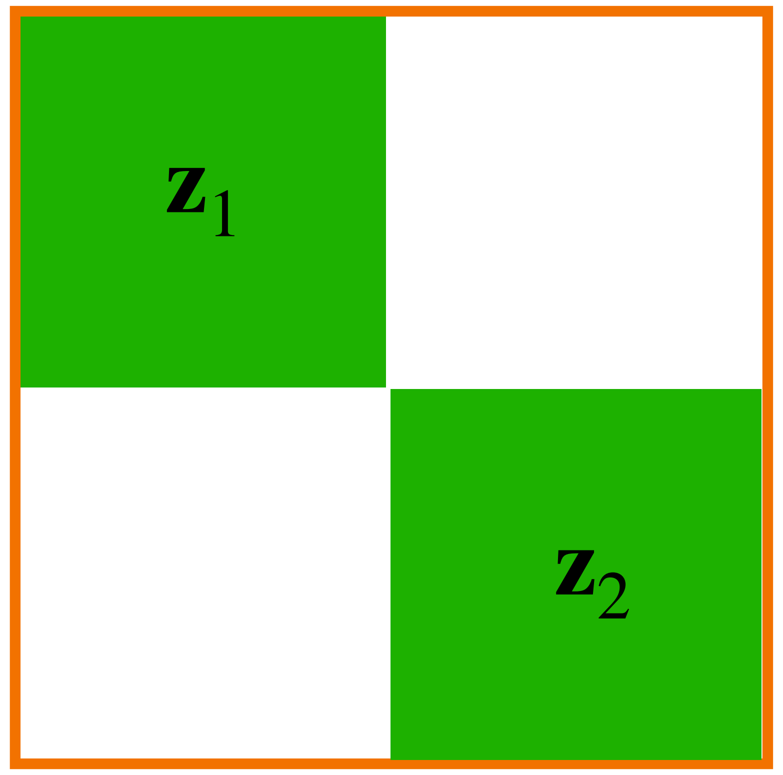
Truncation scheme: ?

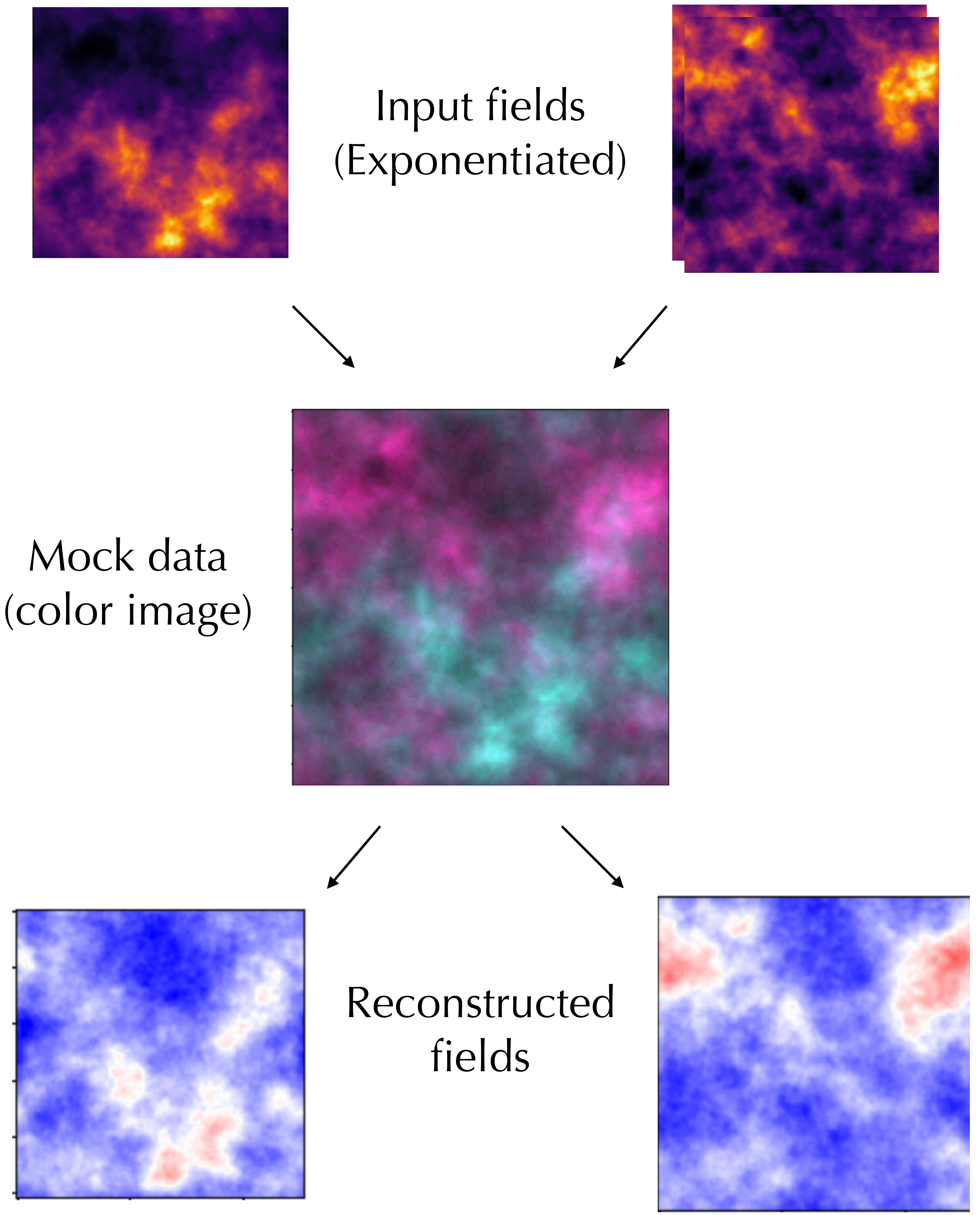
Maybe Proximal nested sampling? Cai+ 2106.03646

