GRAPPA GRavitation AstroParticle Physics Amsterdam

UNIVERSITY OF AMSTERDAM

ML in Cosmology

Towards high-precision deep learning with TMNRE

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)

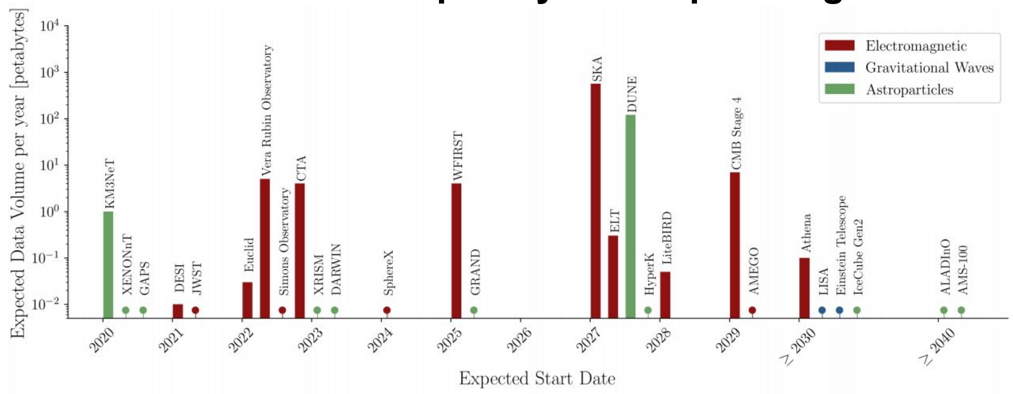
European Research Council
Established by the European Commission

netherlands eScience center

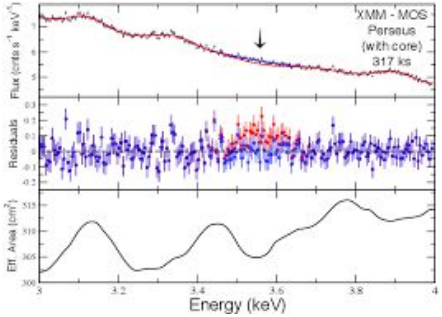
CERN 5th Inter-experimental Machine Learning Workshop, Geneva / virtual
10 May 2022

New-physics searches with astrophysical data

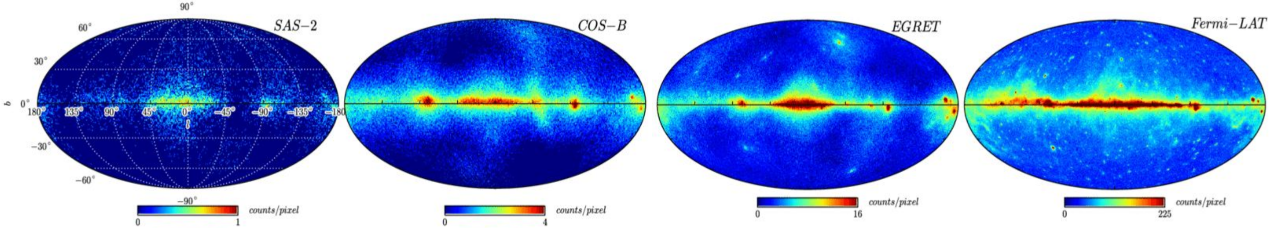
Thousands of petabytes of upcoming data



Statistics → systematics limited searches



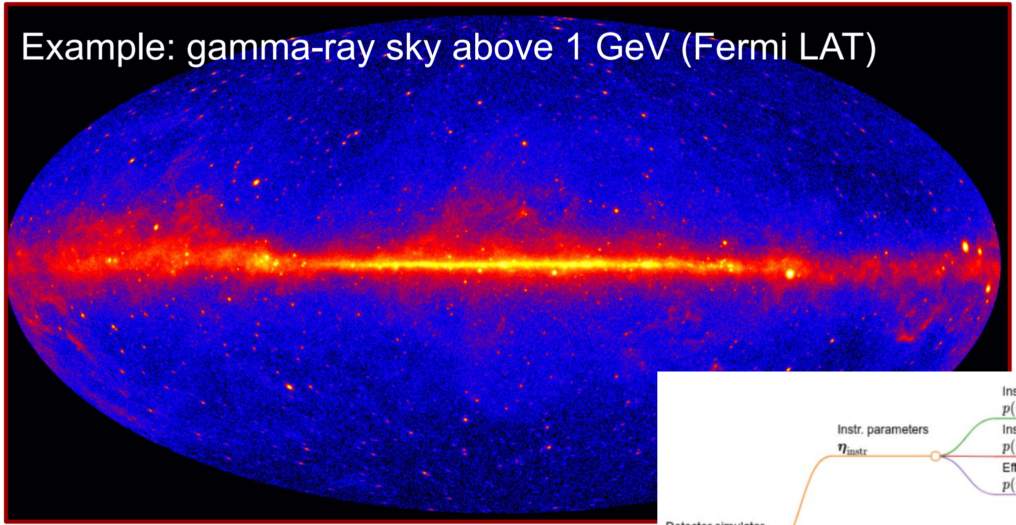
More data → more details



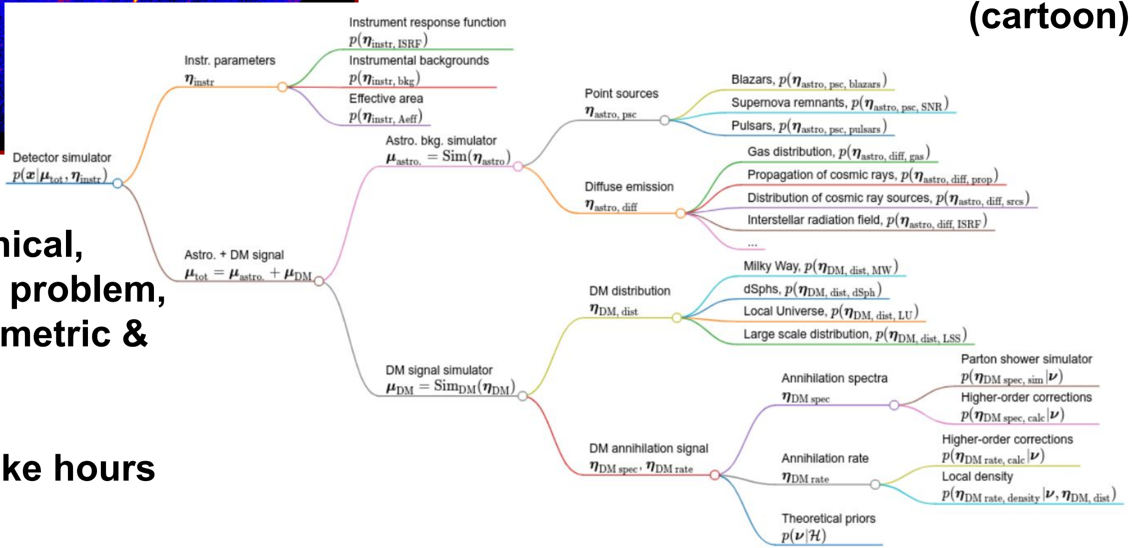
Increased need for high-fidelity astrophysical models & analyses

Image credit: EuCAPT white paper 2021 & A. Morselli, Bulbul+ 1402.2301

Astrophysical models can be really, REALLY complex



Bayes Net for astrophysical model (cartoon)



→ Millions of parameters, hierarchical, trans-dimensional, label switching problem, parameter degeneracies, non-parametric & empirical model components, …

Single forward simulation might take hours

Industry standard: Markov Chain Monte Carlo

Bayes theorem

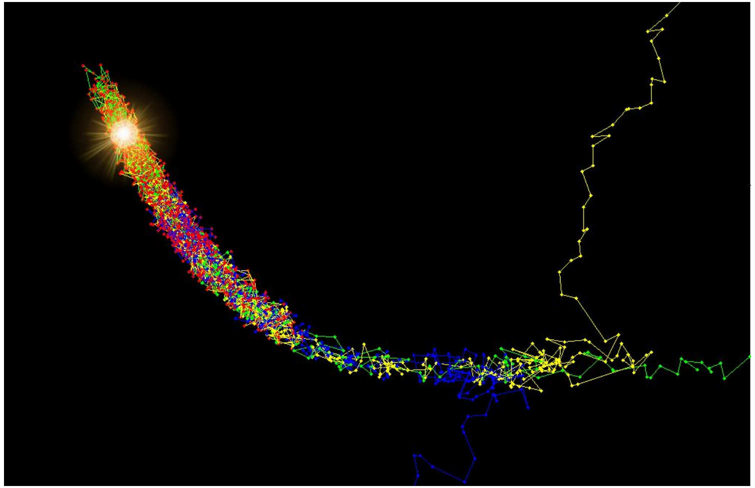
$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{x})}$$

Likelihood

Prior

Posteriors

Evidence



Ex: Metropolis Hastings Algorithm

- Step 1: MC method samples from the **joined high-dimensional posterior for *all* parameters**

