GRAPPA

GRavitation AstroParticle Physics Amsterdam

UNIVERSITY OF AMSTERDAM

Enabling precision dark matter searches with Truncated Marginal Neural Ratio Estimation

Christoph Weniger

James Alvey (UvA), Uddipta Bhardwaj (UvA), Alex Cole (UvA), Adam Coogan (U. Montreal), Androniki Dimitriou (U. Valencia), Elias Dubbeldam (UvA), Mathis Gerdes (UvA), Kosio Karchev (SISSA), Ben Miller (UvA), Noemi Anau Montel (UvA), Roberto Trotta (SISSA)
Gilles Louppe (U. Liège), Anchal Saxena (Groningen), Patrick Forré (UvA), Samaya Nissanke (UvA), Maxwell Cai, Meiert Grootes, Francesco Nattino (eScience)

European Research Council

Established by the European Commission

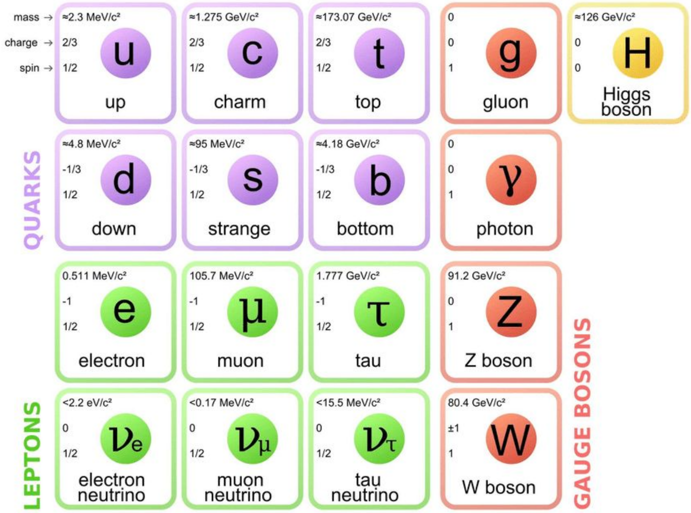
netherlands eScience center

AI goes MAD

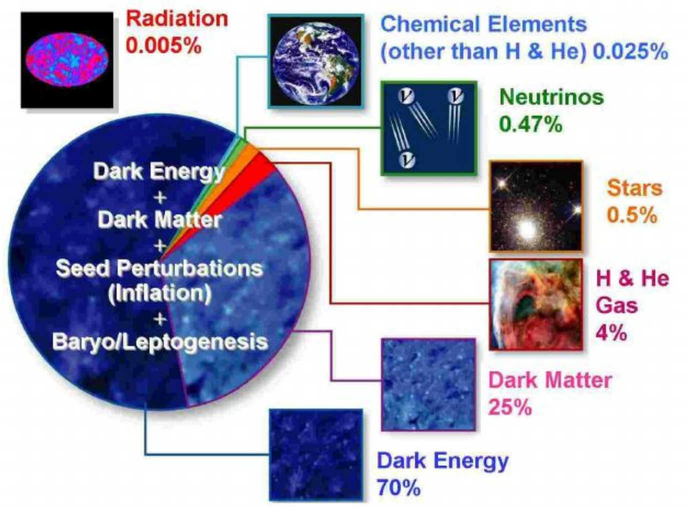
14 June 2022

Standard models and open questions

Particle physics



Cosmology



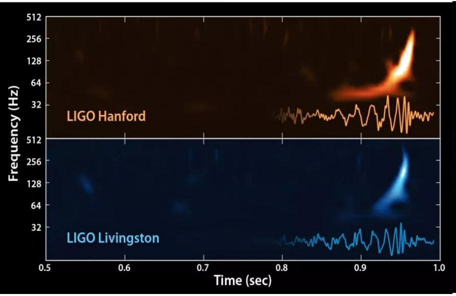
Many option questions: What is dark matter? Is dark energy dynamic? Where is the anti-matter? What caused seed-perturbations? How do black holes grow and merge? How do neutron stars develop? How did the first stars form? What stabilized the electroweak scale? Is there grand unification? How do galaxies form and grow?

Image credits: E. Kolb; Gianluca Bianco

Astrophysical searches for breaks in the standard models

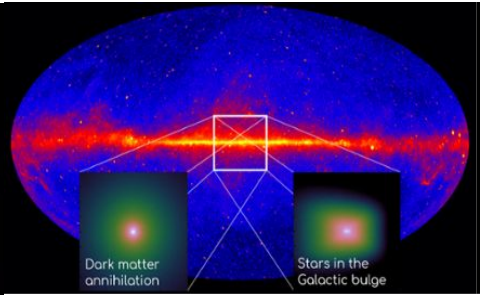
Strategy: Search for deviations from standard-model predictions in order to get answers on some of the open questions.

Image: LIGO



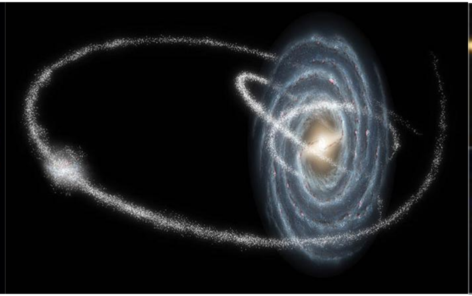
Gravitational waves

Image: Fermi LAT/CW



Gamma rays

Image: NASA/JPL-Caltech/R Hurt (SSC/Caltech)



Stellar streams

Gravitational lensing

Challenge 1) High-dimensional data

Thousands of petabytes of upcoming data
Preprocessed data can have millions of dimensions

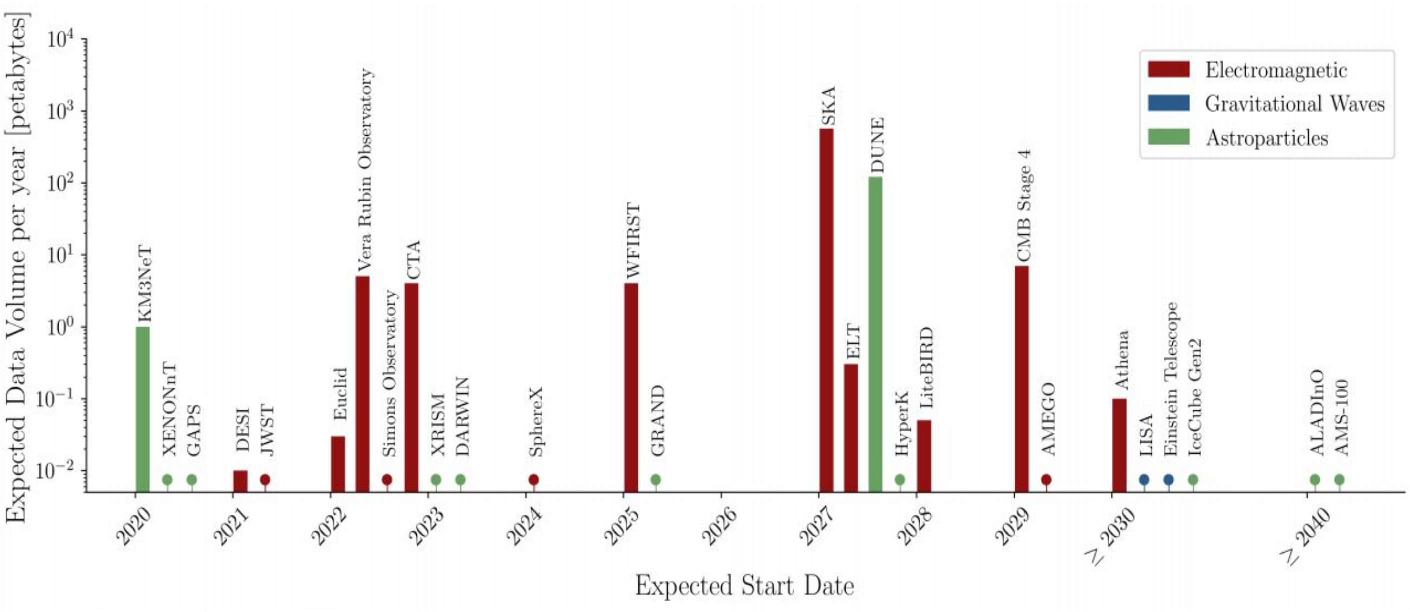
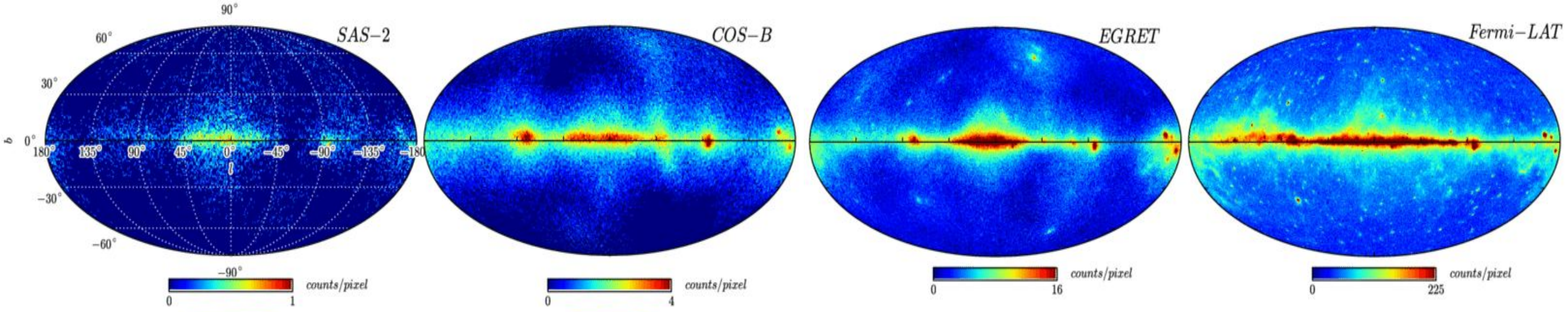


Image credit: EuCAPT white paper 2021

Challenge 2) High-dimensional models



More data → More details → More accurate models

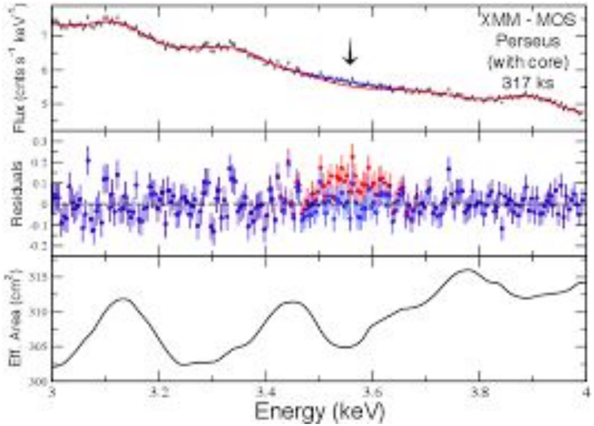
→ **More model parameters**

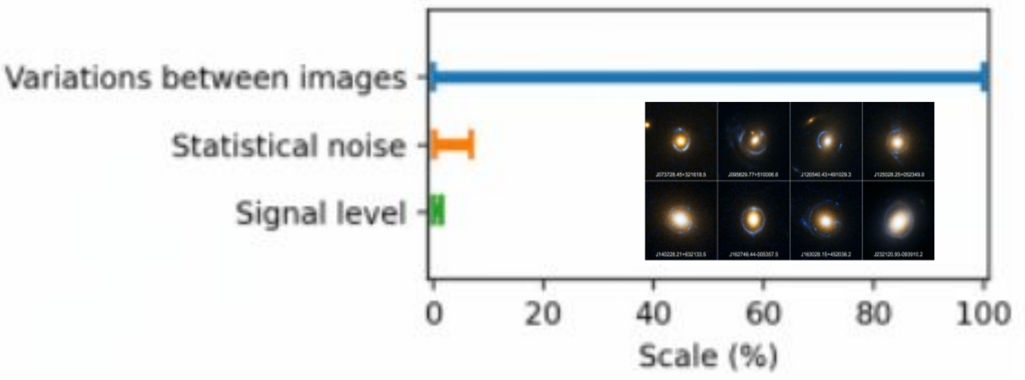
→ **Slower simulations**

Endless statistical analysis challenges

- Hierarchical models (source populations)
- Trans-dimensional models (number of sources)
- Label switching problem (instance detection)
- Parameter degeneracies (distances unclear)
- Non-parametric components (gas maps)
- Inference of fields (diffusion zone)
- Millions of parameters & minutes - hours to evaluate…

Challenge 3) Signals << Background





Any mismodeling of backgrounds, or uncharacterized model uncertainty, or unjustified simplification, does backfire.

New physics searches = Avoiding to shoot yourself in the foot

Image credit: Bulbul+ 1402.2301, Noemi Anau Montel

The inverse problem = the impossible problem?