Example component separation
We learn the two high-dimensional likelihoods of each component, marginalised over the other components

$$\frac{p(\mathbf{x}\,|\,\mathbf{z}_1)}{p(\mathbf{x})} \qquad \frac{p(\mathbf{x}\,|\,\mathbf{z}_2)}{p(\mathbf{x})}$$

Truncation scheme

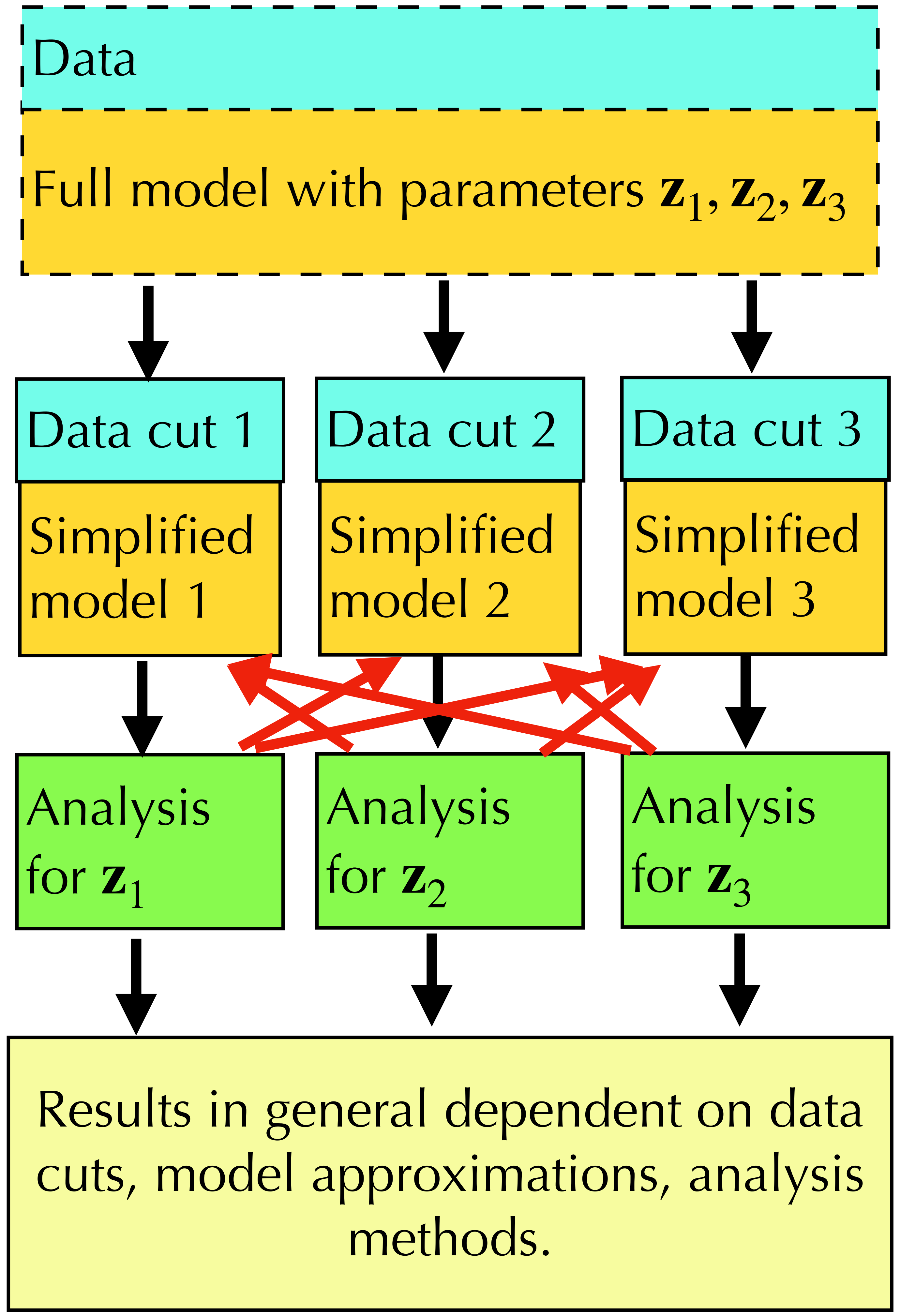


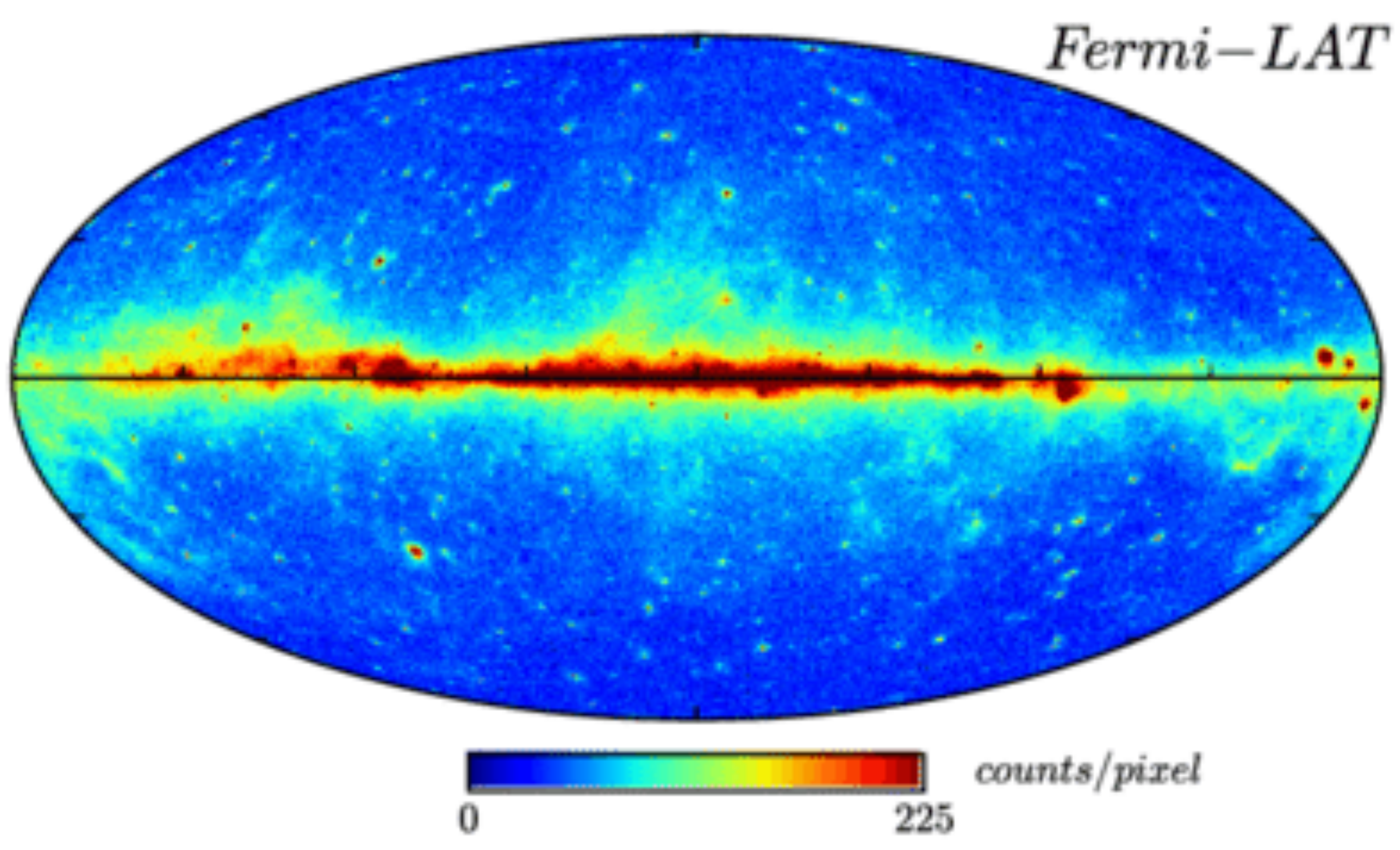


Ongoing work: CW, Anau Montel, List

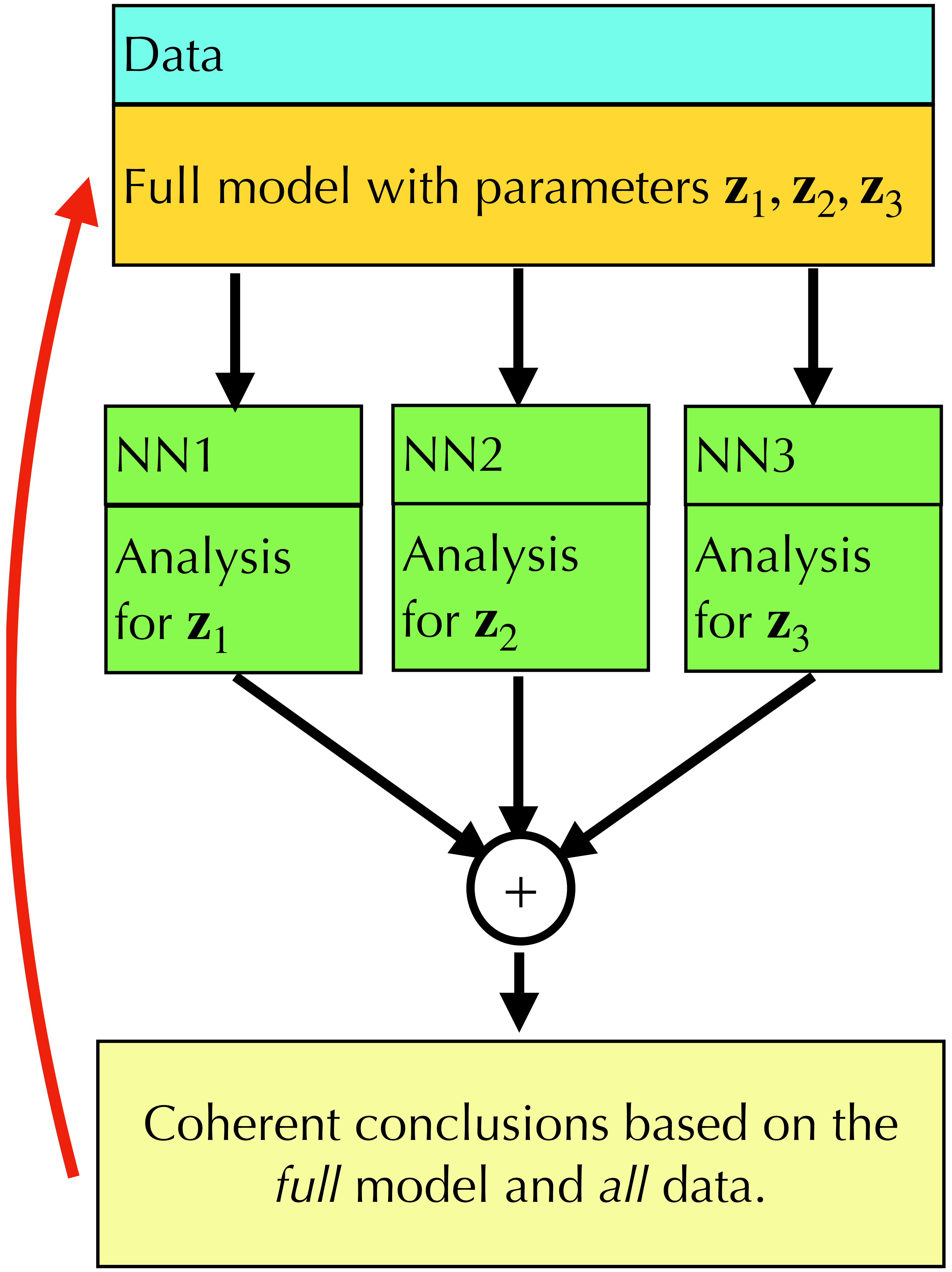
Hope for Large Forward Models?

