Reconstruction of cosmological initial conditions with sequential simulation-based inference

Oleg Savchenko

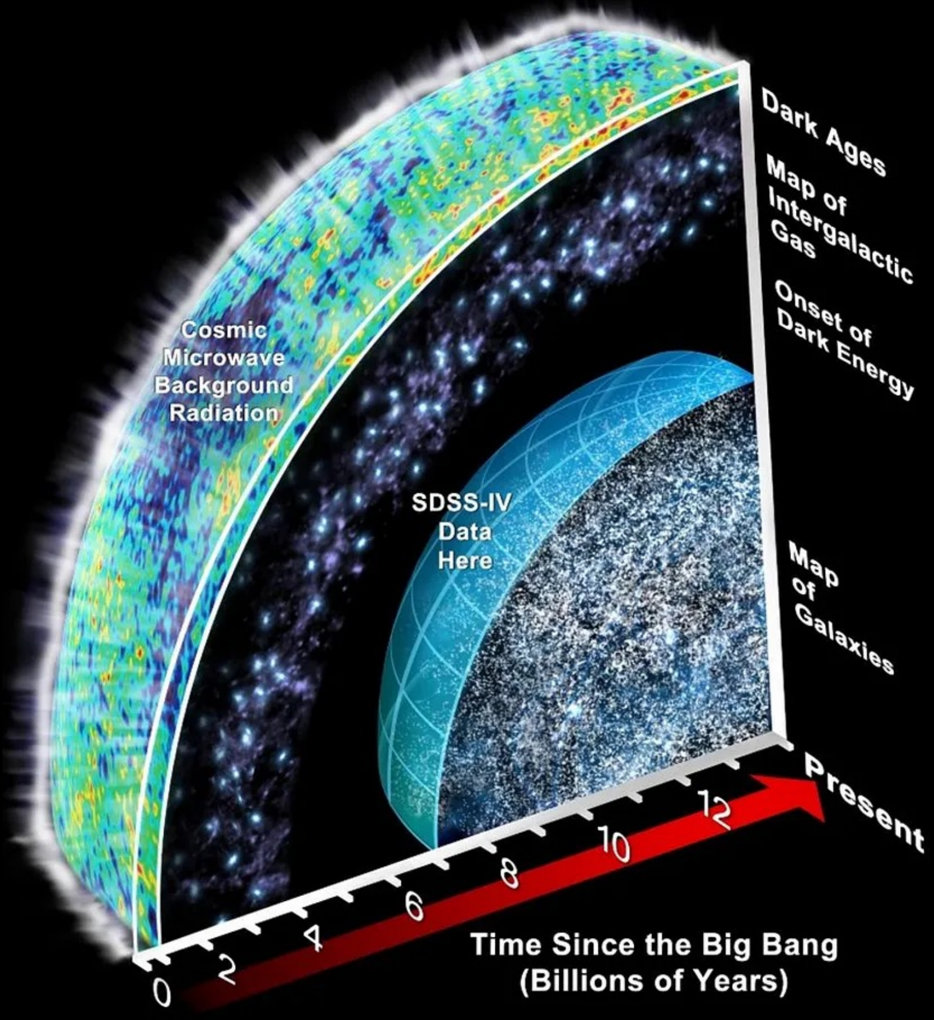
Work with: Guillermo Franco Abellán, Florian List, Noemi Anau Montel, Christoph Weniger