Millions of parameters

Millions of data points

Very expensive simulation model

Posterior with exponentially
large number of modes

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

Bayes theorem

Solving the inverse problem

de facto standard: Markov Chain Monte Carlo

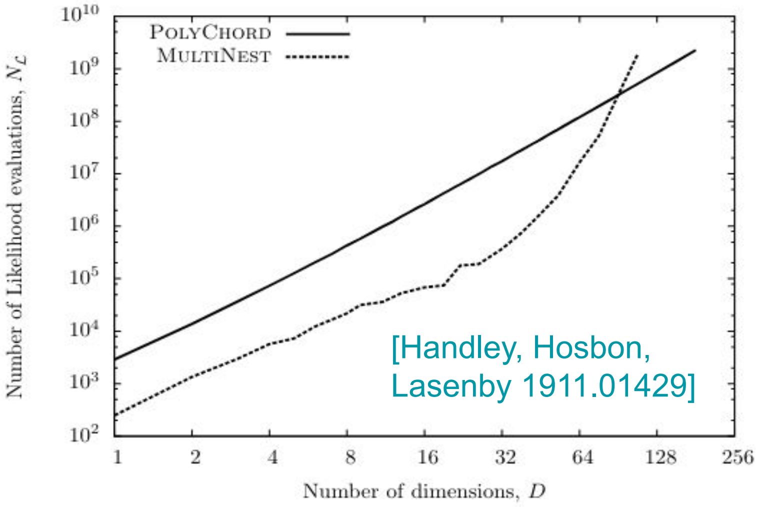
Likelihood

Prior

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

Evidence

Posterior



- Step 1: Samples from joined posterior

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\boldsymbol{x}) \qquad \boldsymbol{\theta} \in \mathbb{R}^D$$

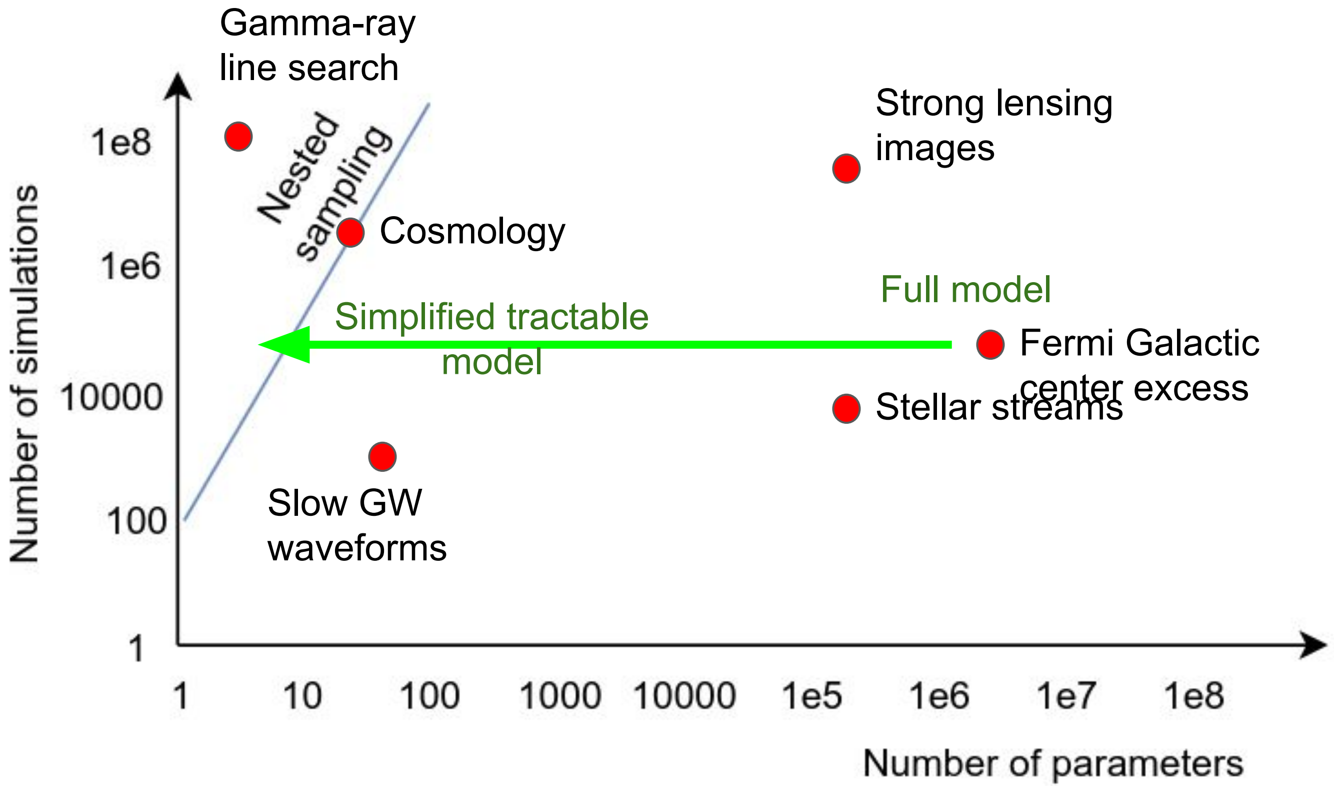
D: Number of parameters

- Step 2: Marginalization to parameters of interest

$$\boldsymbol{\theta} \equiv (\theta_1, \theta_2, \ldots, \theta_D)^T \rightarrow (\theta_i, \theta_j)^T \in \mathbb{R}^2$$

Typical likelihood-based inference algorithms (MH, HMC, VI, …) require a small enough number of parameters.

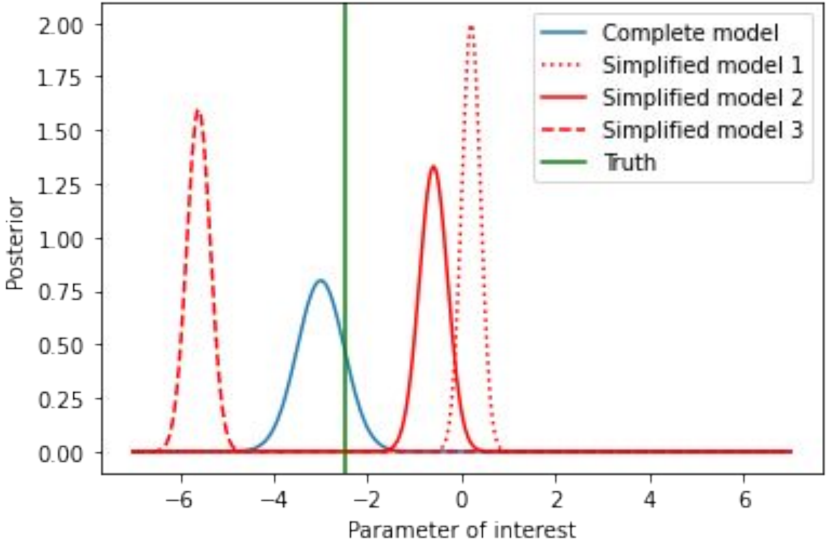
Likelihood-based inference enforces low-dim models



The price of model simplification

Exaggerated (?) illustration of the potential dangers of model simplification

- Biases
- Overly optimistic uncertainties

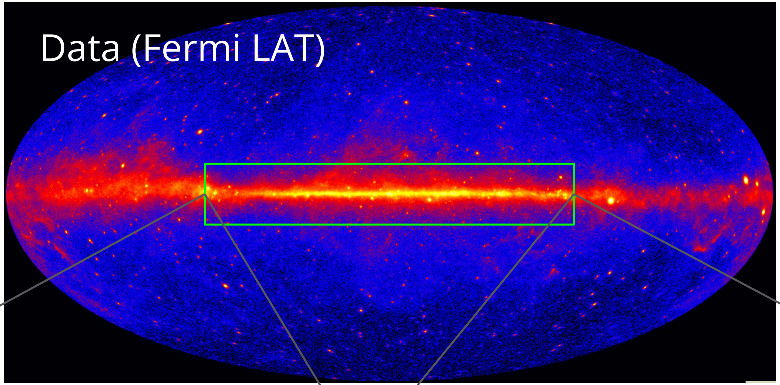


The benefits of high-dimensional models

Almost all existing analysis of Fermi LAT data have these kind of residuals.

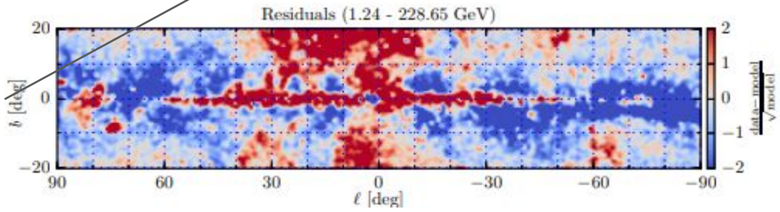
There is no shortage in anomalies in astrophysical data…

Consequences: Large modeling errors because of **simplistic low-dim models**



We pulled this off with **gradient-based optimization.**
Very hard to use in practice, only a handful of examples in the literature.

Residuals
Low-dim model (10 dims)



Storm+ 1705.04065

Residuals
High-dim model (10.000 dims)


nature astronomy ARTICLES
https://doi.org/10.1038/s41550-018-0531-z

The Fermi-LAT GeV excess as a tracer of stellar mass in the Galactic bulge

Richard Bartels[1*], Emma Storm[1], Christoph Weniger[1] and Francesca Calore[2]

Neural simulation-based inference (SBI)

Very active young research field

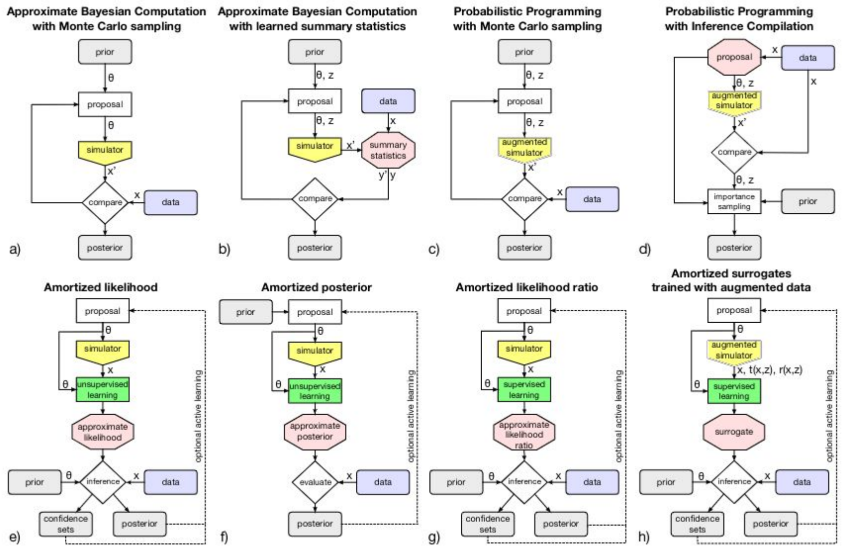


Fig. 3. Overview of different approaches to simulation-based inference.

[Cranmer, Brehmer, Louppe 1911.01429]

General goal: obtain neural network approximator for one of the following:

- Posterior* $p(\boldsymbol{\theta}|\boldsymbol{x})$
- Likelihood* $p(\boldsymbol{x}|\boldsymbol{\theta})$
- Ratios of posteriors and priors = ratios of likelihood and evidence

$$r(\boldsymbol{x},\boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x},\boldsymbol{\theta})}{p(\boldsymbol{x})p(\boldsymbol{\theta})} = \frac{p(\boldsymbol{\theta}|\boldsymbol{x})}{p(\boldsymbol{\theta})}$$

- Various variations of the above quantities…

* require normalization of densities

High-dimensional inference can be simple

	Image, x	*Parameter vector, z*
Examples *x, z ~ p(x|z)p(z)* "Simulated images"		1, 3, 2, **1**, 5, 4, 3, 1, 6, 7, 9, ... 6, 2, 5, **8**, 6, 8, 4, 3,2 1, 3, 4, ... 2, 3, 4, **3**, 1, 7, 8, 9, 5, 3, 2, ... 4, 2, 1, **4**, 6, 8, 6, 4, 3, 2, 4, ... 1, 3, 2, **9**, 5, 4, 3, 1, 6, 7, 9, ... 6, 2, 5, **8**, 6, 8, 4, 3,2 1, 3, 4, ... 2, 3, 4, **1**, 1, 7, 8, 9, 5, 3, 2, ... 4, 2, 1, **2**, 6, 8, 6, 4, 3, 2, 4, ... 1, 3, 2, **4**, 5, 4, 3, 1, 6, 7, 9, ... 6, 2, 5, **4**, 6, 8, 4, 3,2 1, 3, 4, ...
Observed data		?, ?, ?, **8**, ?, ?, ?, ?, ?, ?, ?, ?, ...

Red:
Parameter of interest

Black:
Nuisance parameters
(parametrizing *all* possible background images)

Neural ratio estimation (NRE) in a nutshell

Strategy: Learning to distinguish between **matching (parameter, data) pairs** and **random pairs**.