Traditional inference (usually likelihood-based)





Opportunity offered by Simulation based inference (SBI)



“Inference Assembly”

Inference assembly: Breaking a difficult inference task into digestible components, and combining different inference results into a cohesive whole.

Purpose

1. **Model Integration**
   Merging different models or theories that make predictions about a phenomenon to form a composite model that can make more accurate or comprehensive predictions.

2. **Method Integration**
   Combining different inference strategies (UNets, graph neural networks, CNNs, etc) to coherently perform distinct analysis tasks on the same data.

3. **Data Synthesis**
   Compiling and integrating data from different sources or experiments to increase the robustness and reliability of inferences drawn.

4. **Uncertainty Quantification**
   Assessing the confidence of different inferences and integrating these assessments into the final model or prediction.

Swyft

https://github.com/undark-lab/swyft

- **Swyft**: An opinionated system for scientific simulation-based inference at scale
- **Mission:** Enabling Inference Assembly
- Key components:
  - **Marginal inference**: Performing marginal inference for many parameters parallel
  - **Basic SBI algorithms**: Binary classification/NRE - simple but sufficient for our use-cases
  - **General applicability**: Works for complex hierarchical models with a large number of components

See also: https://github.com/mackelab/sbi & LAMPE

Likelihood-free AMortized Posterior Estimation

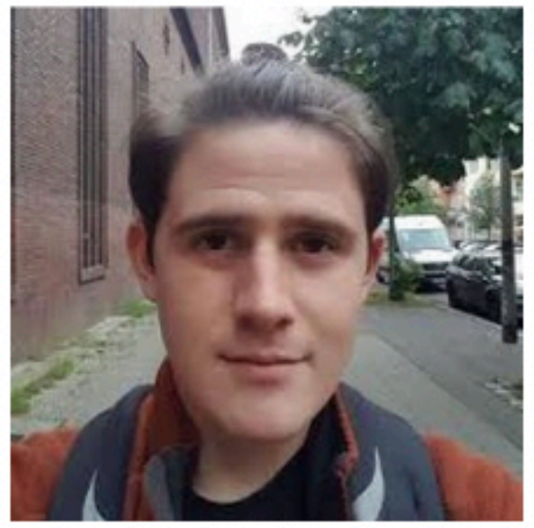
Research group with high pain threshold

Noemi Anau Montel (UvA), Strong lensing

Adam Coogan (Mila, U. Montreal), Strong lensing

Alex Cole (UvA), CMB, 21cm

Elias Dubbeldam (UvA), Strong lensing



Ben Miller (UvA), Algorithm & software

Kosio Karchev (SISSA) SN cosmology, strong lensing

Mathis Gerdes (UvA), Stellar streams, QFT

Androniki Dimitriou (Valencia), Large scale structures

Uddipta Bhardwaj (UvA) Gravitational waves

James Alvey (UvA), Stellar streams, GWs

\+
Gilles Louppe (U. Liège)
Anchal Saxena (Groningen)
Patrick Forré (UvA)
Samaya Nissanke (UvA)
Maxwell Cai, Meiert Grootes, Francesco Nattino (eScience)

Thanks!

(A bit outdated, sorry)

Summary

- Simulation-Based Inference (SBI) is likely becoming a new standard for data analysis in our field, because it enables studying realistic complex models at low computational costs.
- TMNRE has the right set of ingredients to deal with potentially very high-dimensional models.
- Fully exploiting the potential of TMNRE requires new computational infrastructure, and poses interesting software design questions. With Swyft we provide our attempt solution.
- There are tons of potential applications. We are open for discussions and collaborations on all fronts.

Thank you!

Backup

Graphical model for Fermi data (my attempt)

$$\mathbf{z}_{CRsrc} \sim p(\mathbf{z}_{CRsrc})$$

$$\mathbf{z}_{prop} \sim p(\mathbf{z}_{prop})$$

$$\mathbf{z}_{B} \sim p(\mathbf{z}_{B})$$

$$\mathbf{z}_{ISRF} \sim p(\mathbf{z}_{ISRF})$$

$$\mathbf{z}_{CR} \sim p(\mathbf{z}_{CR} \,|\, \mathbf{z}_{CRsrc}, \mathbf{z}_{prop}, \mathbf{z}_{B}, \mathbf{z}_{ISRF})$$

$$\mathbf{z}_{gas} \sim p(\mathbf{z}_{gas})$$

$$\mathbf{z}_{ICS} \sim p(\mathbf{z}_{ICS} \,|\, \mathbf{z}_{CR}, \mathbf{z}_{B}, \mathbf{z}_{ISRF})$$

$$\mathbf{z}_{pi0} \sim p(\mathbf{z}_{pi0} \,|\, \mathbf{z}_{gas}, \mathbf{z}_{CR})$$

$$\mathbf{z}_{pop} \sim p(\mathbf{z}_{pop})$$

$$\mathbf{z}_{psc} \sim p(\mathbf{z}_{psc} \,|\, \mathbf{z}_{pop})$$

$$\mathbf{z}_{tot} \sim p(\mathbf{z}_{pi0}, \mathbf{z}_{ICS}, \mathbf{z}_{psc}, \mathbf{z}_{bub}, \mathbf{z}_{dm})$$

$$\mathbf{z}_{dm} \sim p(\mathbf{z}_{dm})$$

$$\mathbf{x} \sim p(\mathbf{x} \,|\, \mathbf{z}_{tot})$$

$$\mathbf{z}_{bub} \sim p(\mathbf{z}_{bub})$$

Can we miss modes? - It depends...