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

D: Number of parameters

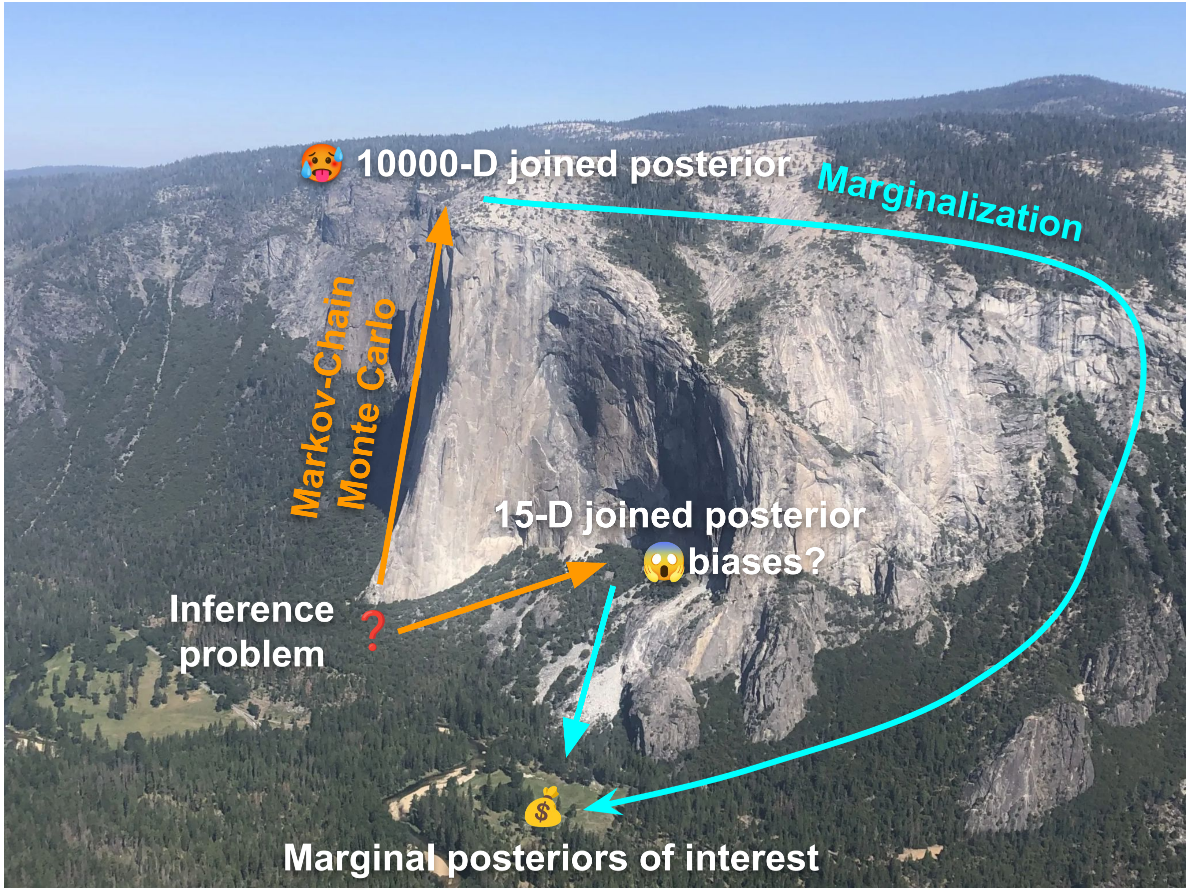
- Step 2: **projection onto parameters of interest**

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

💰**Science result**

Image credit: Wikipedia

Mount joined posterior estimation

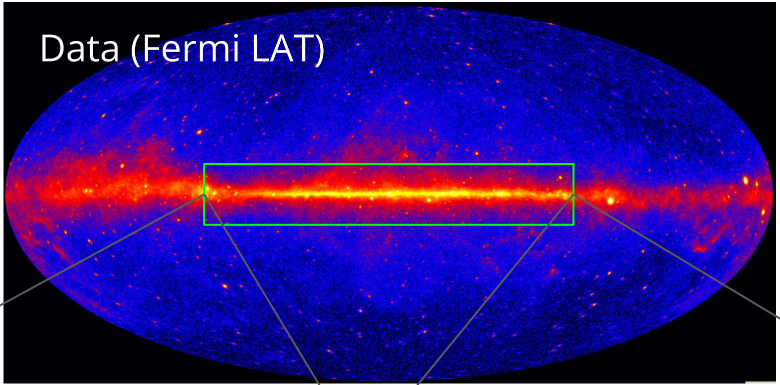


The price of model simplification

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

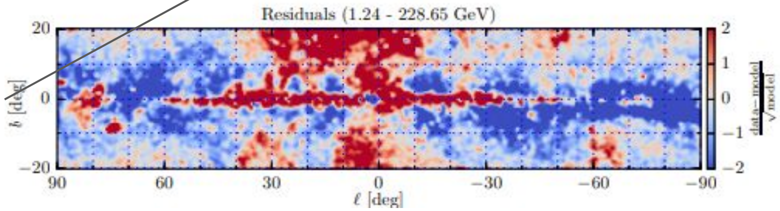
There is no shortage in anomalies in astrophysical data…

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

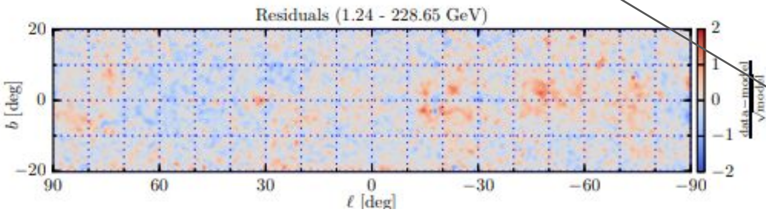


We pulled this off with **gradient-based optimization.**
Very hard to use in practice, only a handful of examples in the literature.

Residuals
Low-dim model (10 dims)

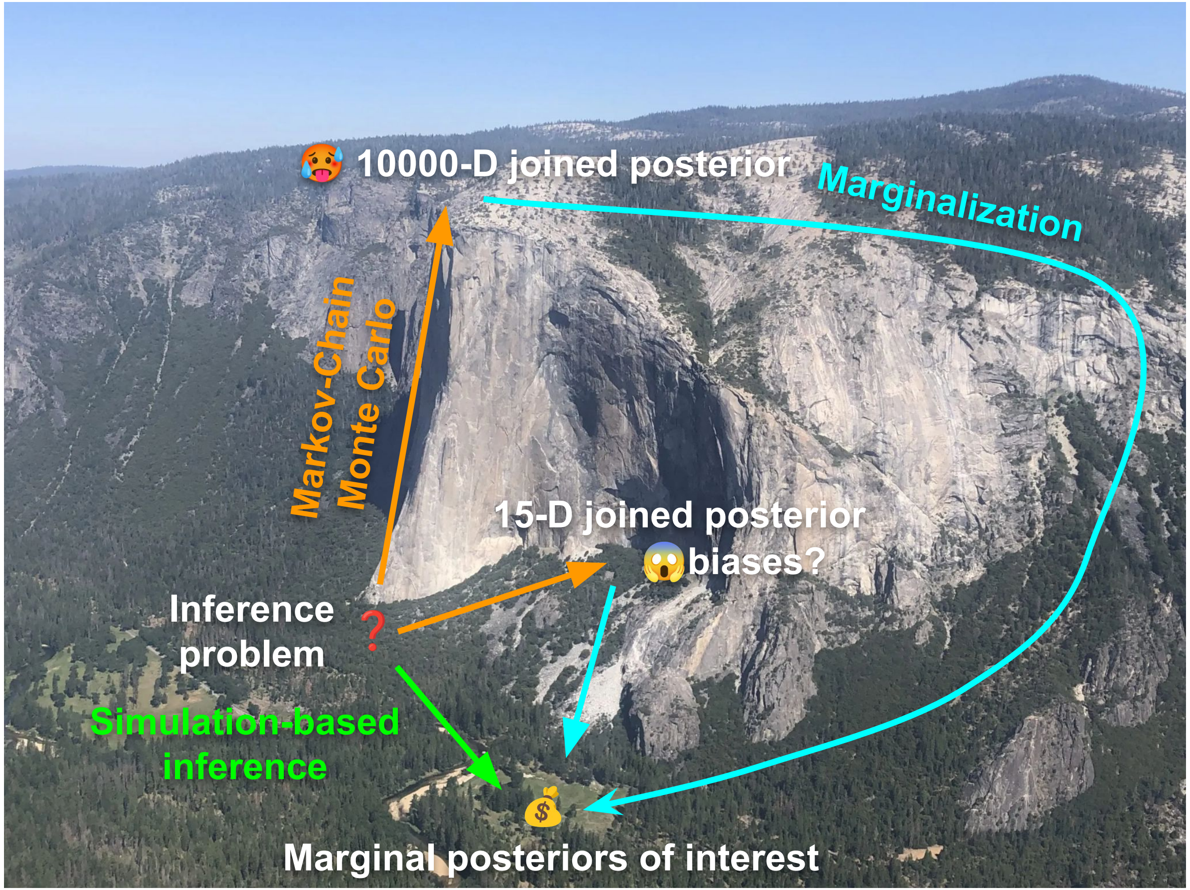


Residuals
High-dim model (10.000 dims)



Storm+ 1705.04065

Mount joined posterior estimation - avoiding the detour



A simulation-based inference thought experiment

"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)

MCMC

⏳

After a
Hubble time*

Neural network

~8

After less than
a second

*10 Million parameters with N^3 scaling for number of likelihood-evaluation, assuming each evaluation takes 1ms.

Neural simulation-based inference (SBI)

Very active young research field

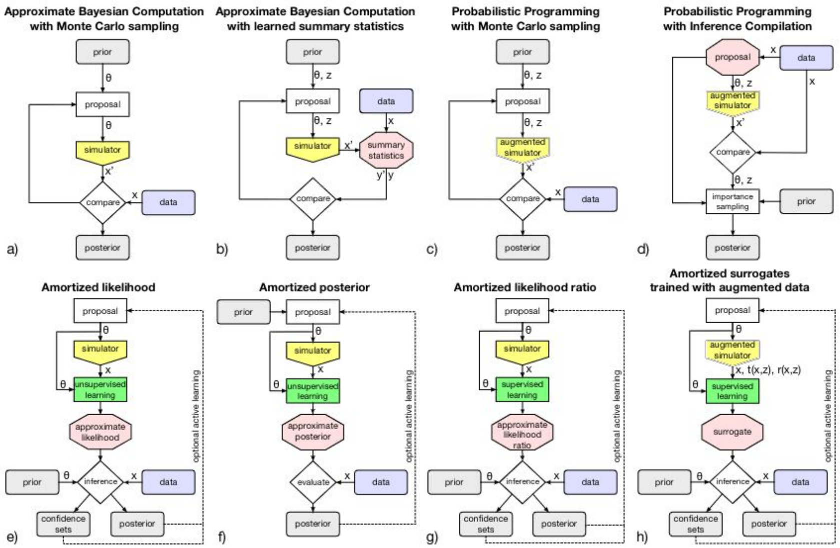


Fig. 3. Overview of different approaches to simulation-based inference.