(1, )

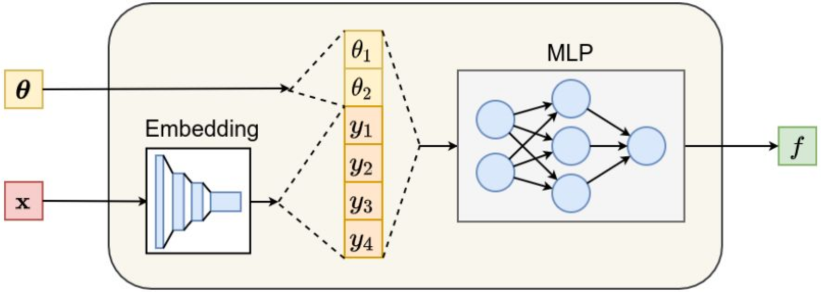
(8, )

(9, )

(4, )

Loss function: Binary cross entropy

$$\ell[f_\phi]_{\mathrm{NRE}} = -\int d\boldsymbol{x}\, d\boldsymbol{\theta}\ \left[p(\boldsymbol{x},\boldsymbol{\theta}) \ln \sigma(f_\phi(\boldsymbol{x},\boldsymbol{\theta})) + p(\boldsymbol{x})p(\boldsymbol{\theta}) \ln\left(1-\sigma(f_\phi(\boldsymbol{x},\boldsymbol{\theta}))\right)\right]$$



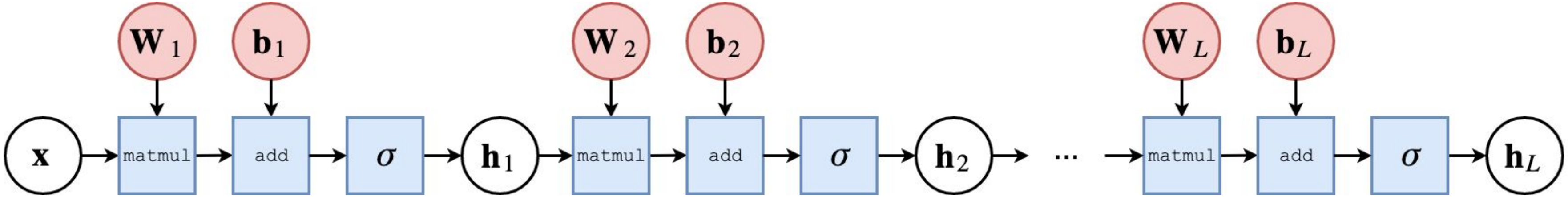
Minimizing network approximates posteriors

$$f_\phi(\boldsymbol{\theta},\boldsymbol{x}) \approx \ln \frac{p(\boldsymbol{x},\boldsymbol{\theta})}{p(\boldsymbol{x})p(\boldsymbol{\theta})} = \ln \frac{p(\boldsymbol{\theta}|\boldsymbol{x})}{p(\boldsymbol{\theta})}$$

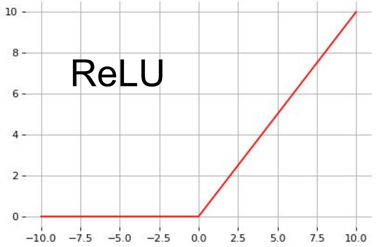
Gutman&Hyvärinen 2010 (as NCE), Mnih&Teh 2012 (self-normalizing), ..., Hermans+ 1903.04057, Miller+ 2107.01214, Cole+ 2111.08030

What is a Multi-Layer Perceptron?

Multi-layer perceptron = MLP = dense network



Activation function



Training would optimize transformation matrices *W* and biases *b.*

Three conjectures for scalable SBI

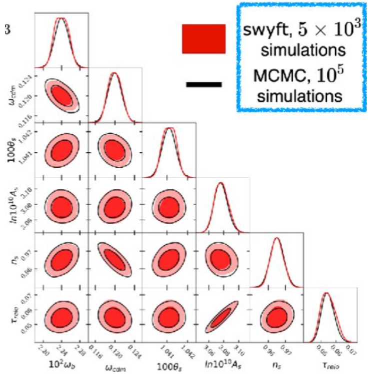
1) Marginal posterior rather than joint posteriors

- A “universal” approach must scale to millions of parameters, and outrageously complex posteriors (transdimensional models, label switching, strong correlations, …)

$$p(z_1, z_2, \ldots, z_{1000000}|\mathbf{x}) \qquad\qquad p(z_1|\mathbf{x}), p(z_2, z_3|\mathbf{x}), p(\max(\mathbf{z})|\mathbf{x}), \ldots$$

Joined: In general intractable
(any approach)

Marginals: Often tractable
(NRE, forward-KL based approaches, …)

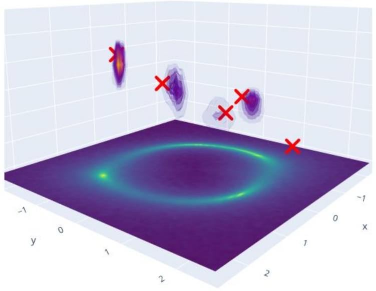


1-dim and 2-dim marginals for corner plots

- Scientifically, we are usually only interested in marginal posteriors anyway
  - Parameter regression: 1-dim marginals
  - Parameter correlations: 2-dim marginals
  - Bayesian model comparison: ratios of marginals
  - Object identification: density functions
  - …

[for discussions see e.g. Alsing+ 1903.01473, Jeffrey+ 2011.05991, Miller+ 2011.13951]

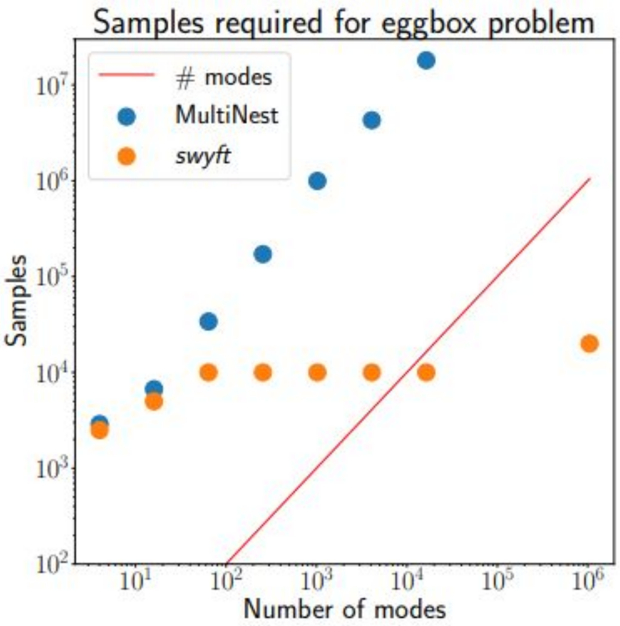
- Caveats: Goodness-of-fit tests, posterior predictive distribution, requires upfront intuition about what matters

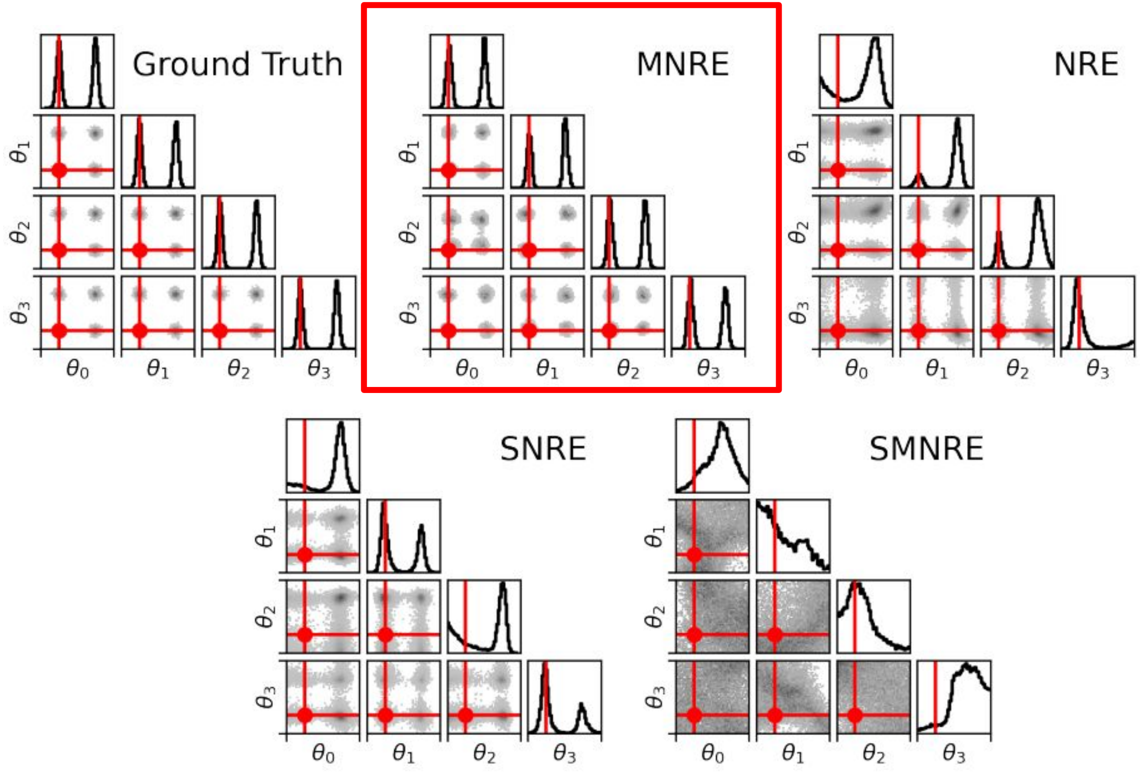


Density functions for object detection

1) <u>Marginal posterior</u> rather than joint posteriors

#modes = 2^#nparams





Estimating marginals breaks naive scaling laws.

[Miller, Cole, Forre, Louppe, **CW** 2107.01214]

2) Truncated priors as sequential proposals

- Sequential techniques are based on targeted training data

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

[Durkan+ 2002.03712 for a discussion]

$$\tilde{p}(\mathbf{z}) \approx p(\mathbf{z}|\mathbf{x_o})$$

- This is fine if the goal is to locally train, e.g., the likelihood (which is prior independent)

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

[Alsing+ 1903.00007 as example (pydelfi)]

- ***But:*** *Marginal* likelihoods/posteriors will be affected by the proposal distribution

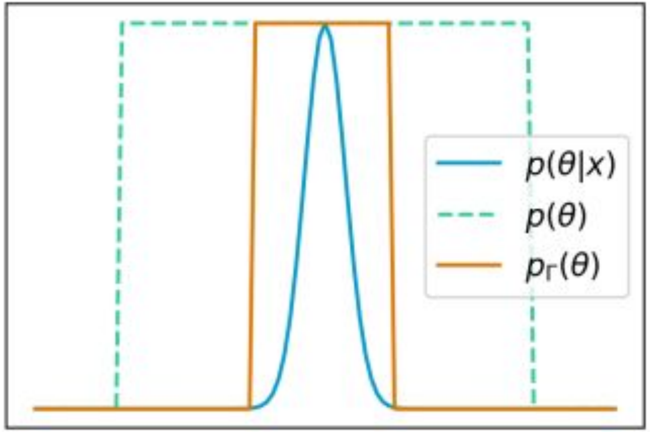
$$p(\mathbf{x}|z_1) = \int dz_2 \dots dz_N \; p(\mathbf{x}|\mathbf{z})\tilde{p}(z_2, \dots, z_N)$$

[see e.g. Alsing+ 1903.01473 for a possible summary-statistics related solution]

- To alleviate this we proposed to use a *truncation scheme*

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

[Miller+ 2011.13951, 2107.01214 - swyft & TMNRE]



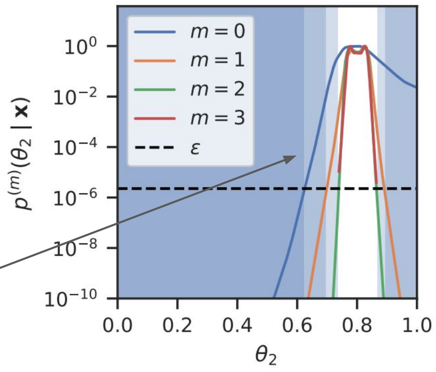
3) Likelihood-to-evidence ratios rather than densities

- Ratio estimation → Binary classification = Battle-proven simplicity

$$f_\phi(\boldsymbol{\theta}, \boldsymbol{x}) \approx \ln \frac{p(\boldsymbol{x}, \boldsymbol{\theta})}{p(\boldsymbol{x})p(\boldsymbol{\theta})} = \ln \frac{p(\boldsymbol{\theta}|\boldsymbol{x})}{p(\boldsymbol{\theta})}$$

[Hermans+ 1903.04057]