Let's consider two simple examples **which have similar likelihood functions around the observation**.

Model A: Tight linear relation

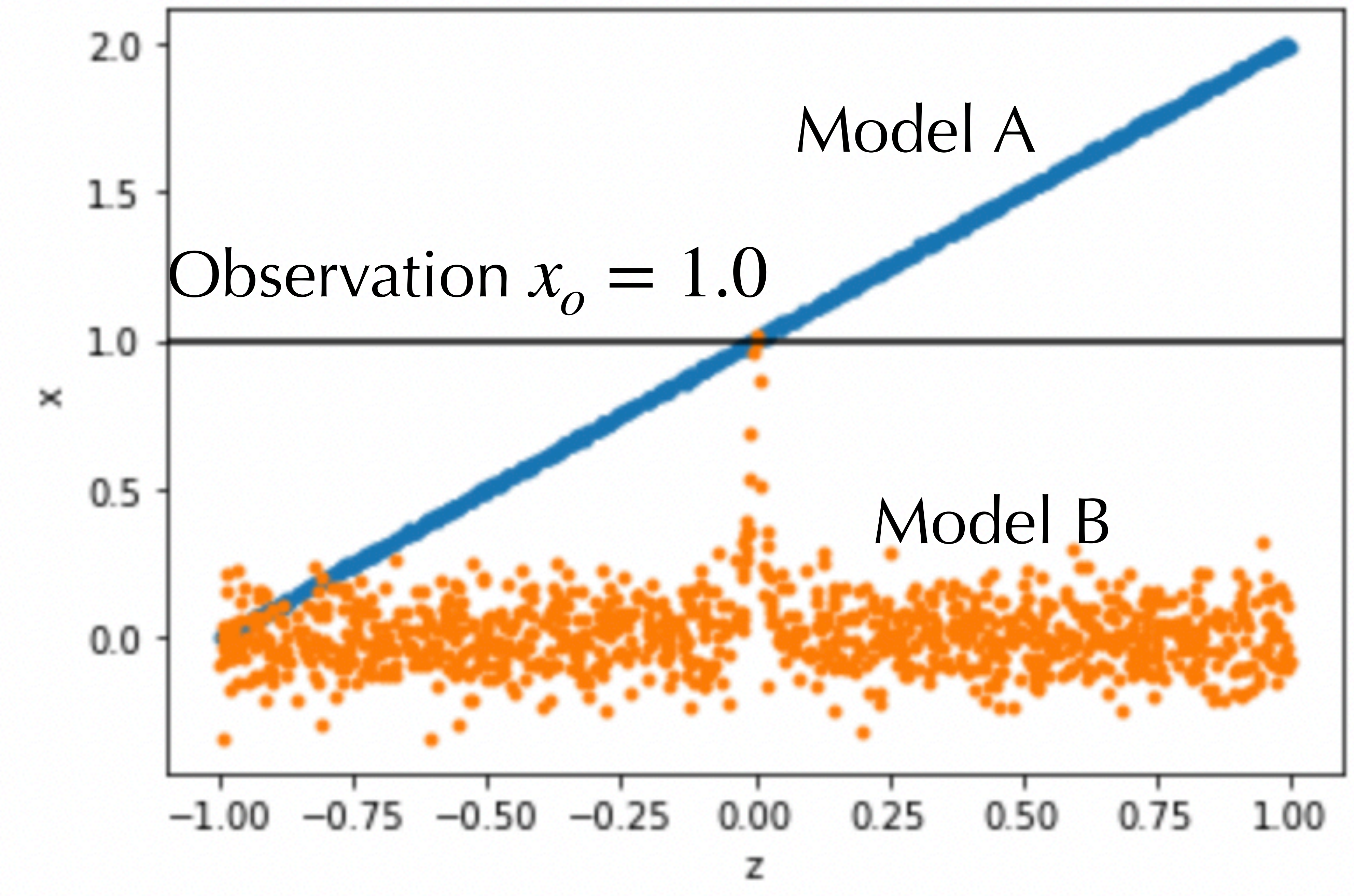
$z \sim \mathcal{U}(-1,1) \qquad \epsilon \sim \mathcal{N}(0,0.005)$