Universiteit van Amsterdam

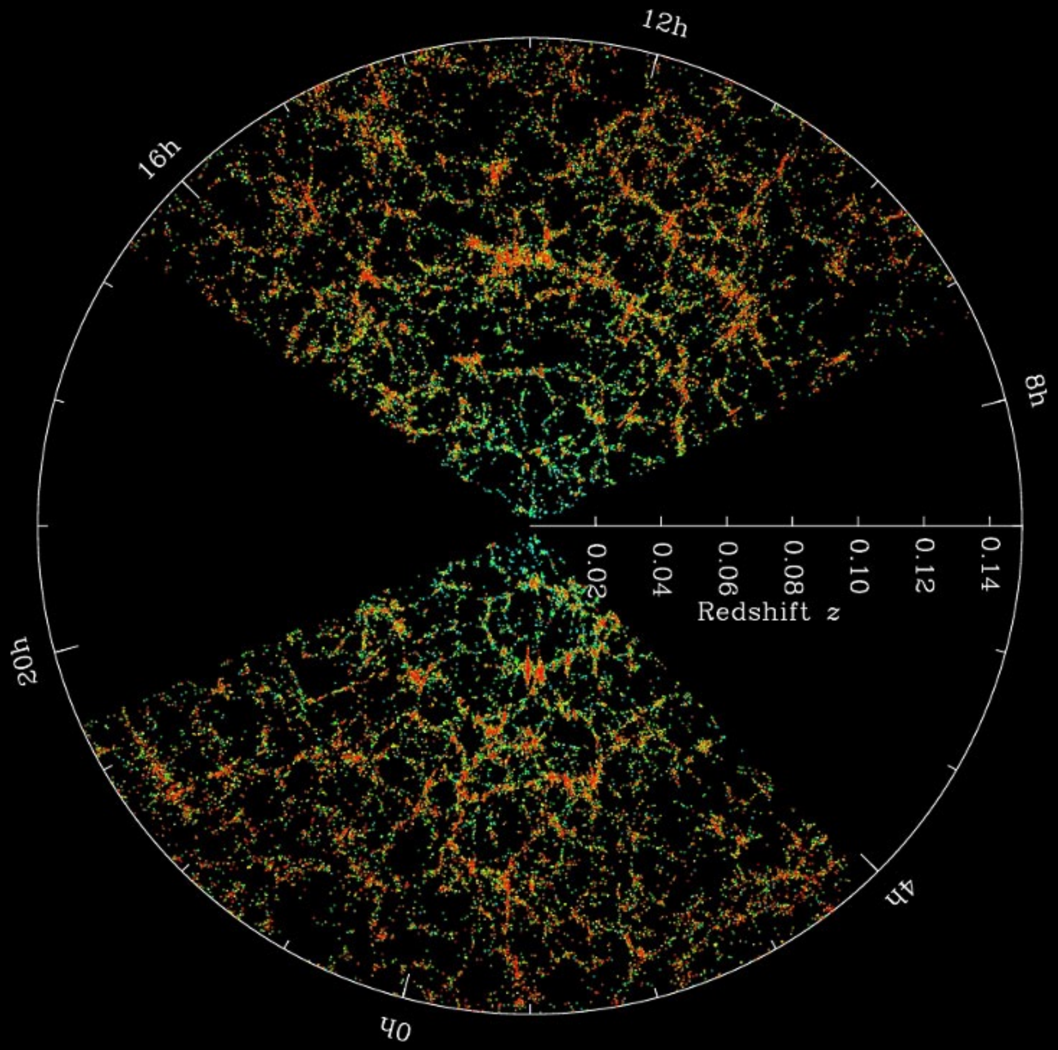
GRAPPA
GRavitation AstroParticle Physics Amsterdam