[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…

* typically based on density estimation with flow-based architectures

Neural ratio estimation (NRE) in a nutshell

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

(**1**, )

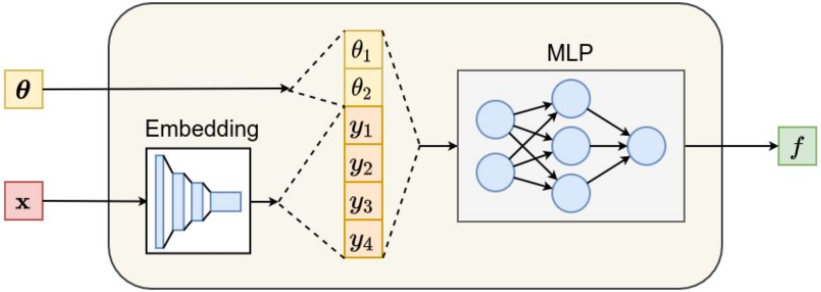
(**8**, )

(**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

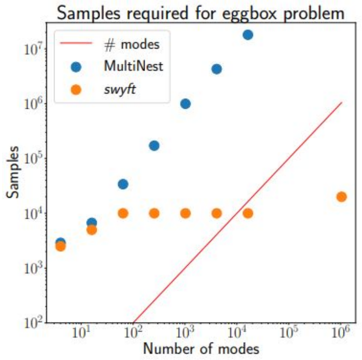
$$f_\phi(\boldsymbol{\theta},\boldsymbol{x}) \approx \ln \frac{p(\boldsymbol{x},\boldsymbol{\theta})}{p(\boldsymbol{x})p(\boldsymbol{\theta})} = \ln \frac{p(\boldsymbol{\theta}|\boldsymbol{x})}{p(\boldsymbol{\theta})}$$

Hermans+ 1903.04057, <u>Miller+ 2107.01214</u>, <u>Cole+ 2111.08030</u>

Truncated Marginal Neural Ratio Estimation (TMNRE)

Competitive performance on standard tasks

Scalable to high-dim models



Key features

1. Focus on Marginals
2. Truncation
3. Neural Ratio Estimation

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	✗	✗	✓	✓	✗	✗	✓

NeurIPS | 2021
Thirty-fifth Conference on Neural Information Processing Systems

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

1) Marginal posterior rather than joint posteriors

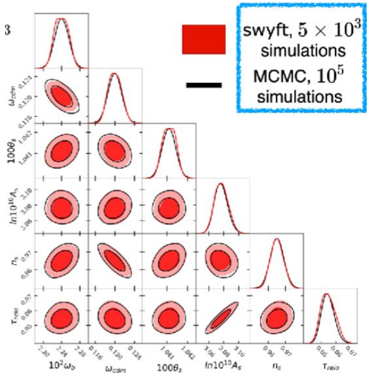
- A “universal” approach must scale to millions of parameters, and outrageously complex posteriors (transdimensional models, label switching, strong correlations, …)

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

Joined: In general intractable
(any approach)

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

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

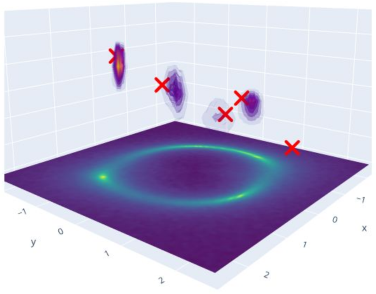


1-dim and 2-dim marginals for corner plots

- 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

2) Truncation rather than sequential

- 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

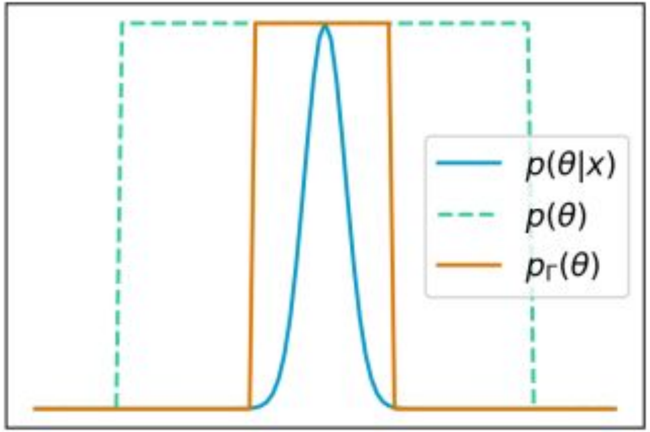
$$p(\mathbf{x}|z_1) = \int dz_2 \ldots dz_N \; p(\mathbf{x}|\mathbf{z})\tilde{p}(z_2, \ldots, z_N)$$

[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]



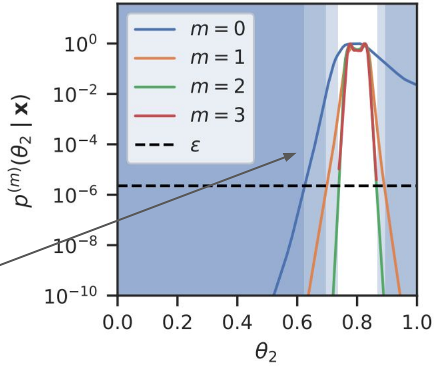
3) Likelihood-to-evidence ratios rather than densities

- Ratio estimation = Binary classification = 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]



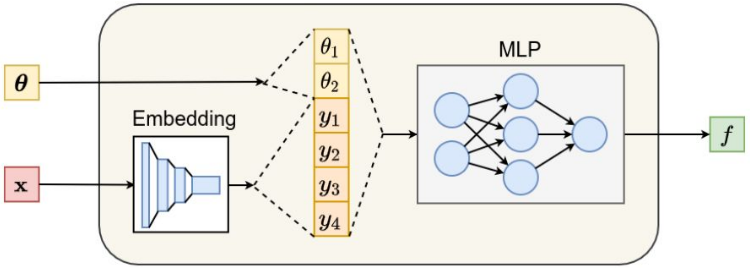
- Usually remains conservative (works well in a truncation scheme)

[but see Hermans+ 2110.06581]

- Ratio estimation automatically generates information maximizing data compression

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

[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).

What TMNRE is not

- TMNRE is **not based on flow-based architectures**, and does not perform density estimation
- TMNRE does **not require pre-optimized summary statistics**,but produces them on-the-fly
- TMNRE **does not require differentiable simulators***
- TMNRE **does not rely on approximations on the form of posteriors**

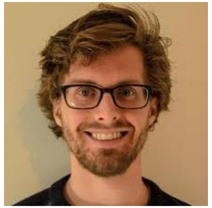
Talk about TMNRE by Ben Miller tomorrow, Wednesday, 3:00 pm

*Initially, we spend A LOT of time trying to exploit gradient-based optimization and differentiable simulators for our applications (strong lensing - we wrote a fully differentiable simulator). However, this turned out to be not fruitful (and quite painful) in numerous ways. Your mileage may vary. Currently we stay away from gradients, but they might come back at some point. Ask me if you are interested in a detailed discussion.

[Chianese+ 1910.06157; Karchev+ 2105.09465; Coogan+ 2010.07032]

Coordinated effort to develop and exploit TMNRE

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

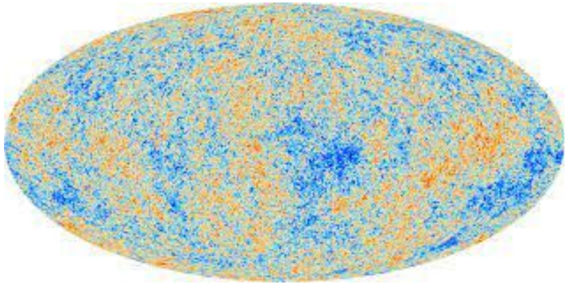
Uddipta Bhardwaj (UvA) Gravitational waves