$$x = z + 1 + \epsilon$$

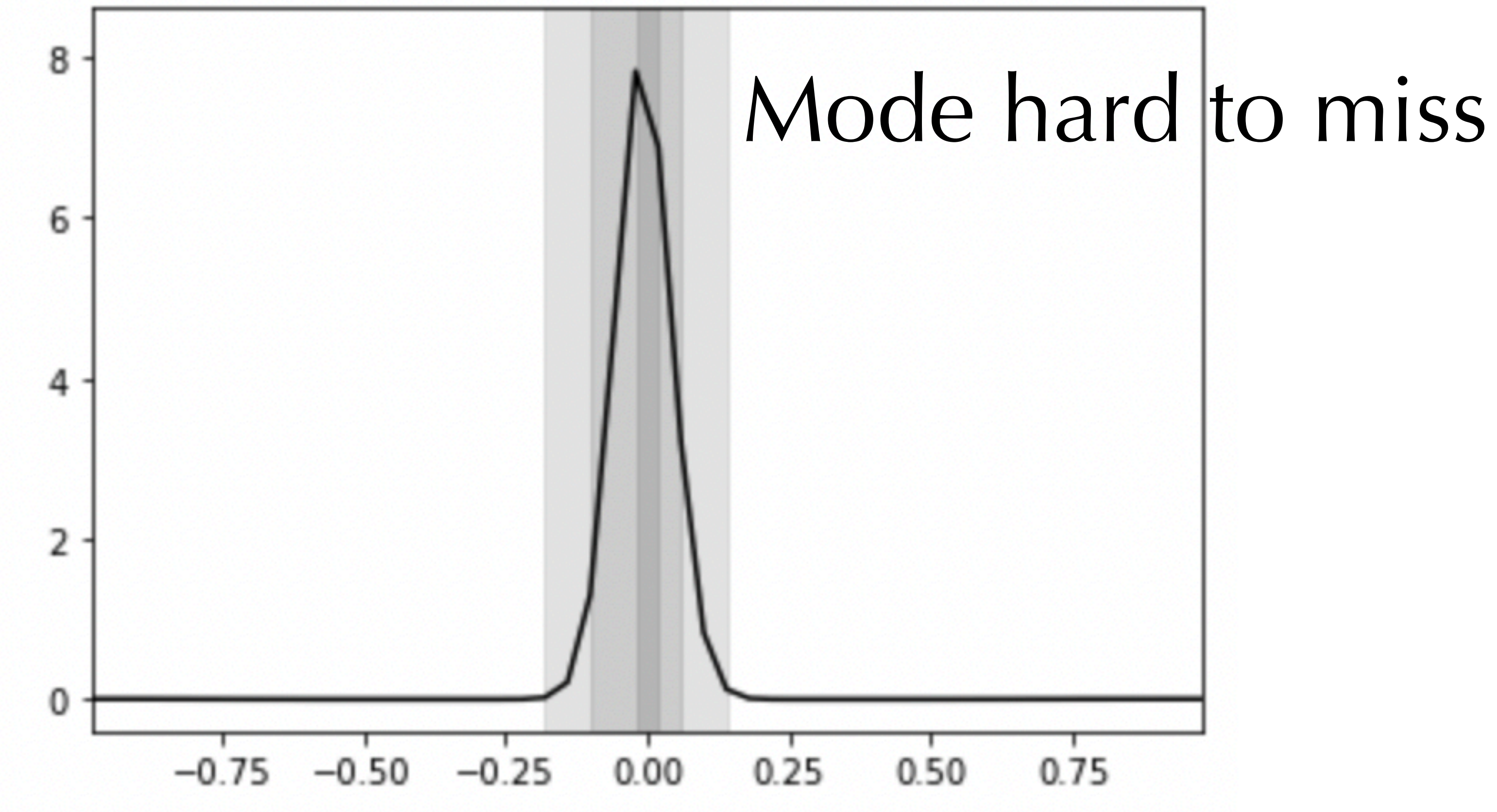
Model B: "Resonance"

$z \sim \mathcal{U}(-1,1) \qquad \epsilon \sim \mathcal{N}(0,0.1)$

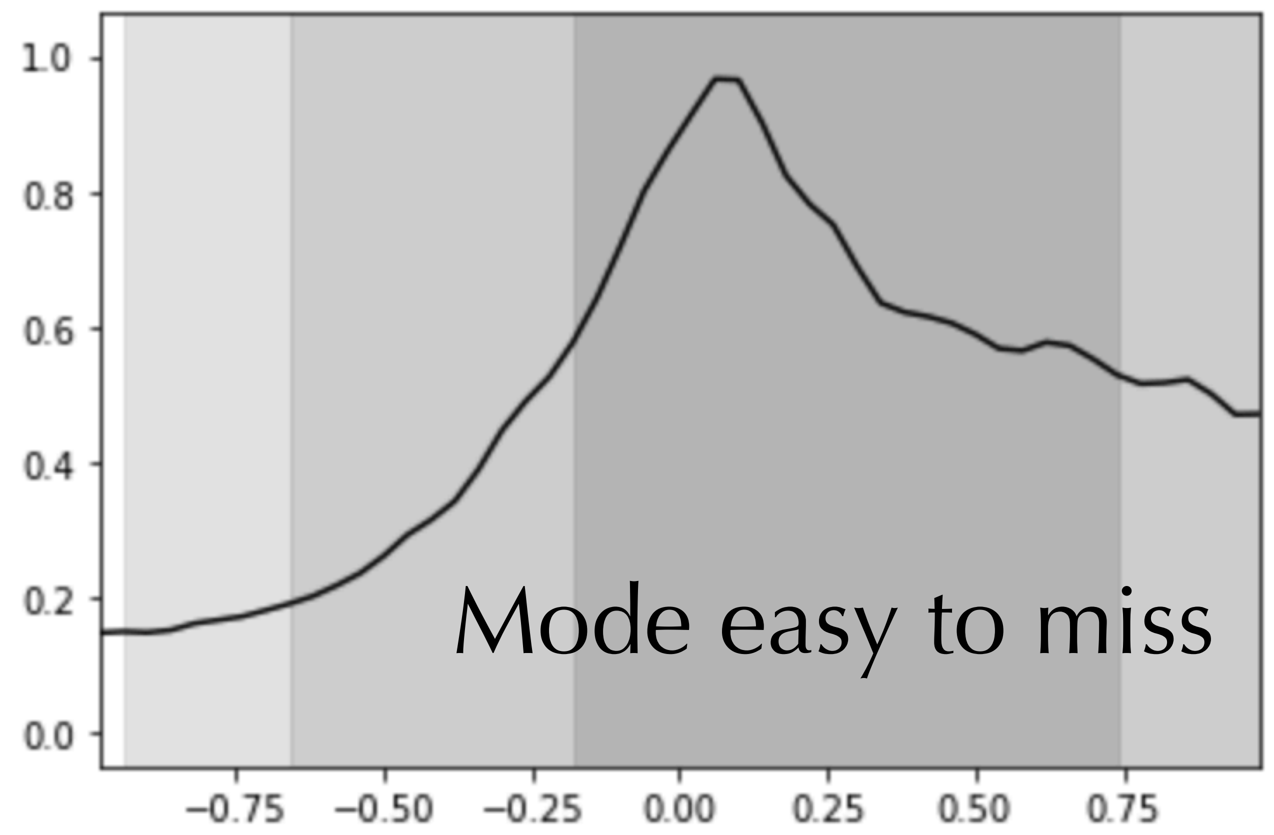
$$x = \frac{1}{1 + \left(\frac{z}{0.01}\right)^2} + \epsilon$$