ICHEP 2024

Large scale structure

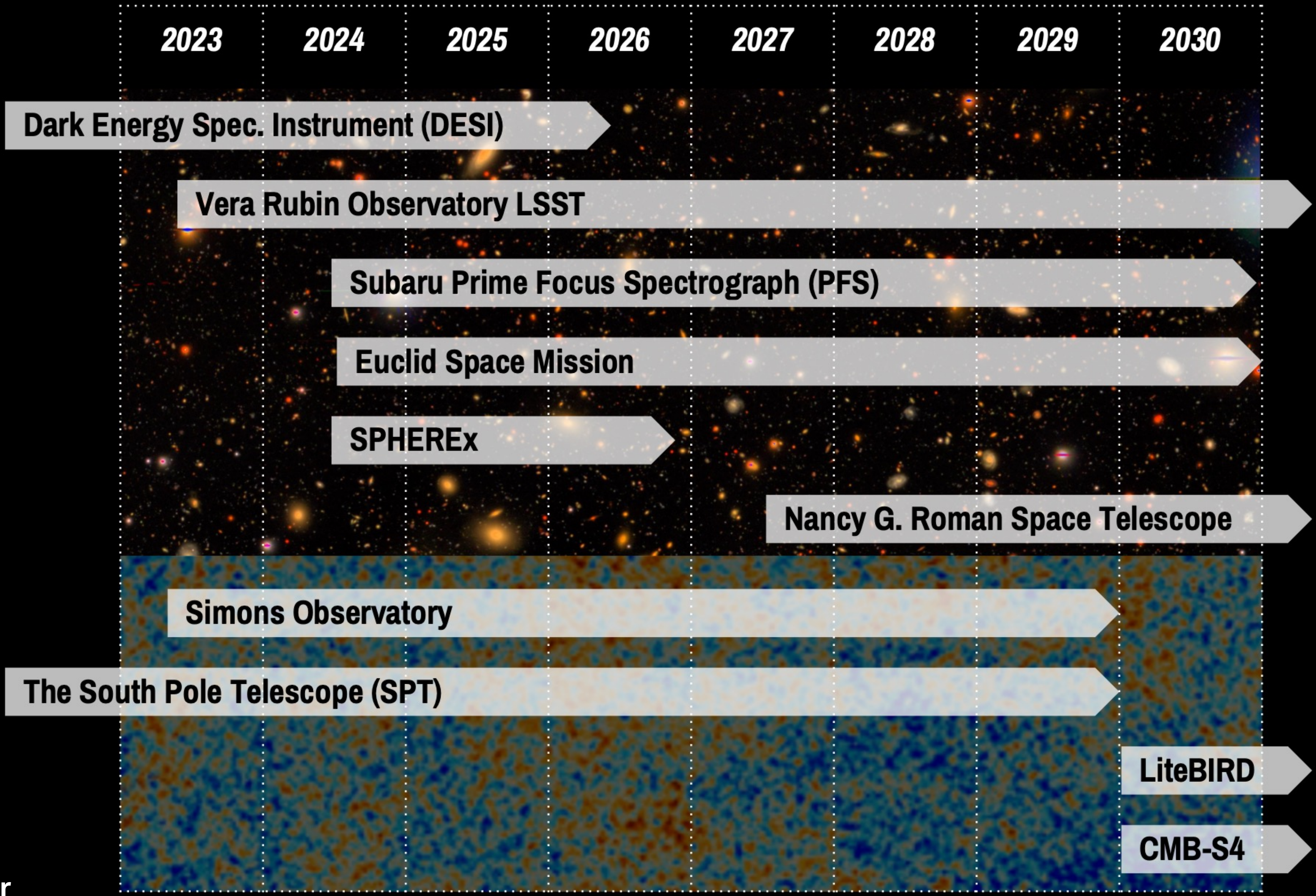


https://mapoftheuniverse.net/



SDSS collaboration

Next decade

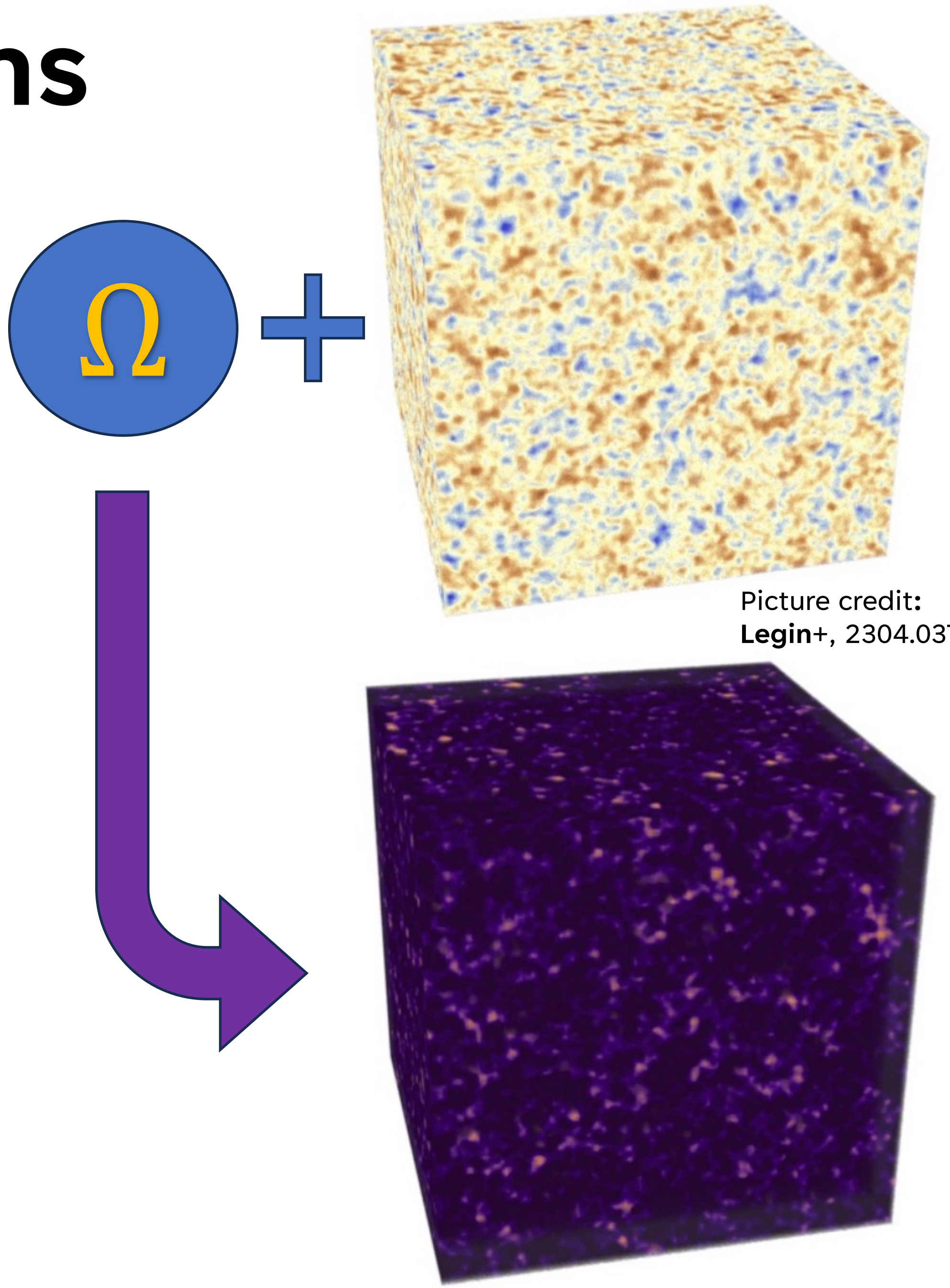


Picture credit: **A. Bayer**

Cosmological simulations

Types:

- LPT
- COLA
- Particle mesh
- N-body
- Hydrodynamical



Picture credit: **Legin**+, 2304.03788

https://quijote-simulations.readthedocs.io/

https://camels.readthedocs.io/

How to analyse LSS data?