[see Cranmer+ 1911.01429 for discussion of many alternatives]



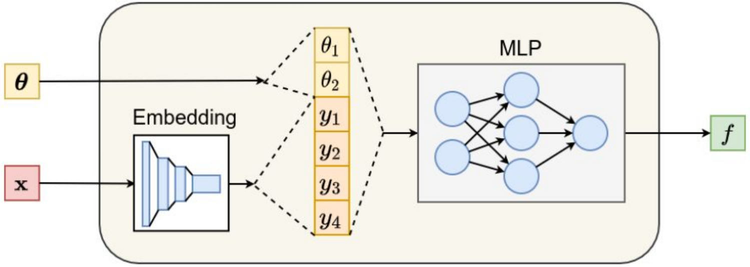
- Usually remains conservative (works well in a truncation scheme)

  [but see Hermans+ 2110.06581]

- Ratio estimation automatically generates information maximizing data compression

$$\ell[\hat{\rho}_\phi] = -2\mathbb{E}_{p(\boldsymbol{x})}\left[\mathrm{JSD}(p(\boldsymbol{\theta}|\boldsymbol{s}(\boldsymbol{x}))||p(\boldsymbol{\theta}))\right] + \text{const}$$

[see Alsing+ 1903.00007 for related discussions in context of likelihood estimation]



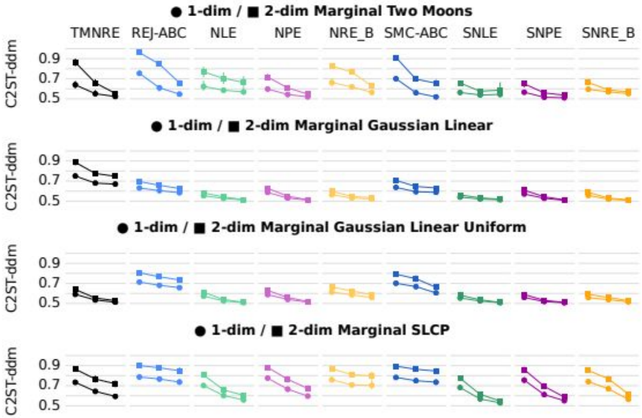
- When focusing on low-dim marginals, sampling is simple (no MCMC or flow-based models required).

1+2+3 = Truncated Marginal Neural Ratio Estimation

NeurIPS | 2021

Thirty-fifth Conference on Neural Information Processing Systems

Competitive performance on standard tasks, but more scalable.



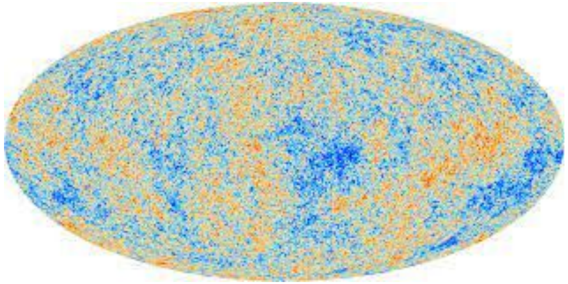
Combination of various properties of existing algorithms

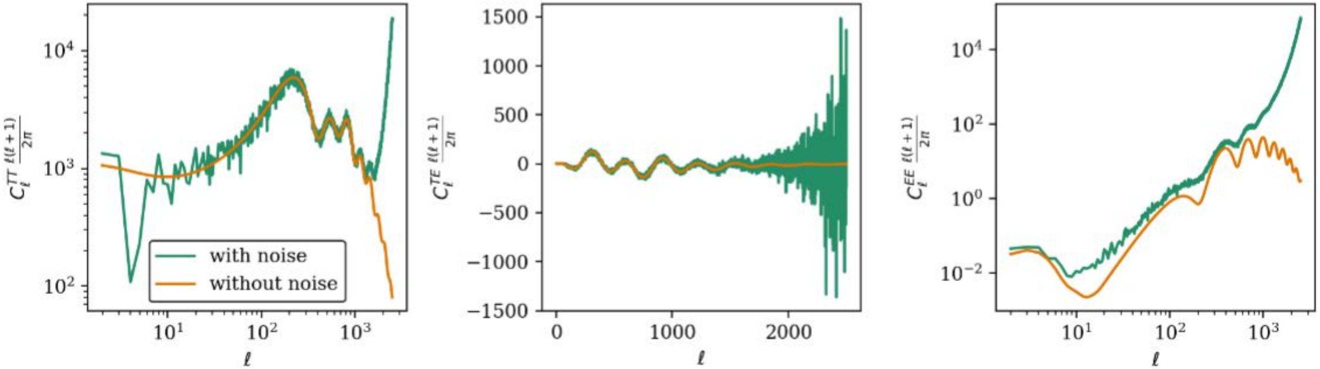
Property / Method	Likelihood-based	ABC	NRE	NPE	SNRE	SNPE	**TMNRE**
Targeted inference	✓	•	✗	✗	✓	✓	✓
Simulator efficient *direct* marginals	✗	✓	•	•	✗	✗	✓
(Local) amortization	✗	✗	✓	✓	✗	✗	✓

Miller, Cole, Forre, Louppe, CW 2107.01214 (NeurIPS)
Miller, Cole, Louppe, CW 2011.13951

Example 1: Cosmic microwave background

Planck cosmology



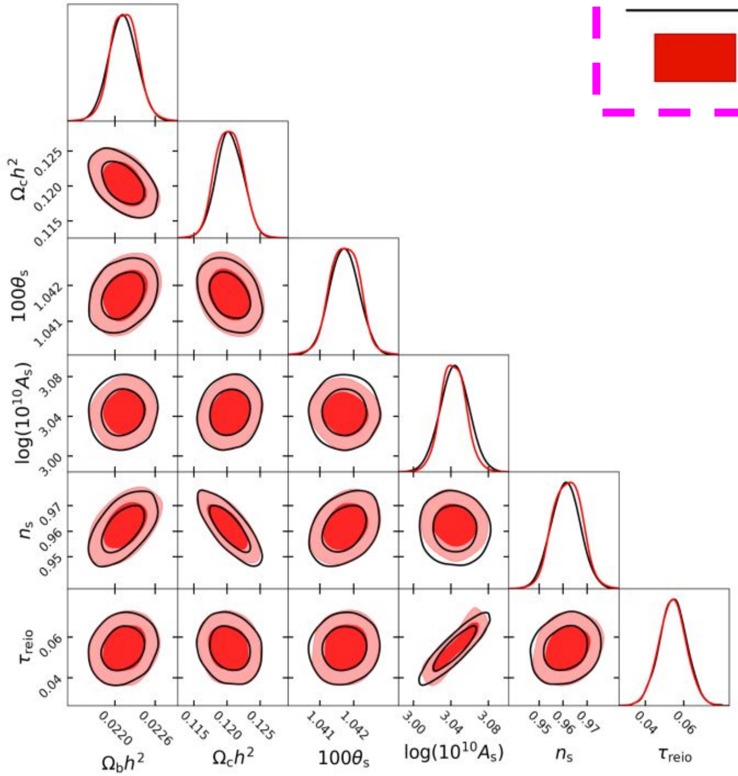


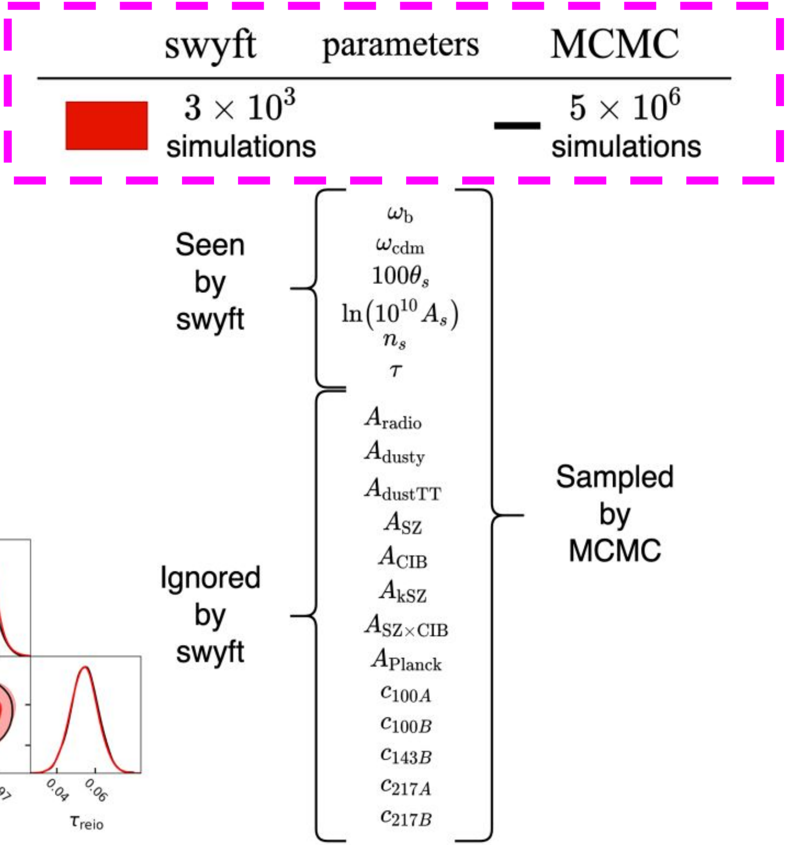
Noise = instrument contribution + cosmic variance

- TT, TE, EE angular power spectrum of CMB with Planck-like noise (Di Valentino+ 2016)
- 6 cosmology parameter to infer, using tight priors (+- 5 sigma Fisher estimate)
- HiLLiPoP likelihood: Planck likelihood,13 varying nuisance parameters [Couchot et al. ‘16]
- Comparison with MCMC is feasible and straightforward
- We use a linear embedding network to go from 7500 → 10 features

[Cole, Miller, Witte, Cai, Grootes, Nattino, CW 2111.08030]

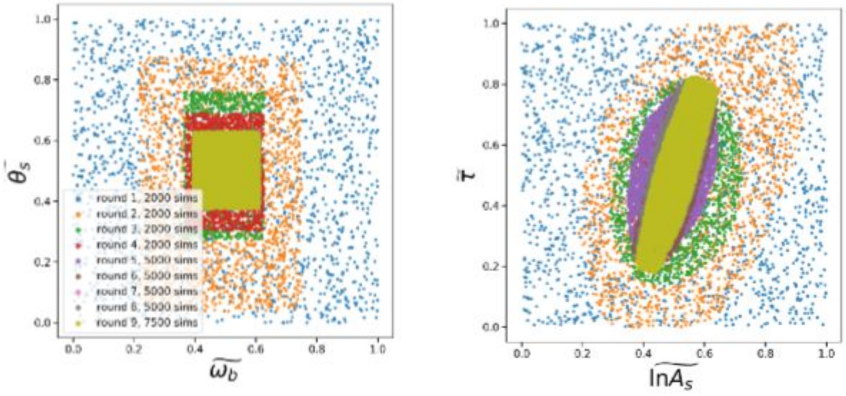
Cosmology with ~1000 times less simultations

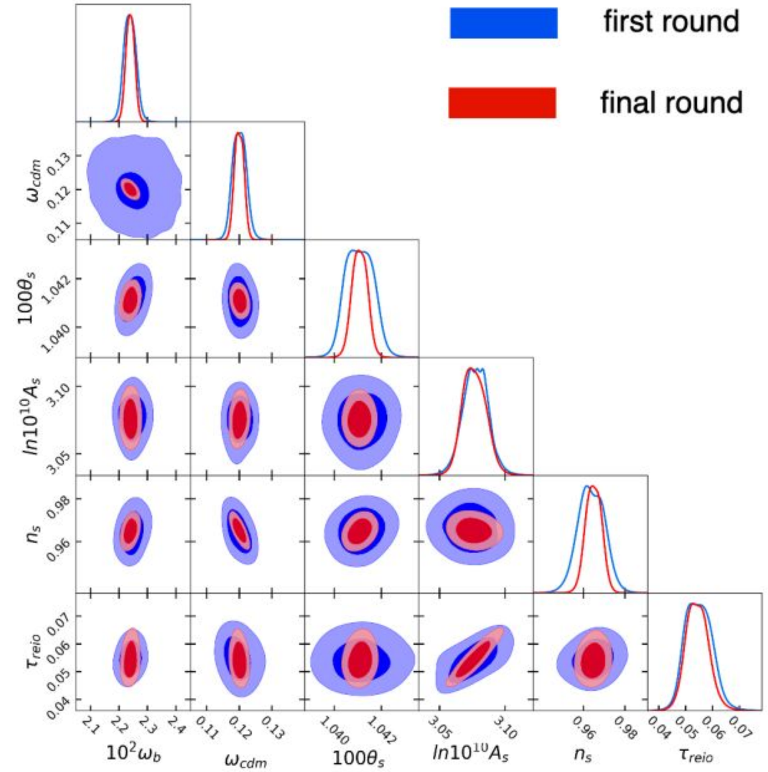




[Cole, Miller, Witte, Cai, Grootes, Nattino, CW 2111.08030]

Simulation efficiency through truncation





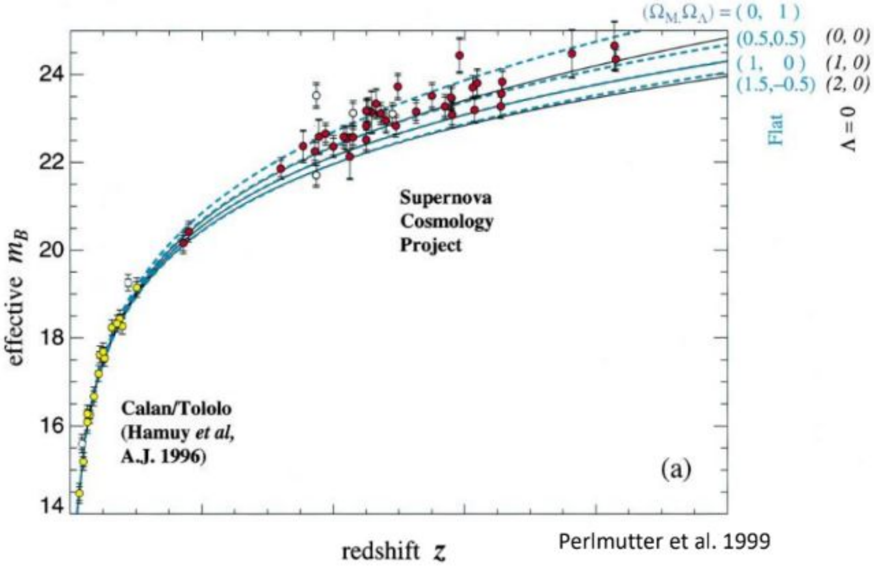
- Demonstration of prior that is “too big” by a factor of 5 for the cosmological parameters
- Truncation effectively identifies region with 20000 extra sims.