Posterior for model A



Posterior for model B

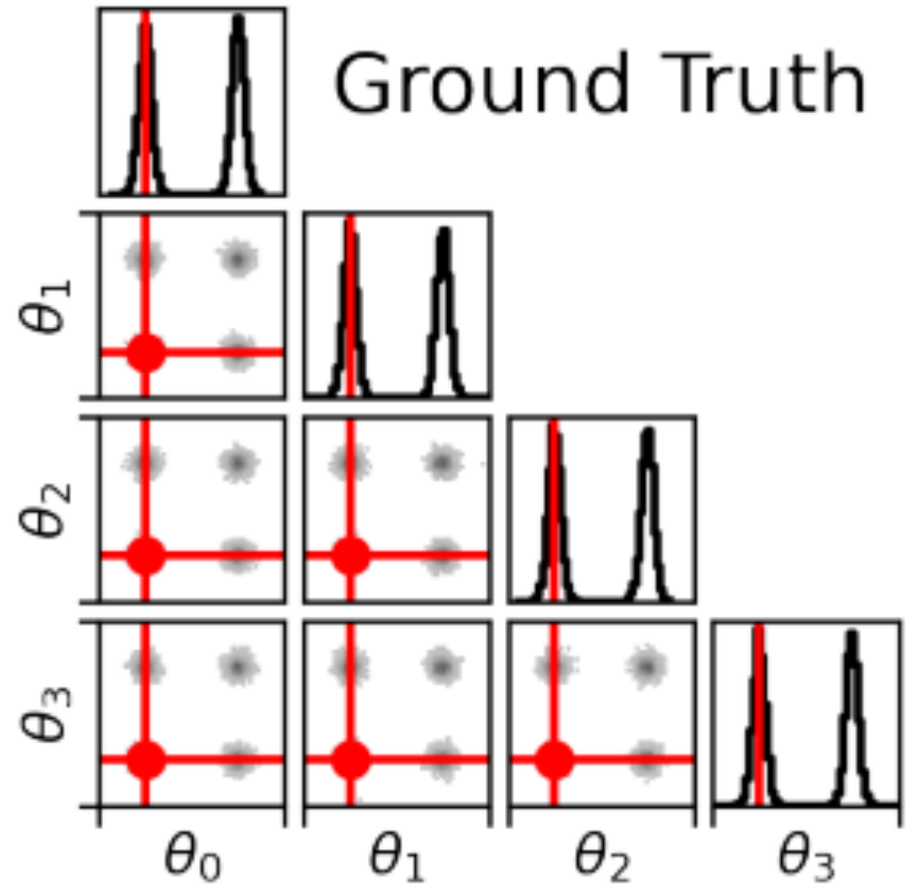


(These are quick&dirty results for illustration)