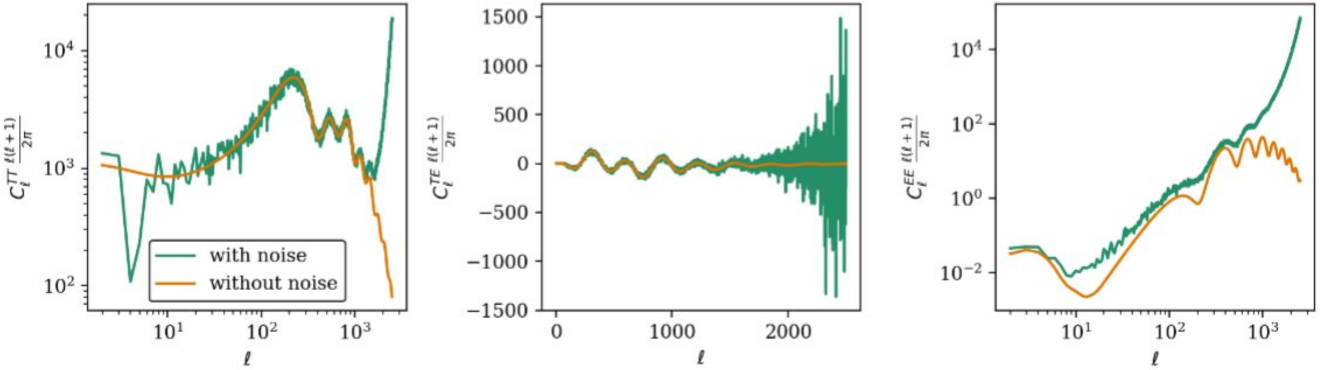
James Alvey (UvA), Stellar streams, GWs

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

Example 1: Cosmic microwave background

CMB forecasting



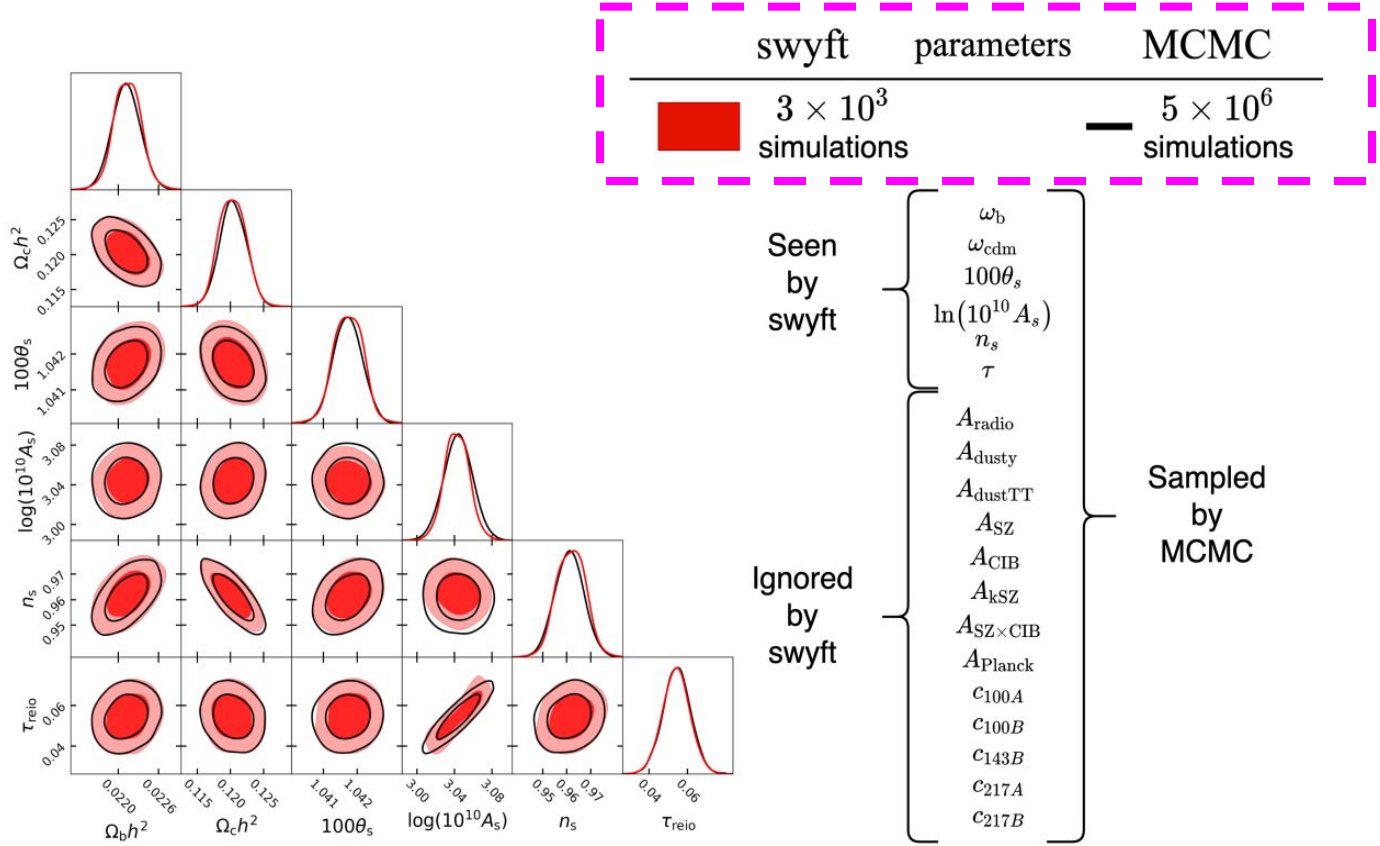


Noise = instrument contribution + cosmic variance

- 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

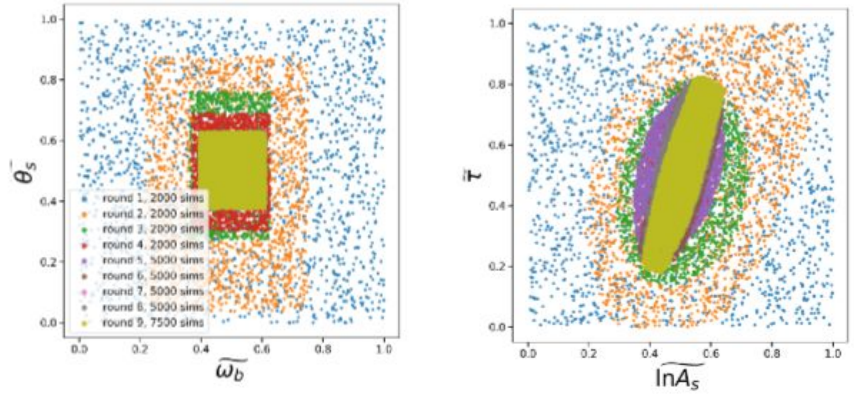
[Cole, Miller, Witte, Cai, Grootes, Nattino, CW 2111.08030]

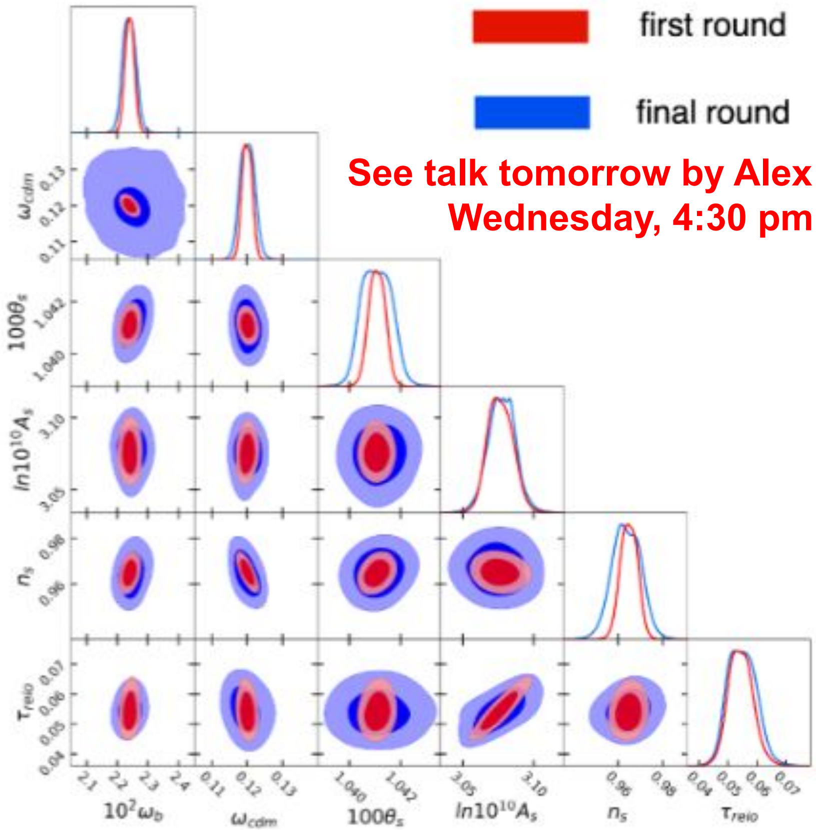
Realistic CMB



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

Importance of truncation





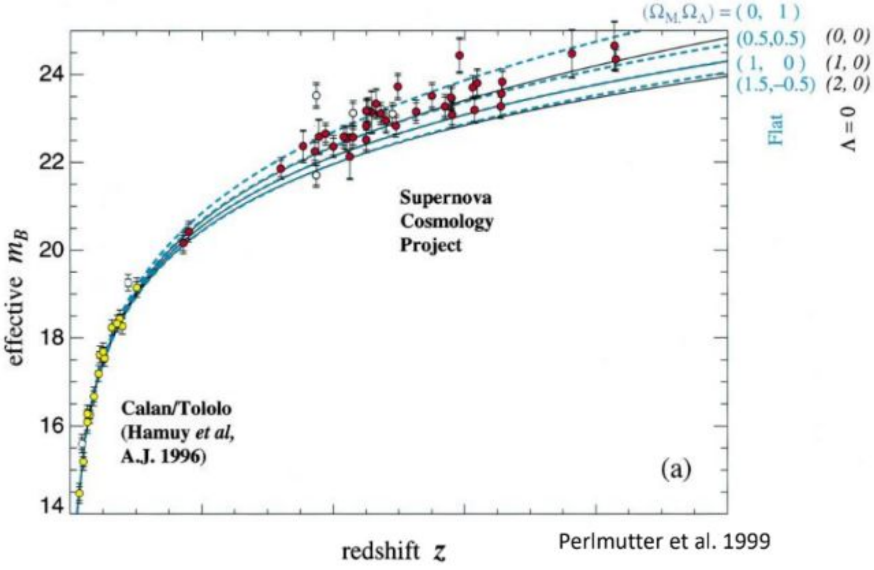
See talk tomorrow by Alex Cole, Wednesday, 4:30 pm