Structure of ratio estimator

- Input: Vector (7500)
- Embedding: Linear (7500 → 10)
- Marginals: MLP (19 1-dim, 15 2-dim)

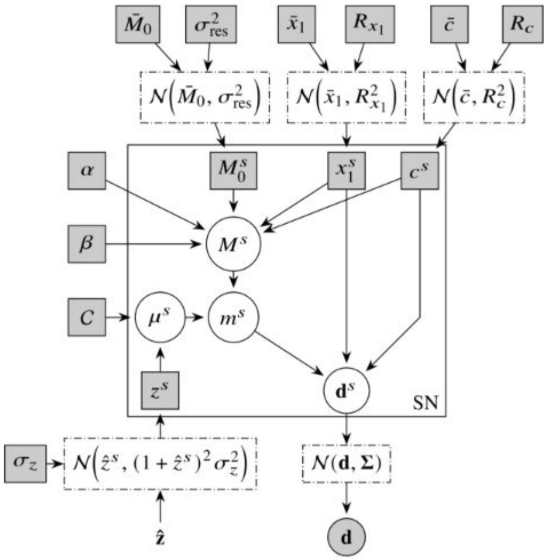
Example 2: Supernova cosmology

Supernova cosmology



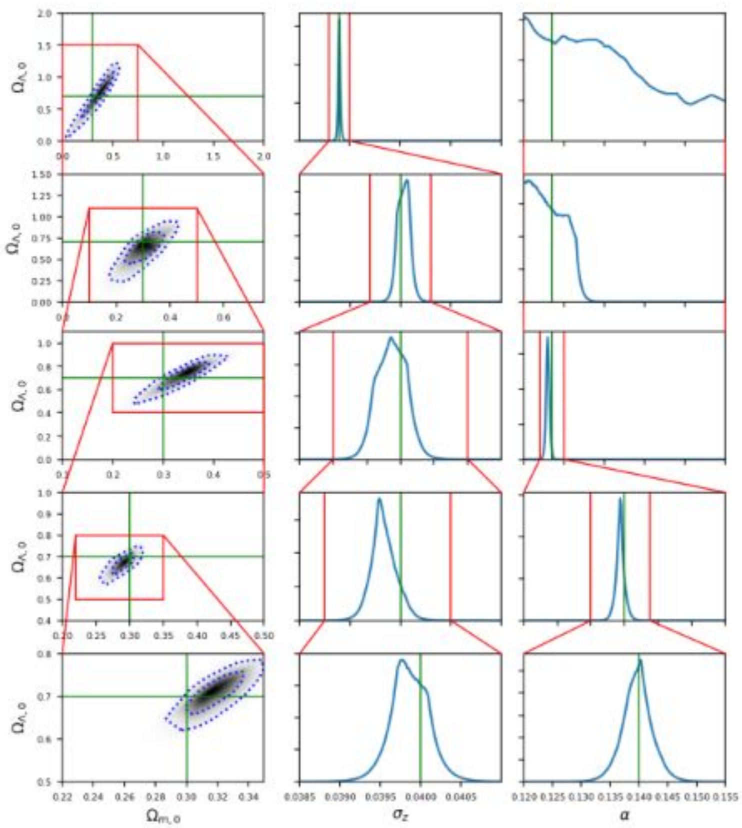
Perlmutter et al. 1999

$$m = M + \mu(z, \mathcal{C}) + \text{"noise"}$$



Ongoing work with Kosio Karchev and Roberto Trotta

Simulation efficiency through truncation



stage 0

train (~1 h)
truncate prior
re-initialise net

stage 1

...

stage 2

stage 3

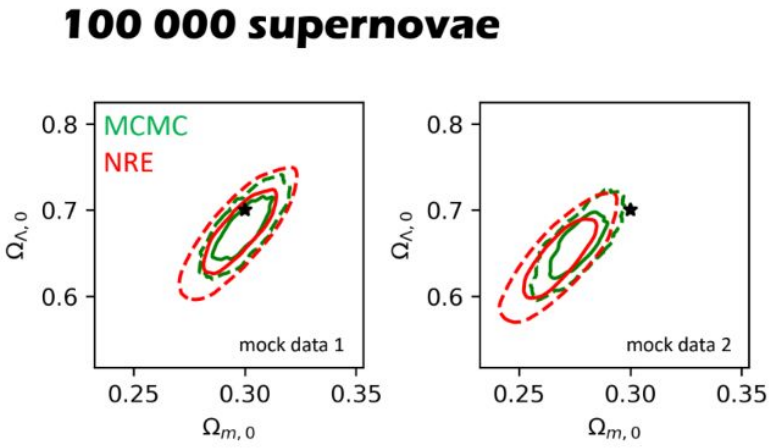
final stage

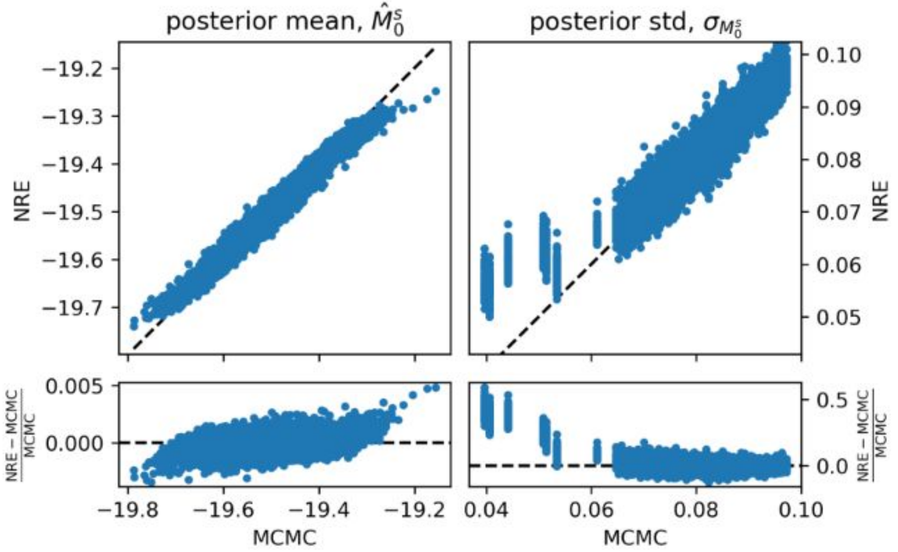
progressively home in on true values

Ongoing work with Kosio Karchev and Roberto Trotta

(Marginal) measurements for 100000 parameters

MALFOI: marginal likelihood-free object-by object inference





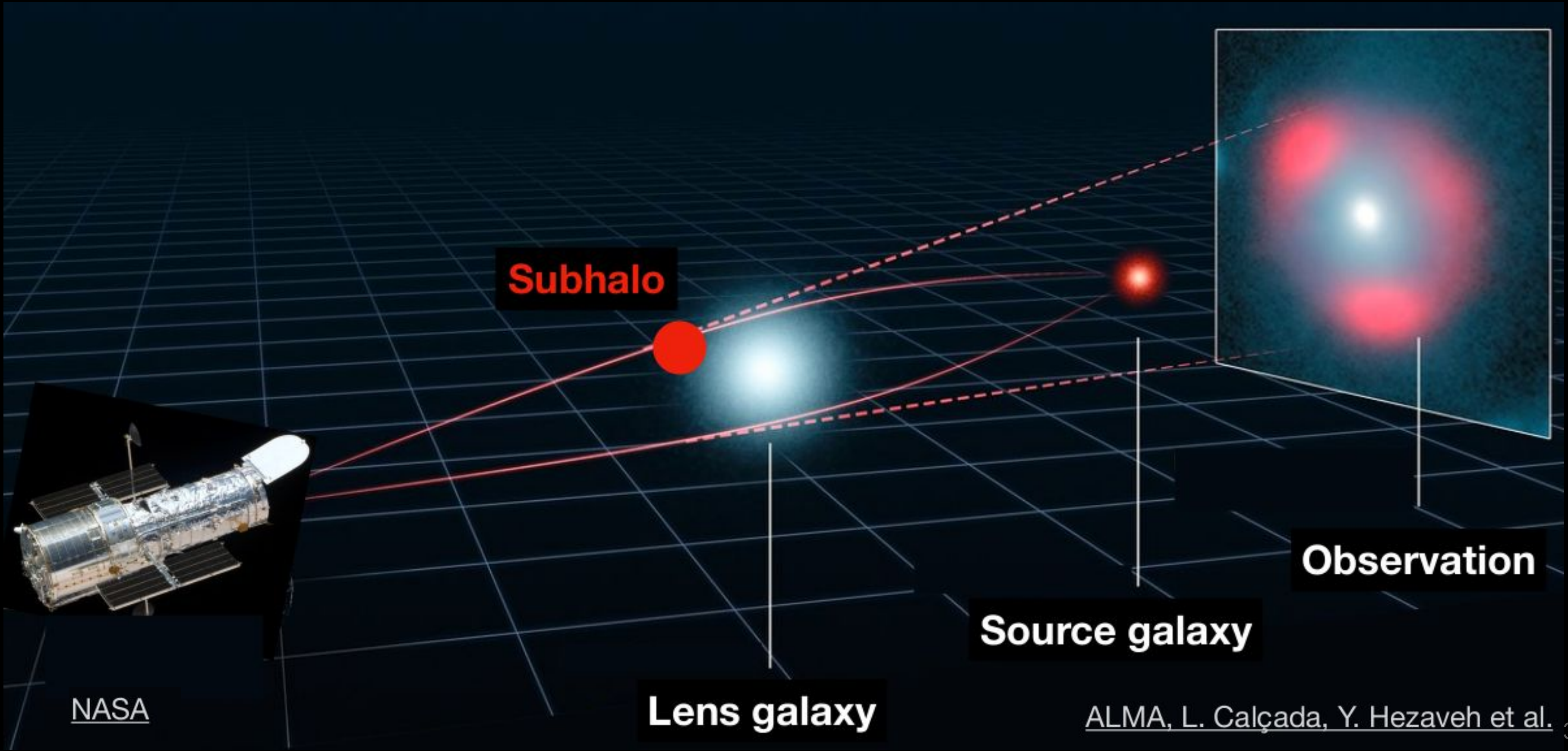
- "MCMC" results were obtained using pre-marginalized likelihoods (only possible under assumptions of Gaussianity and SN independence).
- Instead, NRE marginalizes automatically and *assumption-free.*

Structure of ratio estimator
- Input: 100.000 Spectra (100000, 3)
- Embedding: Linear (300000 → 256)
- Marginals: MLP (100009 1-dim, 1 2-dim)