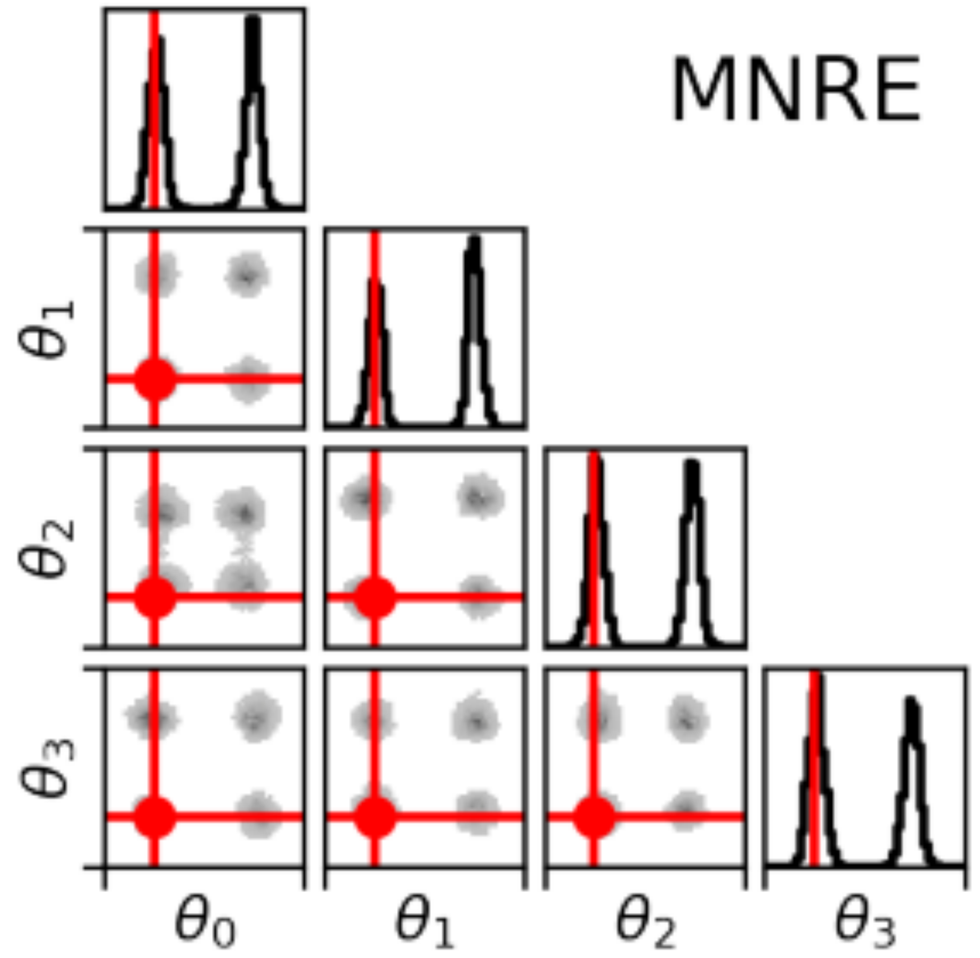
Marginal inference with SBI

Miller+ 2011.13951

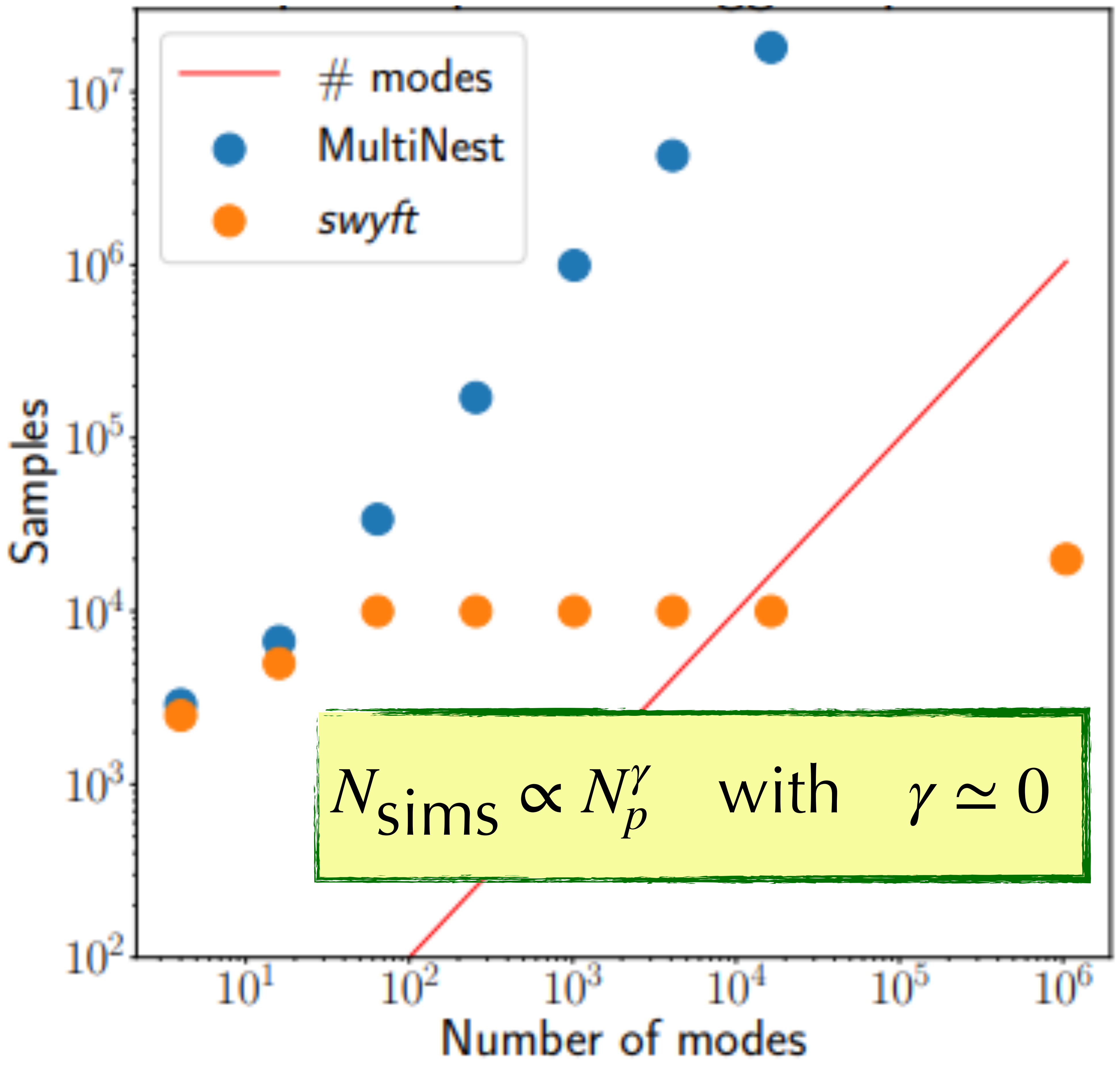
Correct:



MNRE* result



*We trained 190 2-dim posterior estimators for this. That is much better than training a single 20-dim posterior and then marginalising.



$$N_{\text{sims}} \propto N_p^\gamma \quad \text{with} \quad \gamma \simeq 0$$

$N_p = 1$ $N_p = 10$ $N_p = 20$

Estimating marginals is trivial with SBI

Training a neural network $f(\mathbf{x}; \mathbf{z})$ with $\mathbf{x}, \mathbf{z} \sim p(\mathbf{x}, \mathbf{z})$

—> estimates the joined posterior, $p(\mathbf{z} \mid \mathbf{x})/p(\mathbf{z})$.

Training a neural network $f(u; \mathbf{z})$ with $\mathbf{x}, u(\mathbf{z}) \sim p(\mathbf{x}, \mathbf{z})$

—> estimates the marginal posterior, $p(u \mid \mathbf{x})/p(u)$.

In a likelihood-based setting we would have to perform the high-dim integral:

$$p(u \mid \mathbf{x}) = \int d\mathbf{z}\, \delta(u - u(\mathbf{z})) p(\mathbf{z} \mid \mathbf{x})$$