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

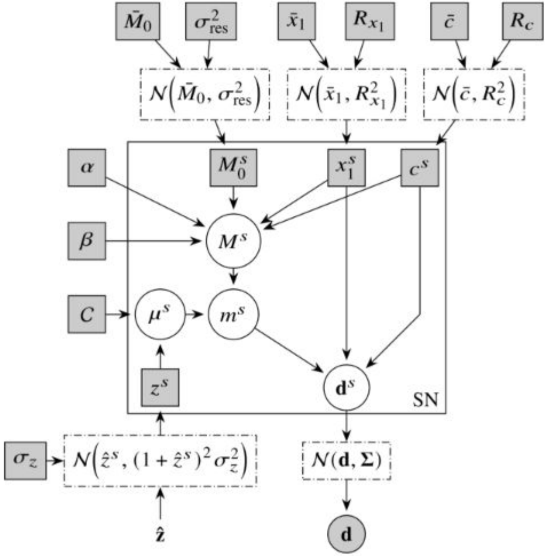
Structure of ratio estimator
- Input: Vector (7500)
- Embedding: Linear (7500 → 10)
- Marginals: MLP (19 1-dim, 15 2-dim)

Example 2: Supernova cosmology

Supernova cosmology

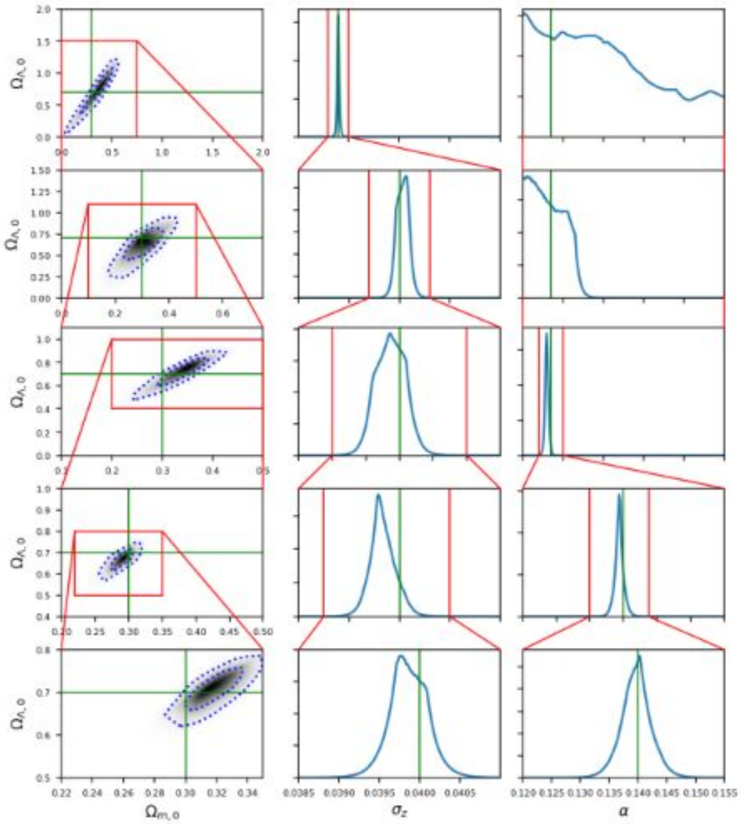


$$m = M + \mu(z, \mathcal{C}) + \text{"noise"}$$



Ongoing work with Kosio Karchev and Roberto Trotta

Truncated marginal NRE



stage 0

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

stage 1

...

stage 2

stage 3

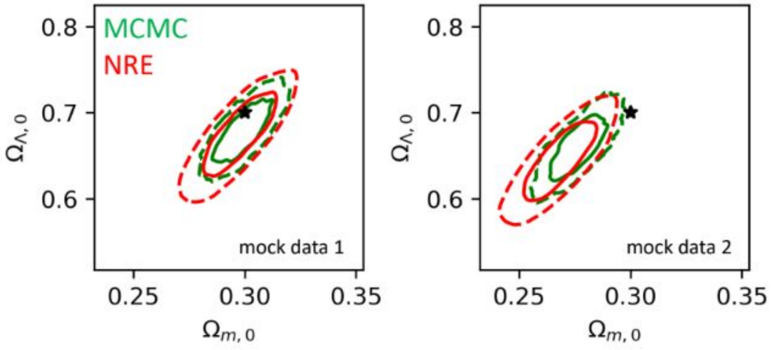
final stage

progressively home in on true values

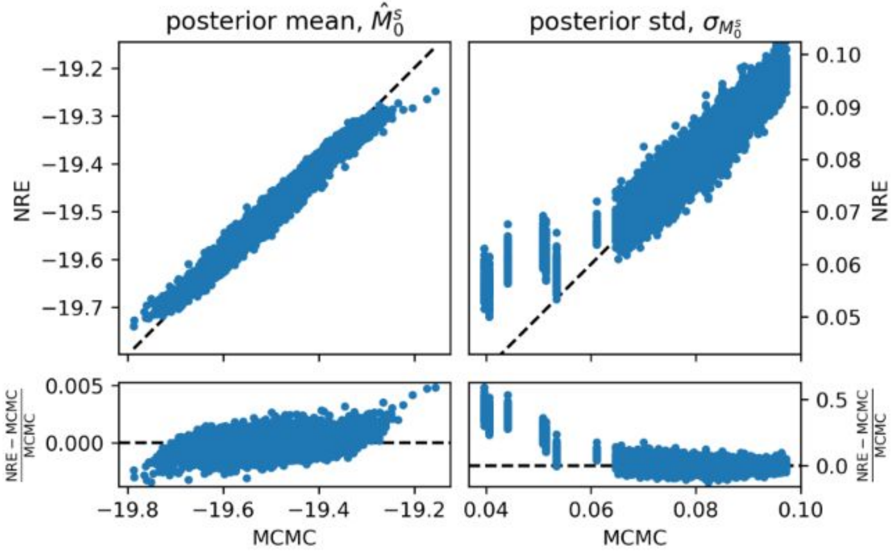
Ongoing work with Kosio Karchev and Roberto Trotta