Ongoing work with Kosio Karchev and Roberto Trotta

Example 3: Strong lensing

Strong galaxy-galaxy lensing



Warm vs cold dark matter

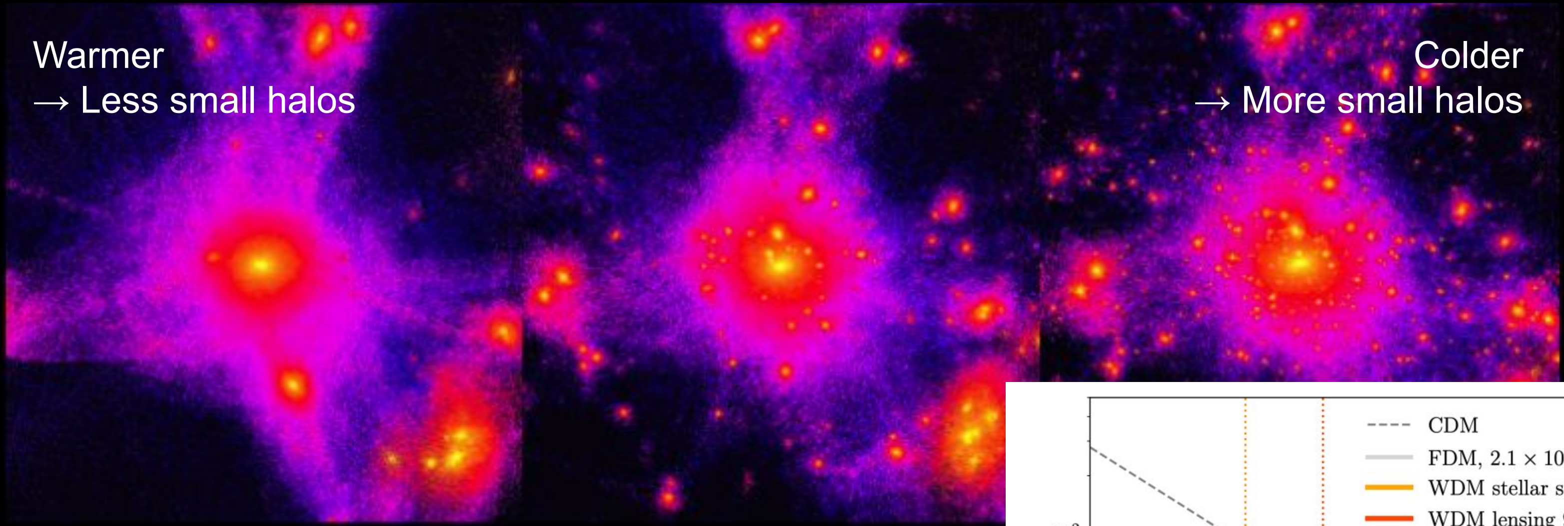
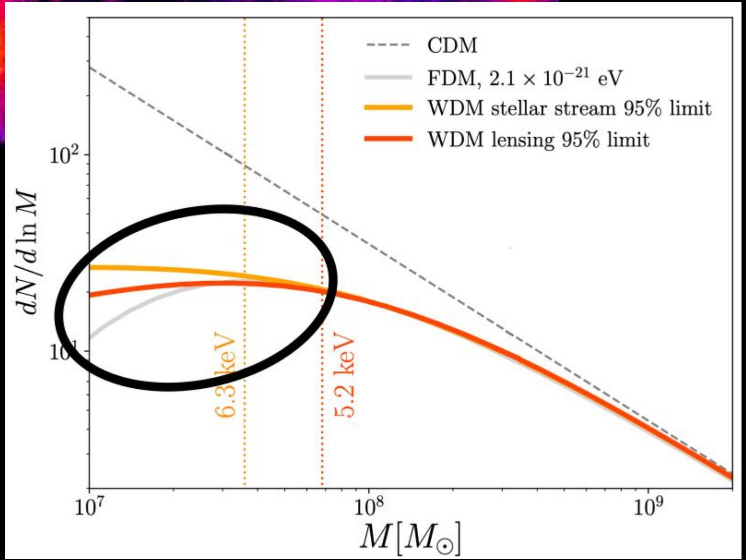


Image credit: ITC @ University of Zurich



Strong lensing animated

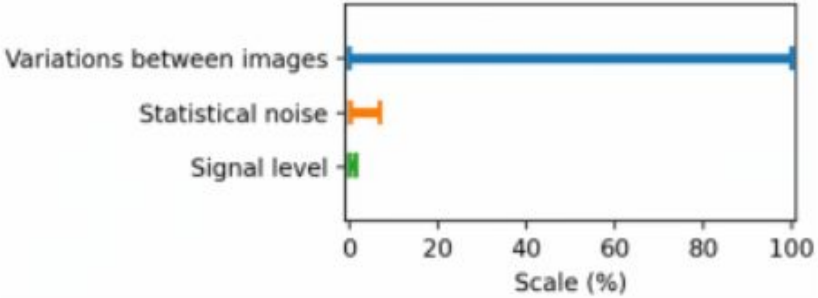
https://adam-coogan.github.io/lensing-multisub/

Inference challenges

- Signal is small compared to noise and variations between images

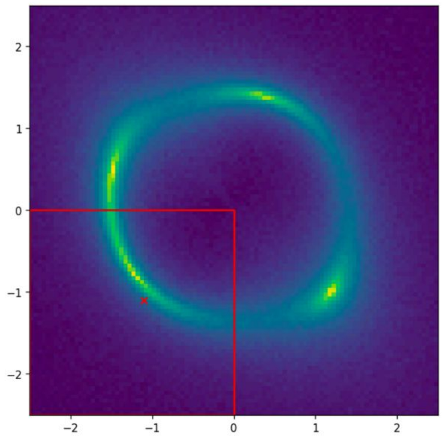
- Marginalization over numerous source, lens and halo parameters

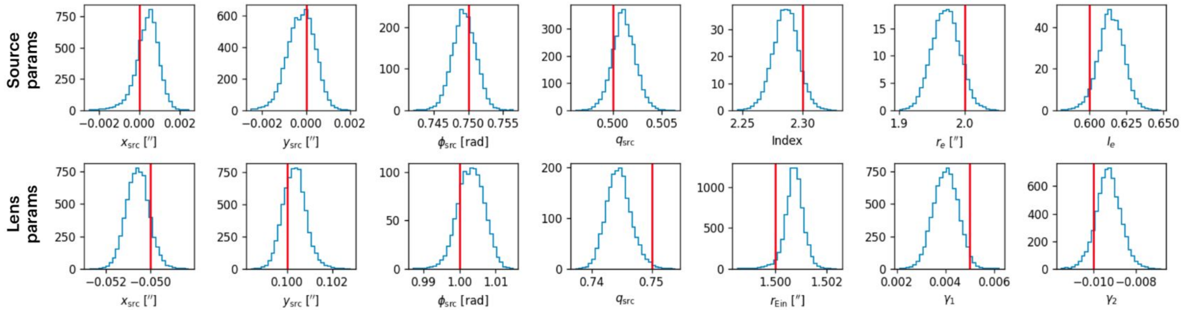
- Joint posterior has $\sim N_{sub}!$ modes; likelihood can be intractable

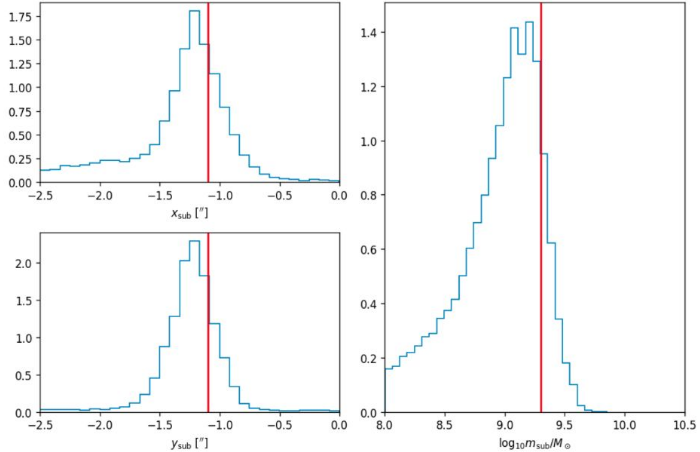


Slide credit: Noemi Anau Montel

A) Single subhalo, simple source model







Structure of ratio estimator

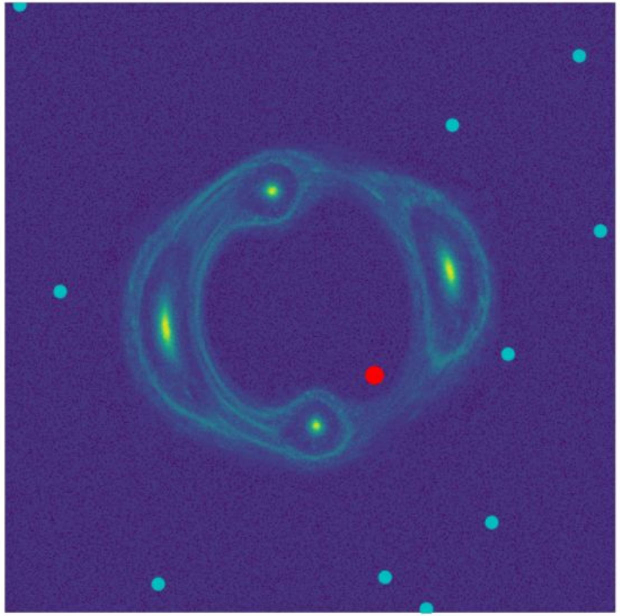
- Input: Images (typically 200x200)
- Embedding: CNN
- Marginals: MLP (17 1-dim)

Ongoing work led by Adam Coogan

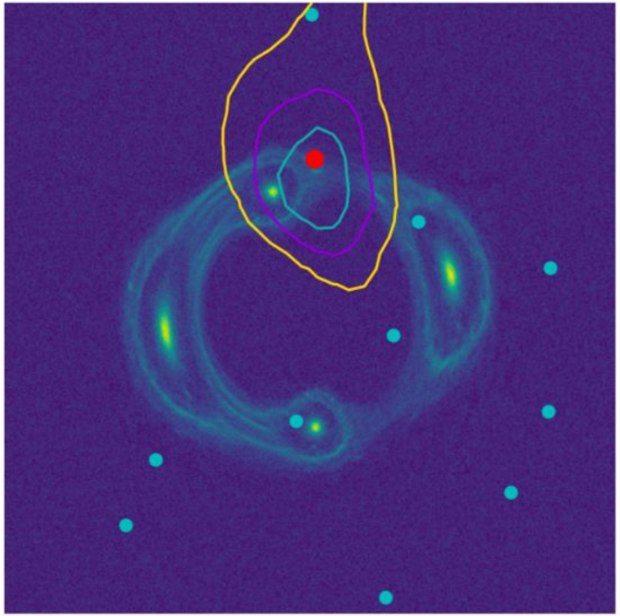
Slide credit: Noemi Anau Montel

B) Multiple subhalos, complex source model

Training data



Inference



• = $5 \times 10^9\, M_\odot$, • = $10^8 - 10^9\, M_\odot$

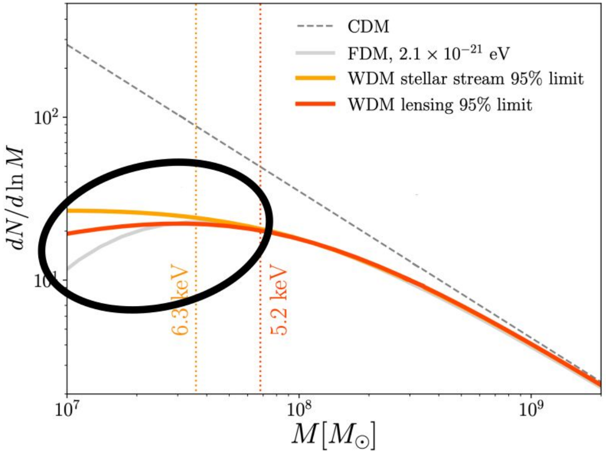
Marginalized over source, lens and halo population

Structure of ratio estimator

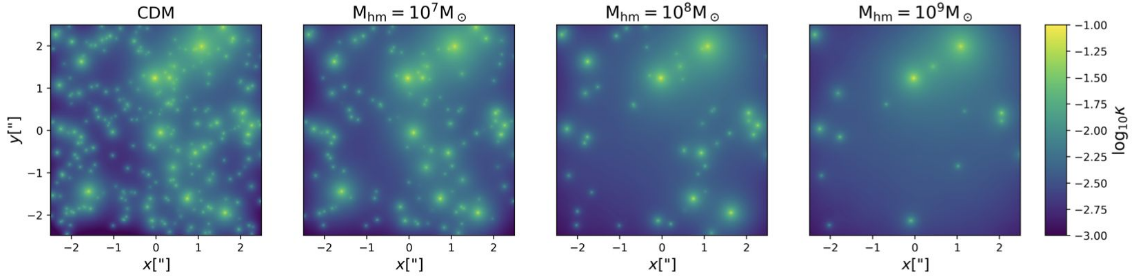
- Input: Images (typically 200x200)
- Embedding: CNN
- Marginals: MLP (2-dim)