Marginal posteriors

100 000 supernovae



MALFOI: marginal likelihood-free object-by object inference



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

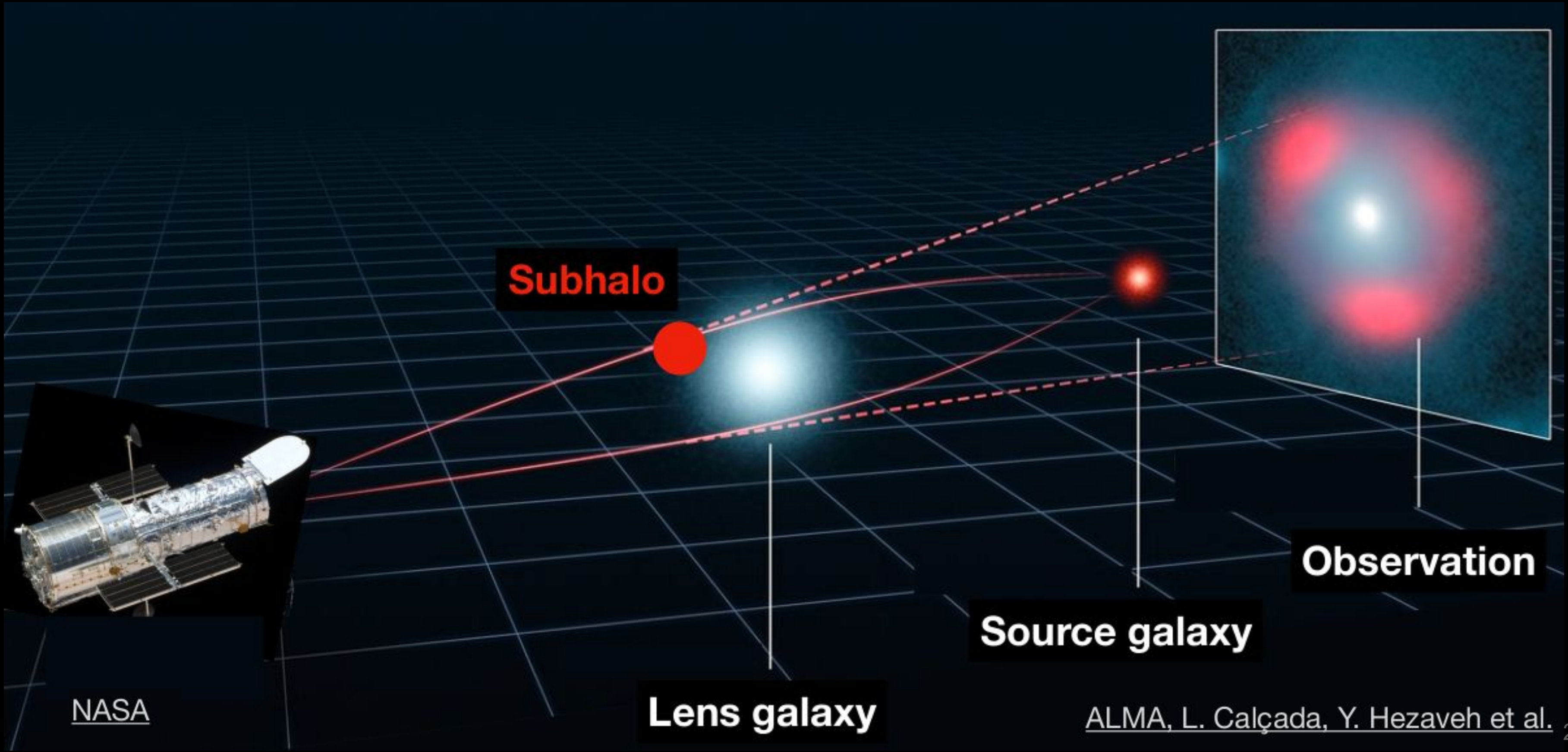
Structure of ratio estimator

- 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

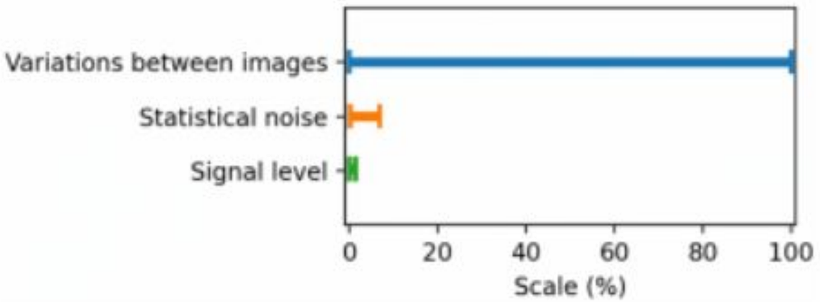
Example 3: Strong lensing

Strong galaxy-galaxy lensing



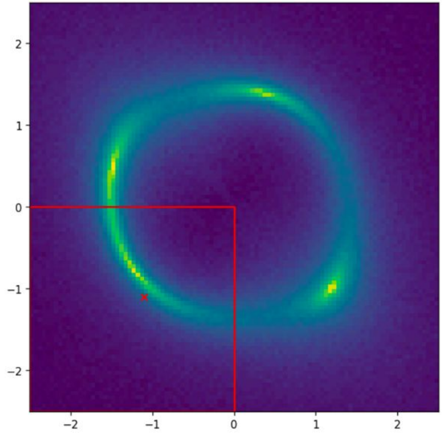
Inference challenges

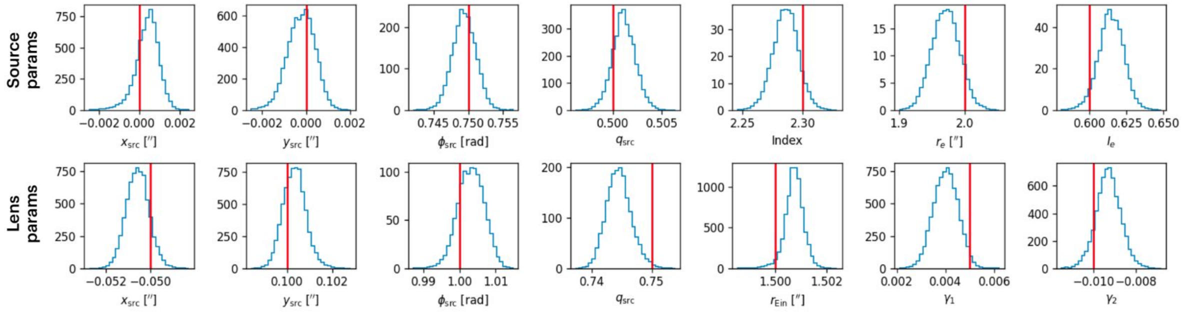
- Signal is small compared to noise and variations between images
- Marginalization over numerous source, lens and halo parameters
- Joint posterior has $\sim N_{sub}!$ modes; likelihood can be intractable

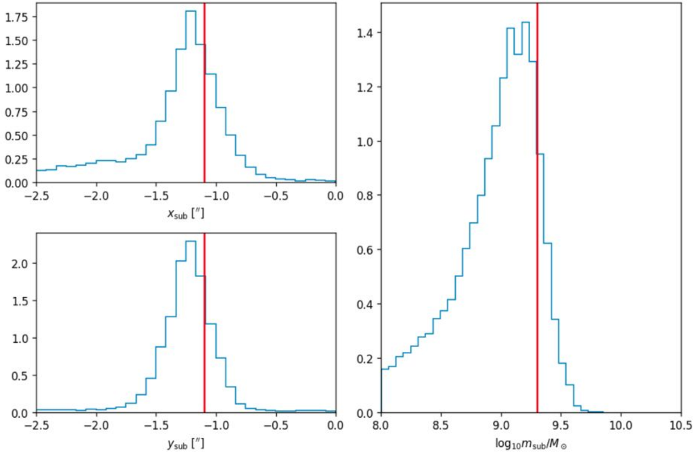


Slide credit: Noemi Anau Montel

Single subhalo, simple source model







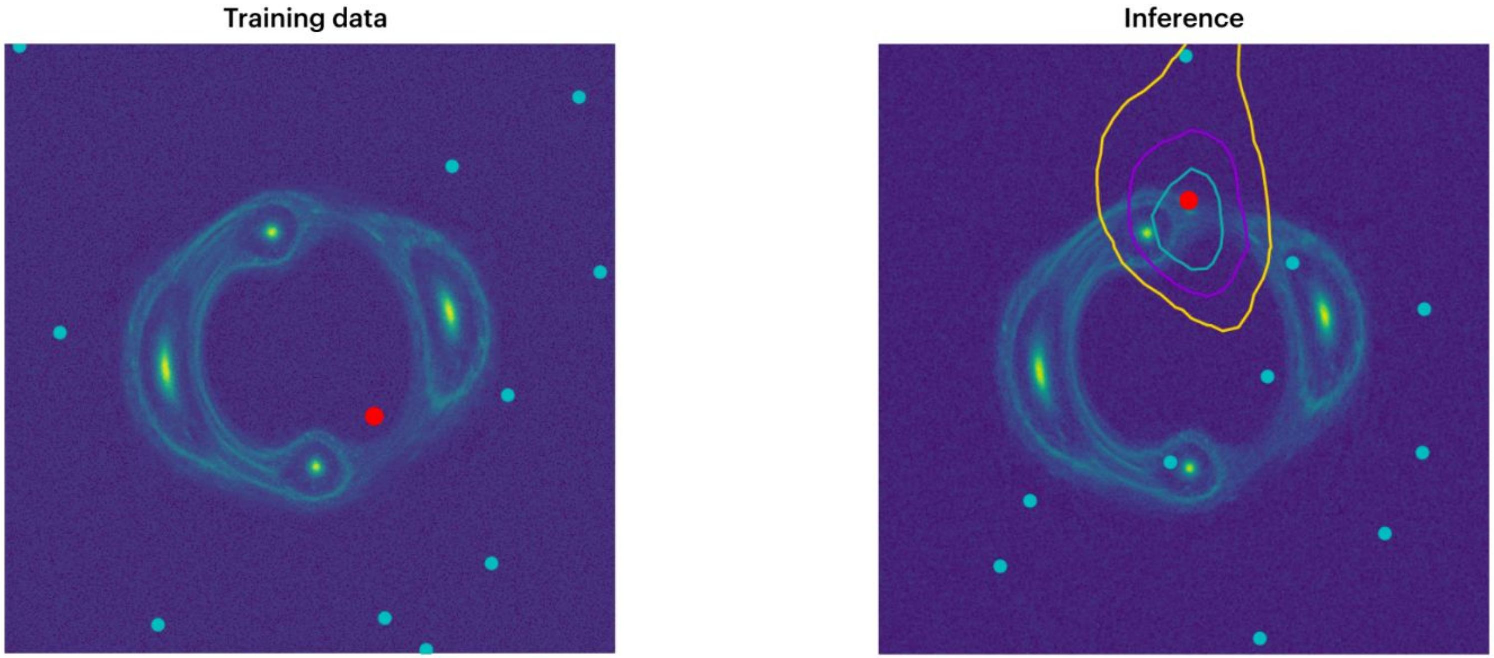
Structure of ratio estimator

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

Ongoing work led by Adam Coogan

Slide credit: Noemi Anau Montel

Multiple subhalos, complex source model



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

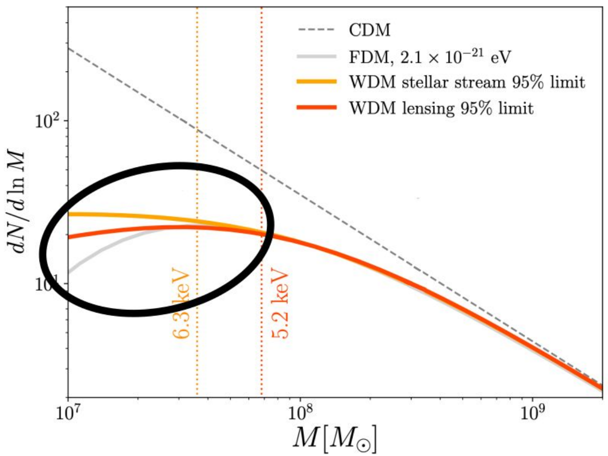
Marginalized over source, lens and halo population

Ongoing work led by Adam Coogan

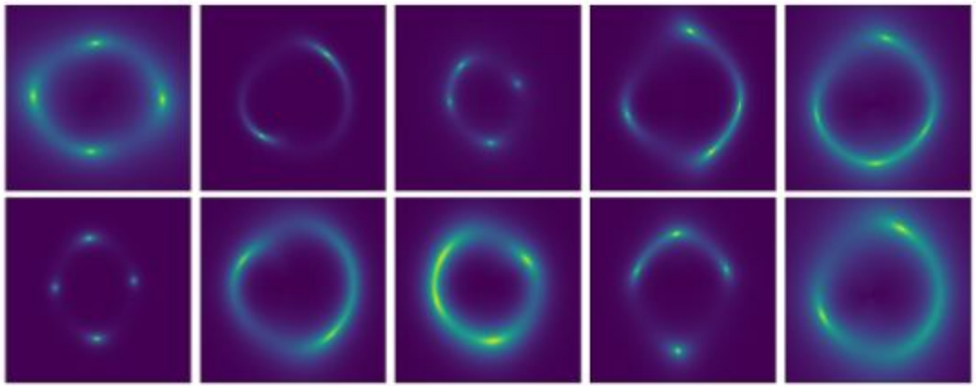
Structure of ratio estimator

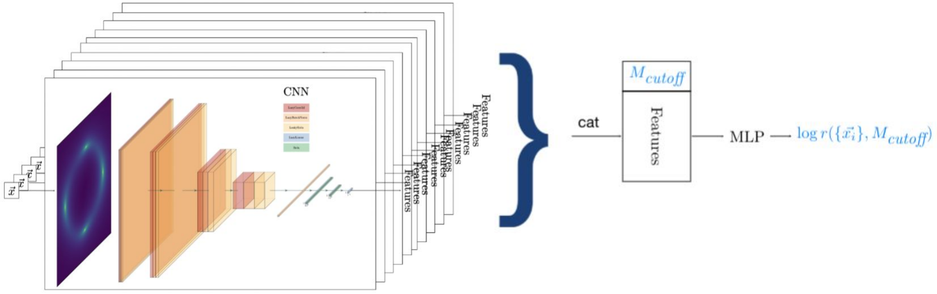
- Input: Images (typically 200x200)
- Embedding: CNN
- Marginals: MLP (2-dim)

Infer subhalo mass function cutoff $p(M_{\mathrm{cutoff}} \mid \{\vec{x}_i\}_{i=1,\ldots,10})$



Combining observations to reduce subhalo shot noise





Structure of ratio estimator

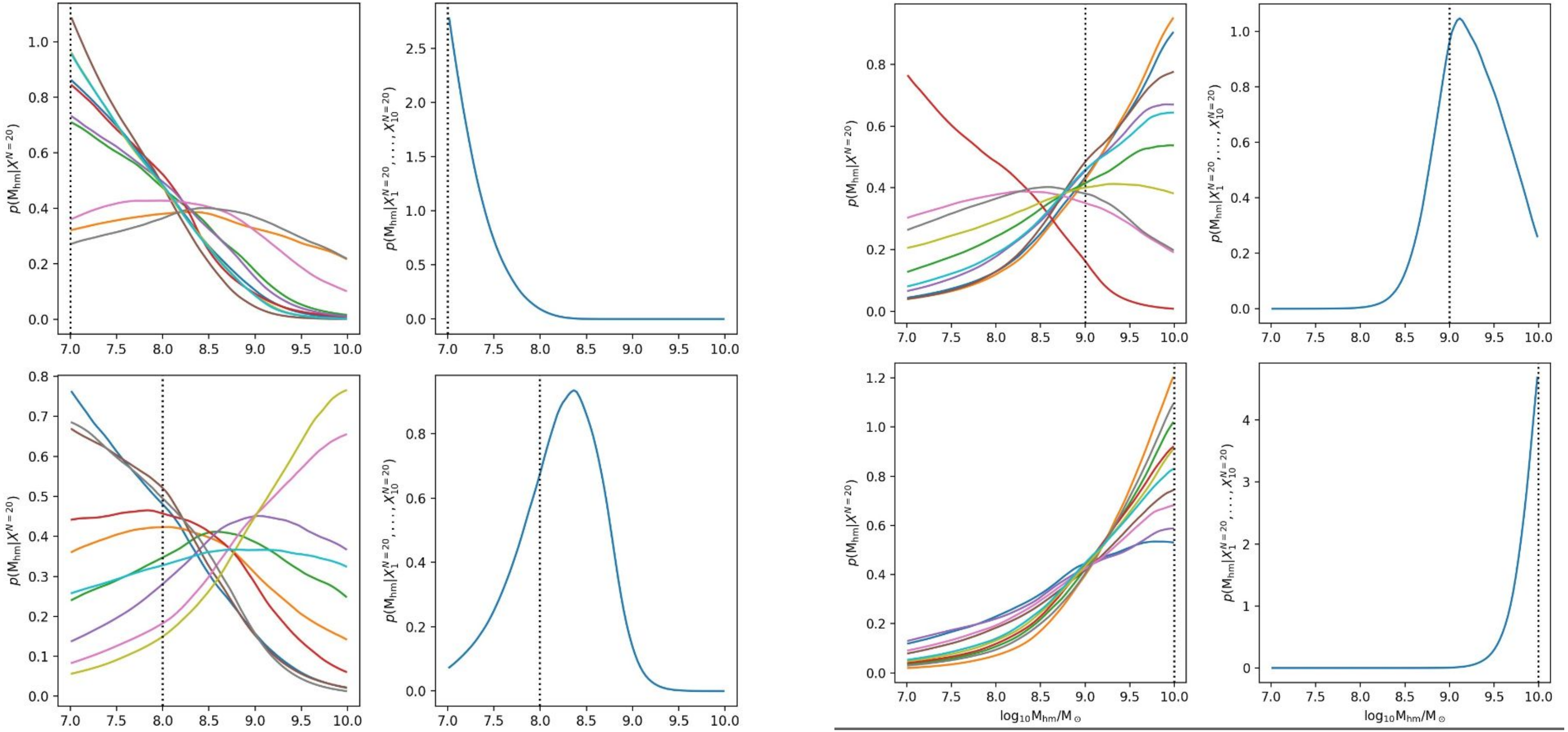
- Input: 10 Images (10x100x100)
- Embedding: Stack of CNNs
- Marginals: MLP (1-dim)

Ongoing work led by Noemi Anau Montel

Infer subhalo mass function cutoff $p(M_{\text{cutoff}} | \{\vec{x}_i\}_{i=1,\ldots,10})$

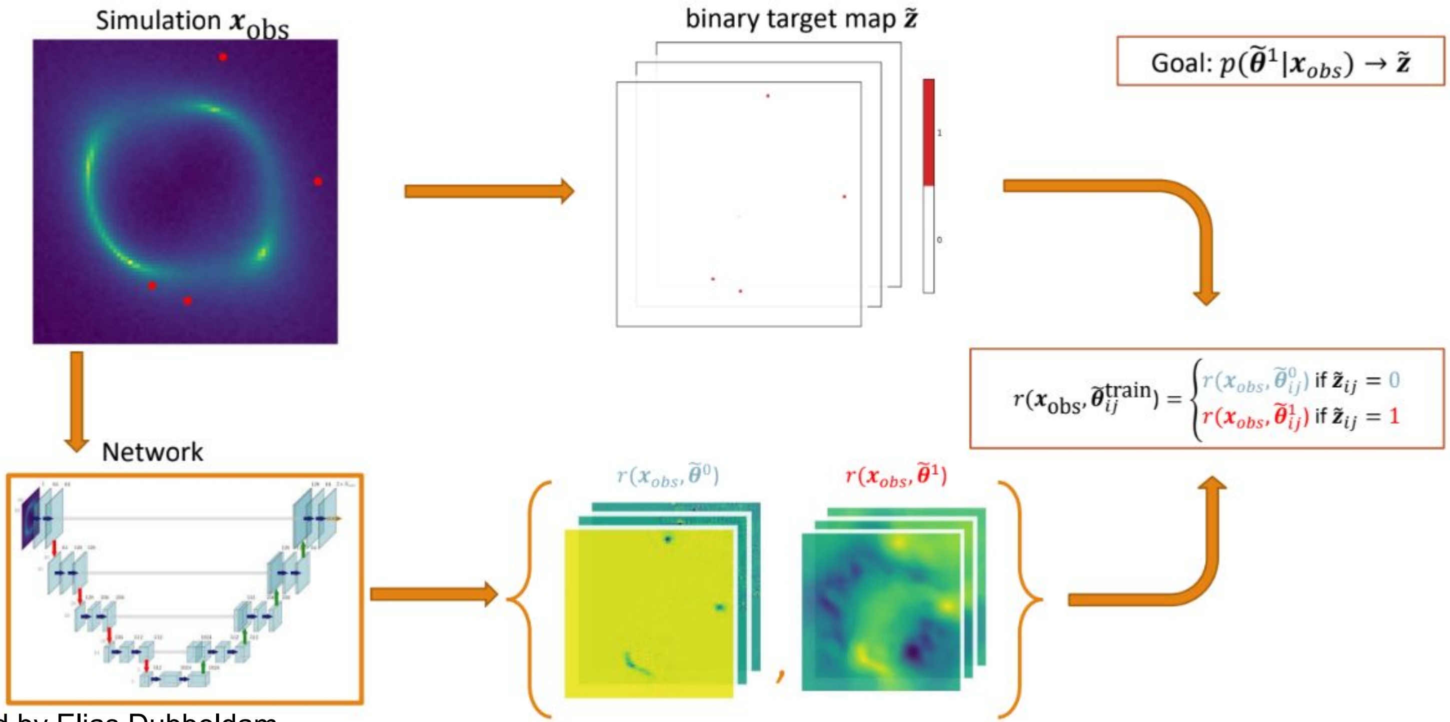
Ongoing work led by Noemi Anau Montel

Combining 100 images (10x10x100x100 images) should lead to tight posteriors.