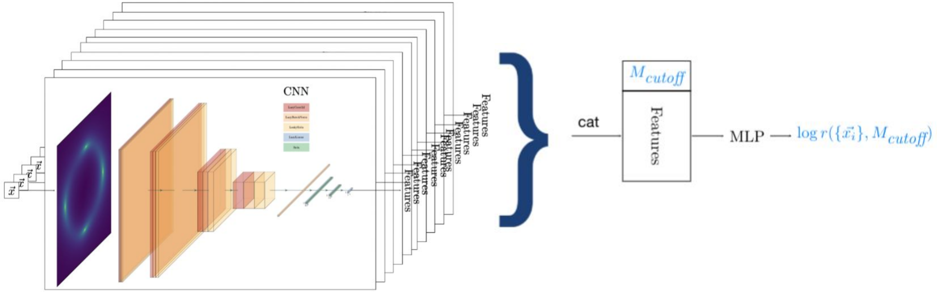
Ongoing work led by Adam Coogan

C) Subhalo mass function cutoff - Combined analysis



Combining observations to reduce subhalo shot noise





Structure of ratio estimator

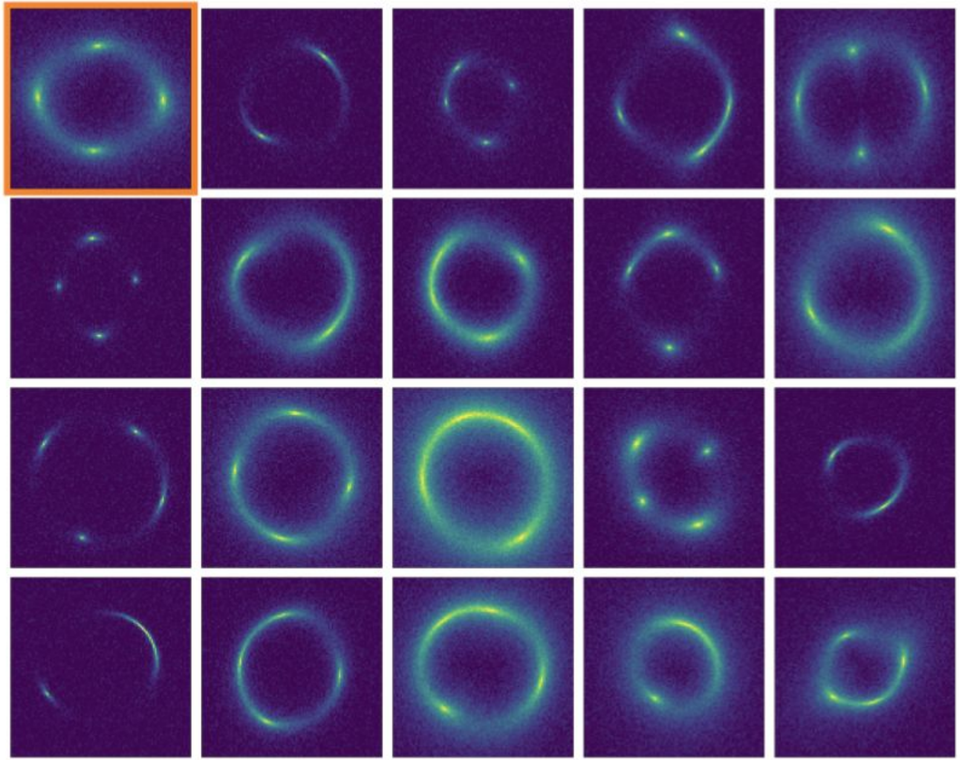
- Input: 10 Images (10x100x100)
- Embedding: Stack of CNNs
- Marginals: MLP (1-dim)

Anau Montel+ 2205.09126

Related work: He+ 2010.13221 (similar in spirit, using ABC), Wagner-Carena+ 2203.00690 (constraining subhalo mass function normalization)

C) Subhalo mass function cutoff - Combined analysis

Subset of 20 target images in our analysis.

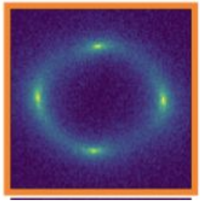


Let’s focus on this one, and call it “A”

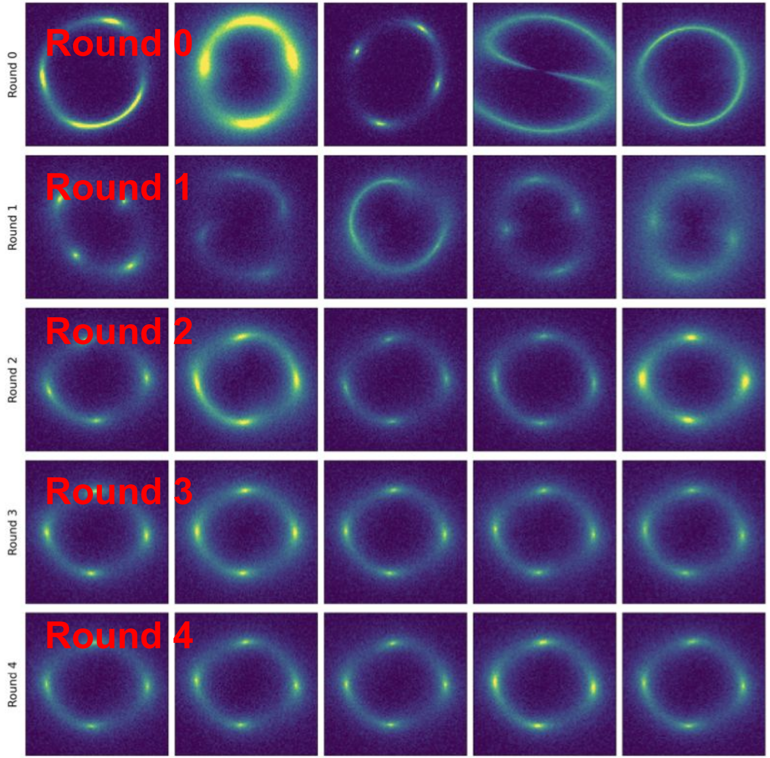
Anau Montel+ 2205.09126

C) Subhalo mass function cutoff - Prior truncation

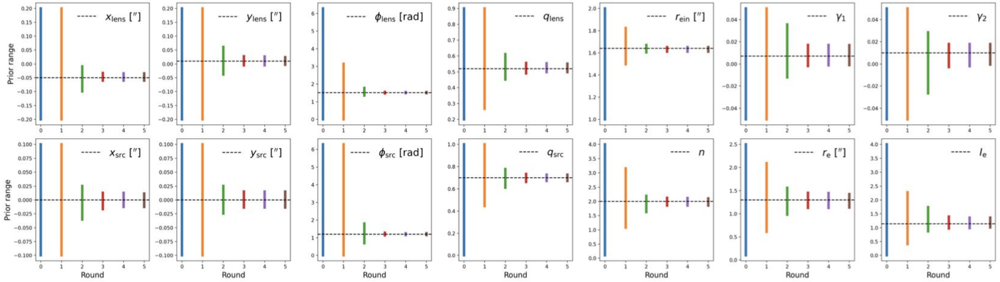
Target image "A"



Training data for target image "A"



Constrained prior ranges (round 0 - round 4) for all 14 main lens & source parameters

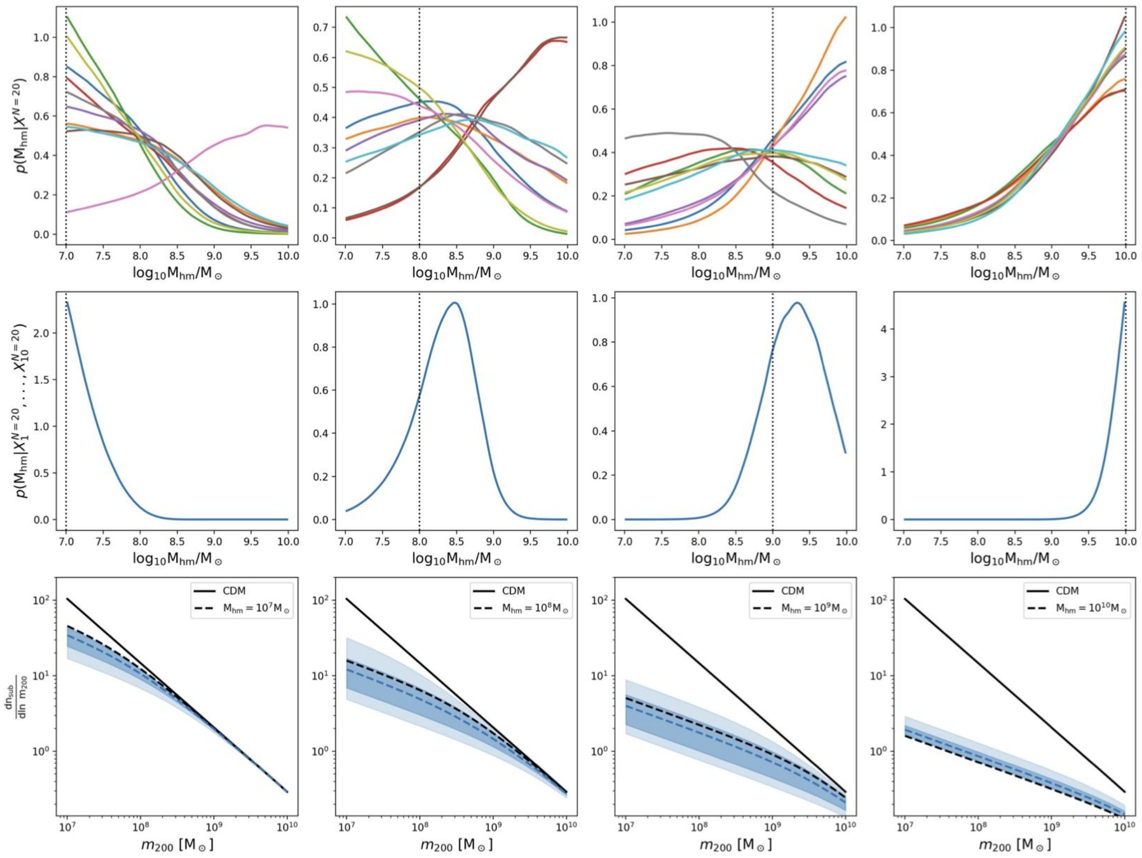


Anau Montel+ 2205.09126

C) Subhalo mass function cutoff - Results

Generating targeted training data for 100 images and combining their constraining power gives tight constraints on the subhalo mass function.

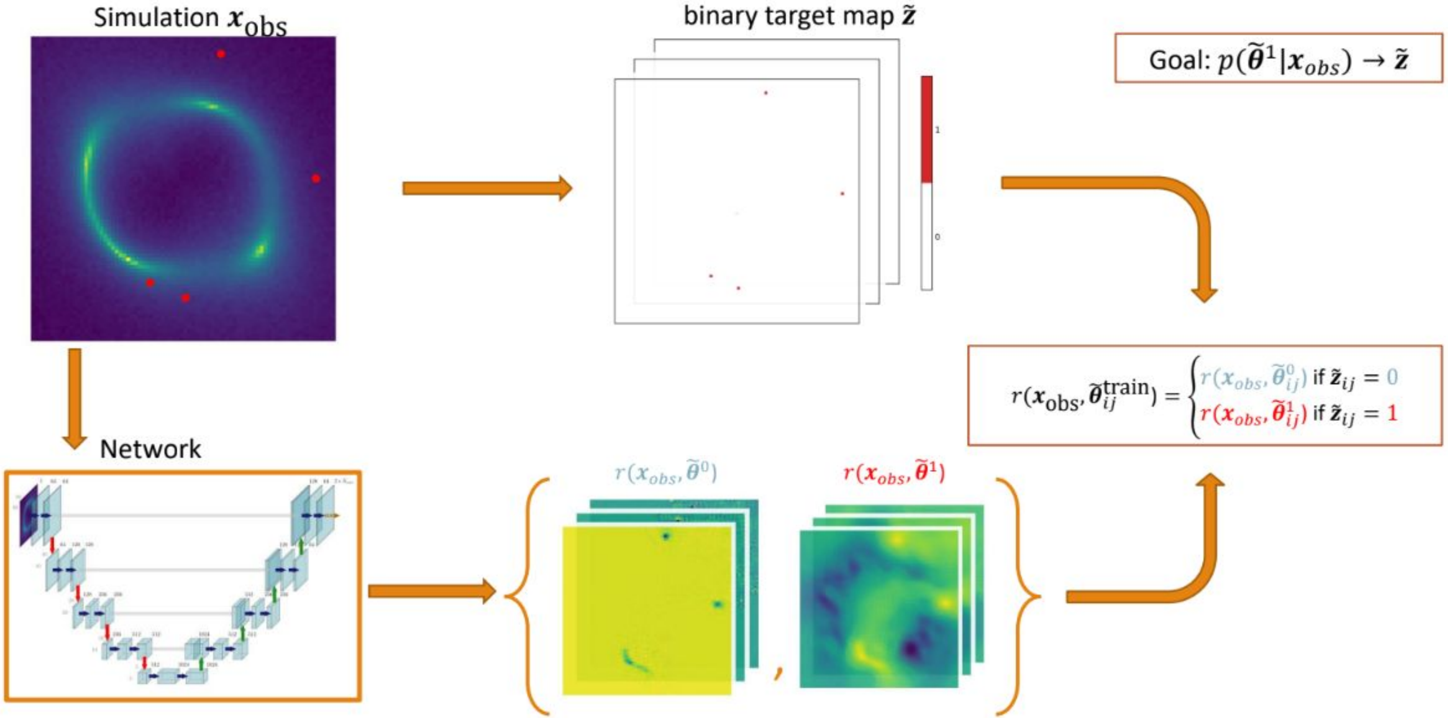
$$p(M_{\text{cutoff}} \mid \{\vec{x}_i\}_{i=1,\ldots,10})$$