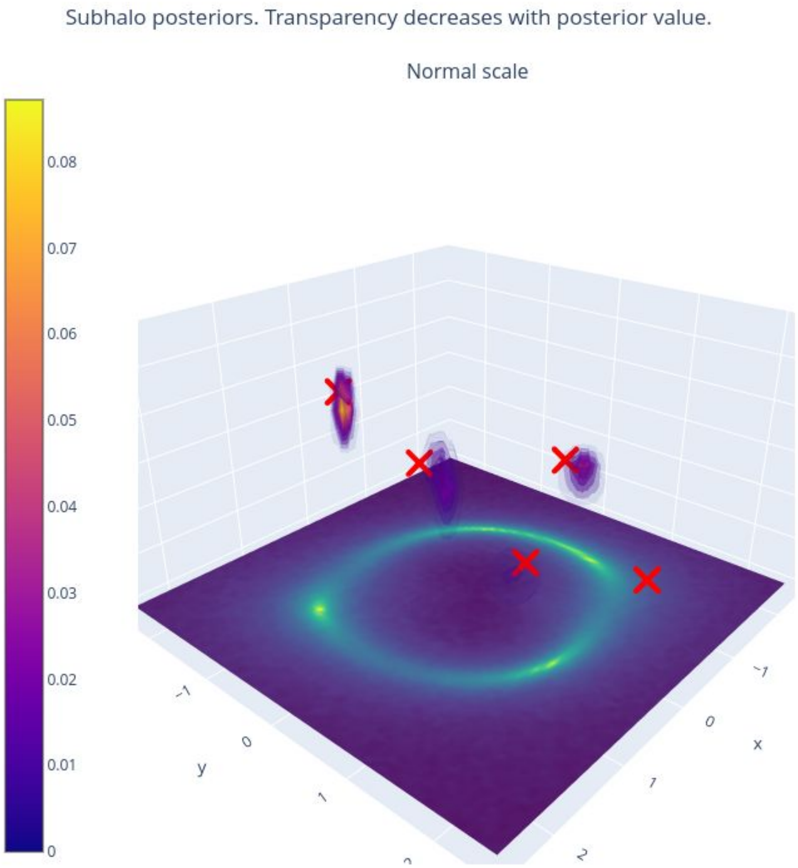
Probabilistic image segmentation

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

Probabilistic image segmentation



Structure of ratio estimator

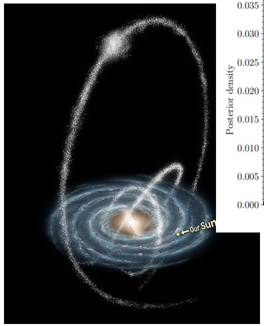
- Input: Image (typically 100x100)
- Embedding: U-Net
- Marginals: Binary marginals 100x100x10 (ten mass bins)

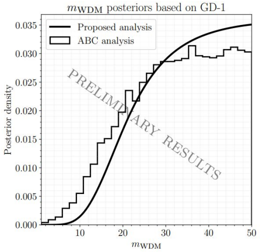
https://dm-lensing-parislfi.github.io/

Ongoing work led by Elias Dubbeldam

TMNRE/SWYFT appear to be broadly applicable

Stellar streams

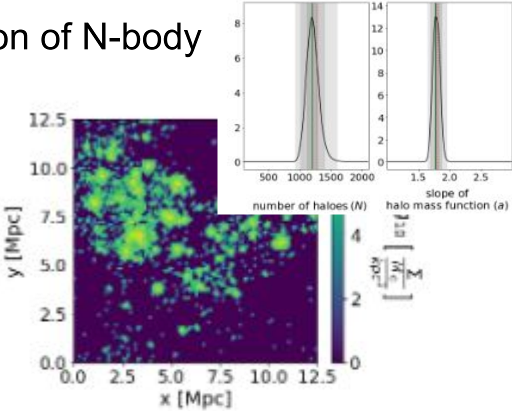




Hermans et al., 2020
James Alvey, Mathis Gerdes, in progress

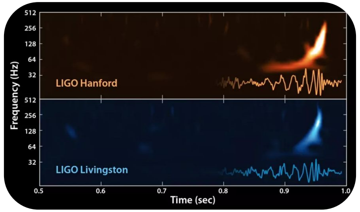
TMNRE/SWYFT

Interpretation of N-body simulations



Androniki Dimitriou+, soon

Gravitational waves



Delaunoy+ 2020
Uddipta Bhardwaj+, in progress

21 cm cosmology
LHC pheno fits
…

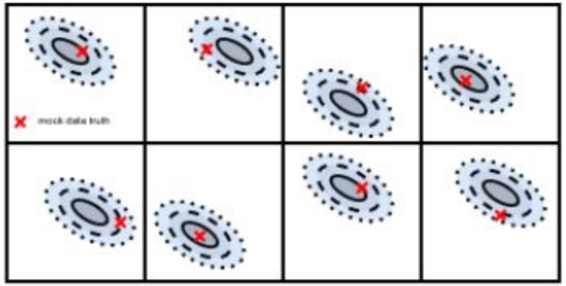
How can one trust results?

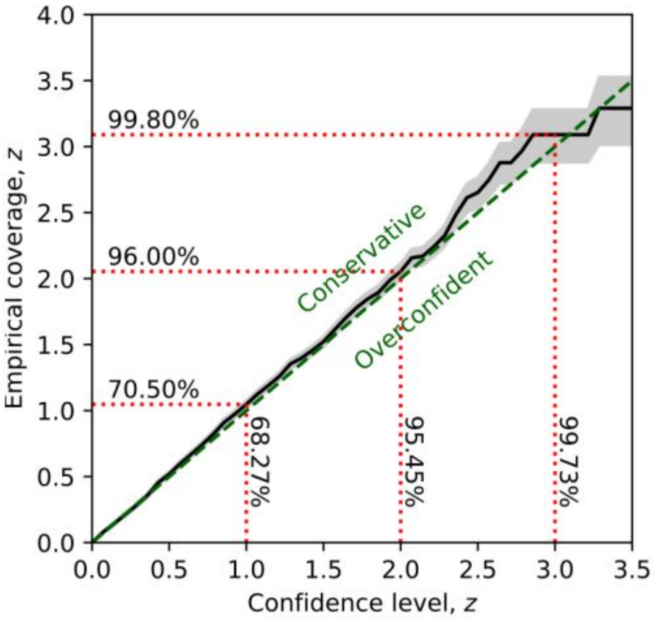
Credibility of inference results can be tested

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

Expected coverage of the 68% and 95%

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

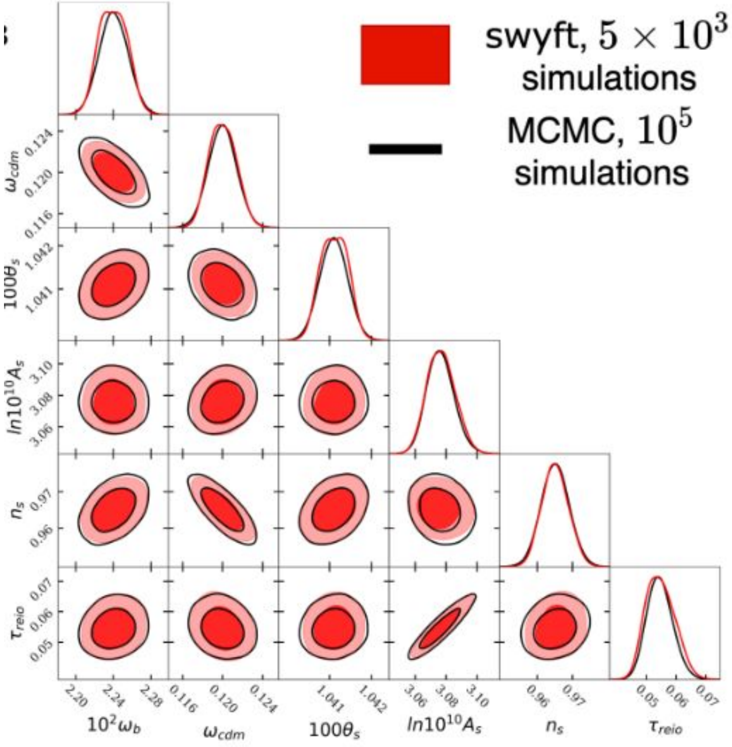




[Cole, Miller, Witte, Cai, Grootes, Nattino, CW 2111.08030]

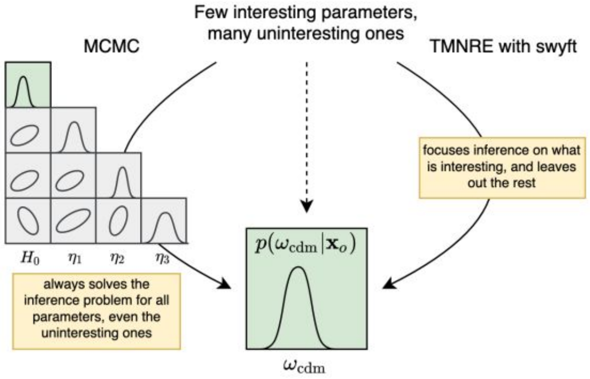
See also Hermans, Delaunoy, Rozet, Wehenkel, Louppe 2110.06581

Coverage tests!

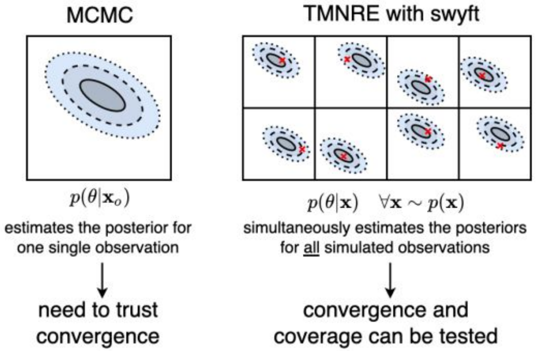


Open source package SWYFT

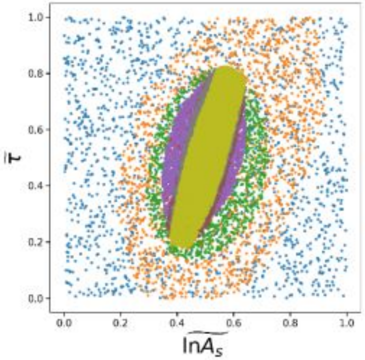
Estimating marginals of interest



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**.
- I 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**
- I 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.

Thank you!

Backup

Example for truncation scheme