Anau Montel+ 2205.09126

D) Halo detection - Probabilistic image segmentation

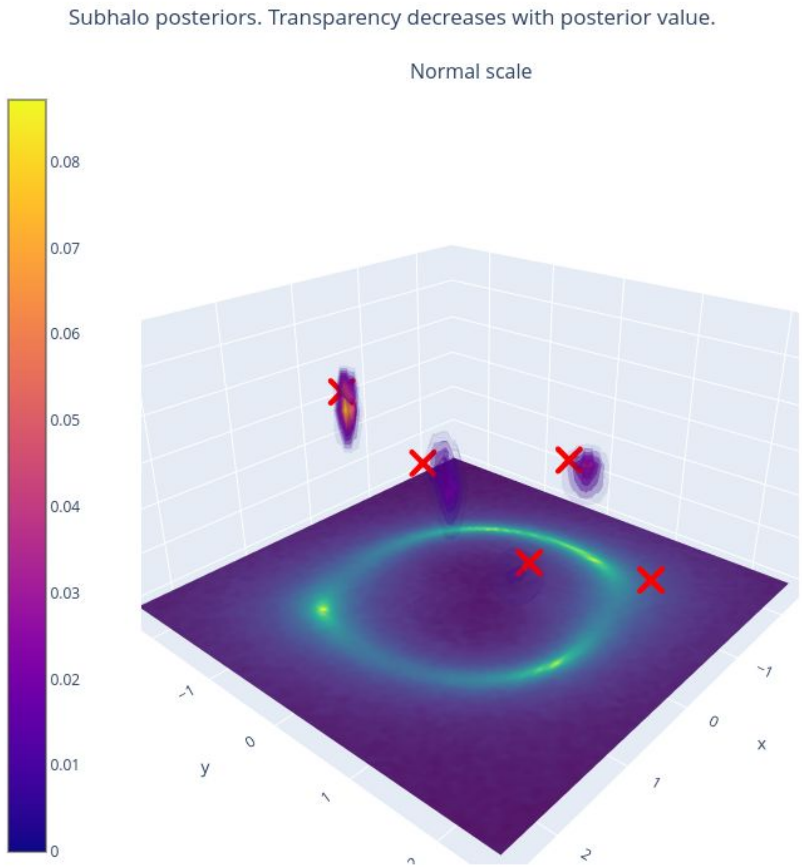
In the presence of multiple subhalos, we can also estimate the subhalo density function (which can be understood as marginal of the more complex joined subhalo distribution).



Ongoing work led by Elias Dubbeldam

Related work: 2009.06639 Ostdiek+ (using U-Net for subhalo detection)

D) Halo detection - Probabilistic image segmentation



Structure of ratio estimator

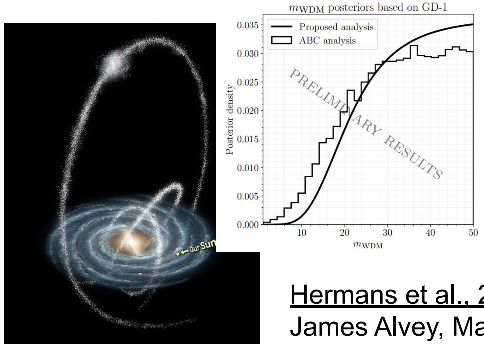
- Input: Image (typically 100x100)
- Embedding: U-Net
- Marginals: Binary marginals 100x100x10 (ten mass bins)

https://dm-lensing-parislfi.github.io/

Ongoing work led by Elias Dubbeldam

TMNRE/SWYFT appear to be broadly applicable

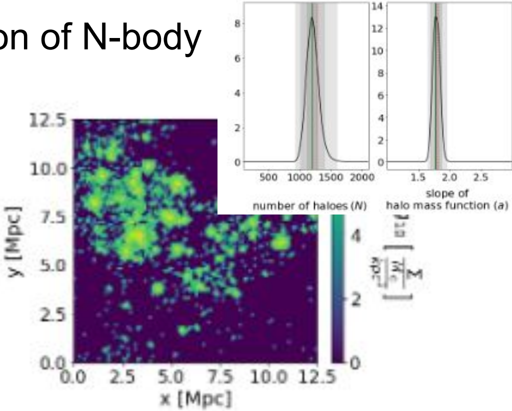
Stellar streams



Hermans et al., 2020
James Alvey, Mathis Gerdes, in progress

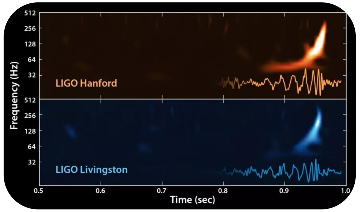
TMNRE/SWYFT

Interpretation of N-body simulations



Androniki Dimitriou+, soon

Gravitational waves



Delaunoy+ 2020
Uddipta Bhardwaj+, in progress

21 cm cosmology
LHC pheno fits
Fermi data
...

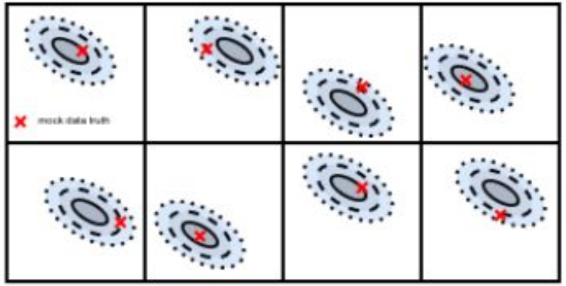
How can one trust results?

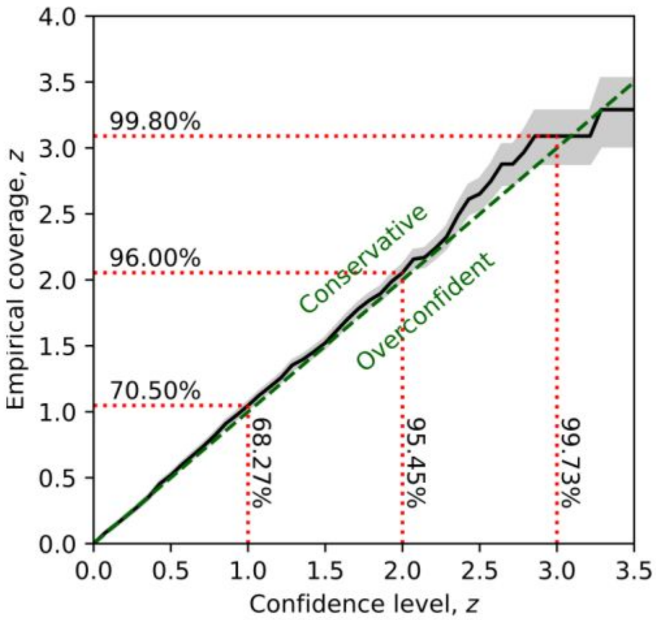
Credibility of inference results can be tested

Let $\Theta_{p(\boldsymbol{\vartheta}|\boldsymbol{x})}(1-\alpha)$ denote the $1-\alpha$ highest posterior density region

Expected coverage of the 68% and 95%

$$1-\hat{\alpha} = \mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\mathbb{1}\left[\boldsymbol{\vartheta} \in \Theta_{\hat{p}(\boldsymbol{\vartheta}|\boldsymbol{x})}(1-\alpha)\right]\right]$$

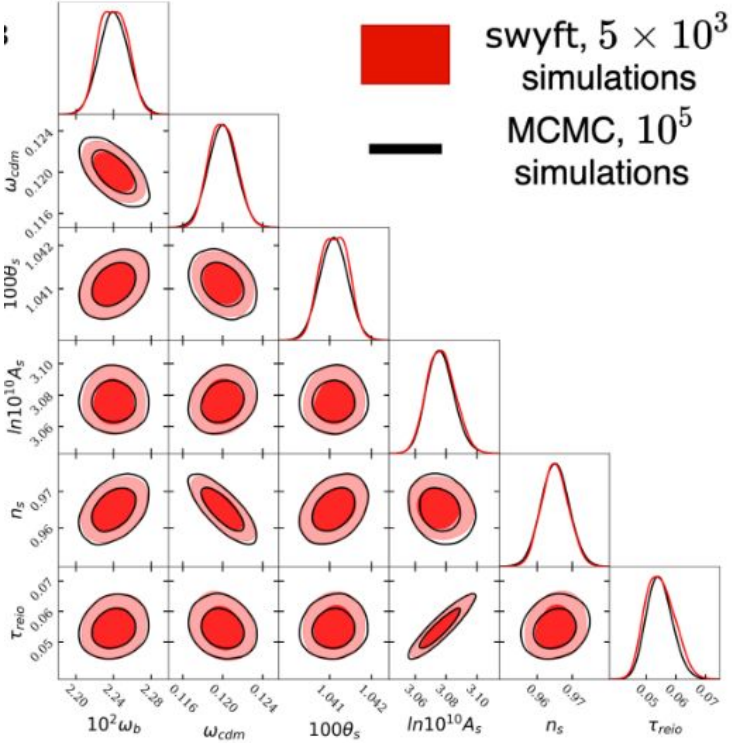




[Cole, Miller, Witte, Cai, Grootes, Nattino, CW 2111.08030]

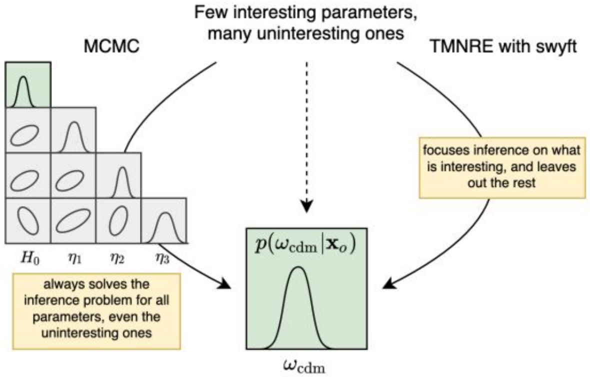
See also Hermans, Delaunoy, Rozet, Wehenkel, Louppe 2110.06581

Coverage tests!

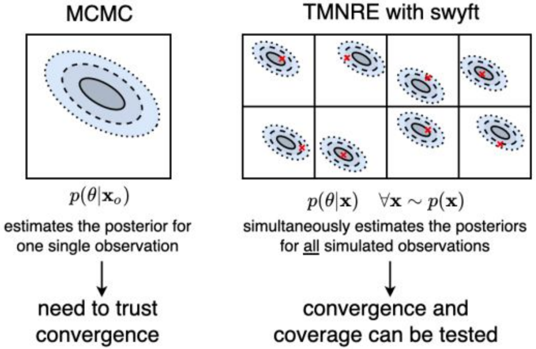


Open source package SWYFT

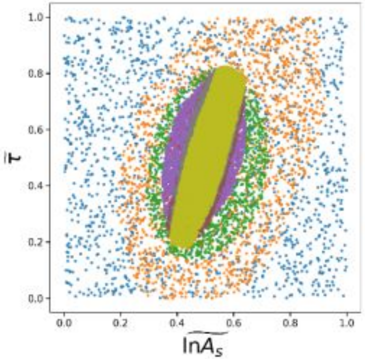
Estimating marginals of interest



Coverage tests



Truncation schemes



Check it out on: https://github.com/undark-lab/swyft
(under heavy development)

PyTorch

PyTorch Lightning

Conclusions

Conclusions

- Simulation-based inference (SBI) has the potential to deal with **ultra-high dimensional inference problems**.
- We discussed a few components that we found very useful in practice, and which are part of **TMNRE**
  - **Neural ratio estimation** offers flexibility and simplicity
  - Focus on **marginal posteriors** rather than the joint
  - **Prior truncation**
- We demonstrated that this framework is promising in tackling a wide range of astrophysical / cosmological data analysis problems. Domain knowledge enters the analysis in terms of network architectures.
  - **CMB Cosmology**
  - **SN Cosmology**
  - **Strong lensing image analysis**
- We provide a **software implementation for TMNRE ("swyft")**, which we currently use for a much wider range of dark-matter-related analysis problems.