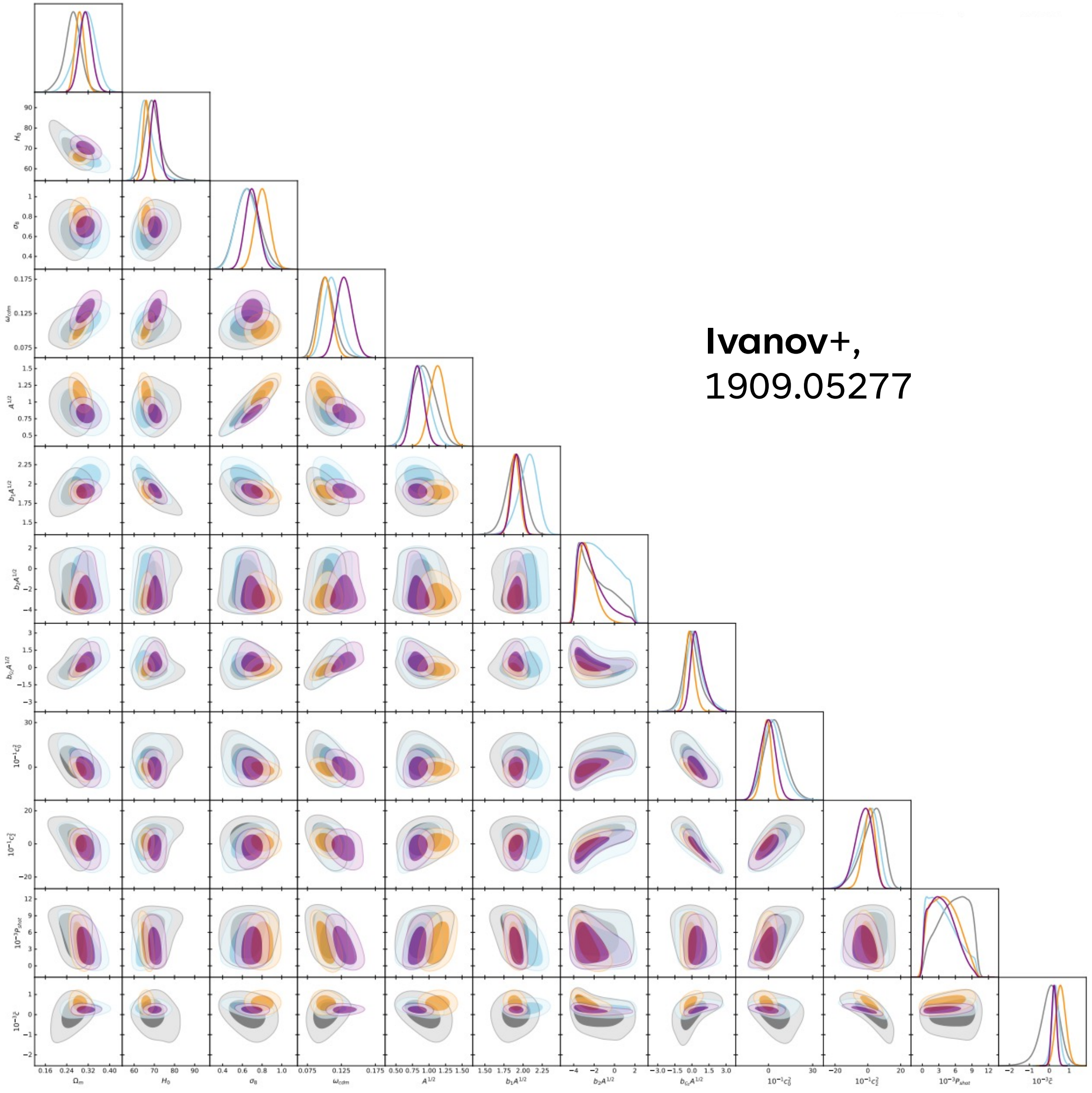


Figure 11: The triangle plot for cosmological and nuisance parameters of four independent BOSS datasets.

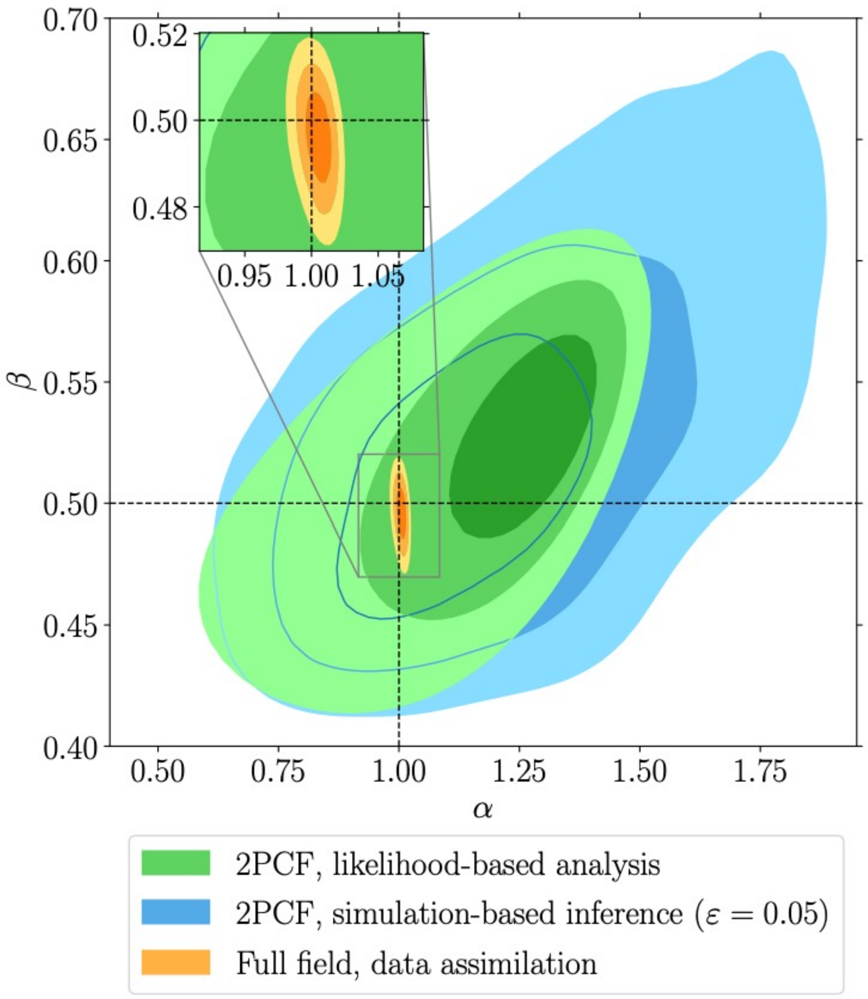
Ivanov+,
1909.05277

Classic approach:

- Come up with some summary statistic ***s****(data)*
- Develop theory predictions for ***s***
- Construct an analytic likelihood model (usually Gaussian)
- Use Bayes theorem and run MCMC:

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})}{p(\boldsymbol{x})} p(\boldsymbol{z}).$$

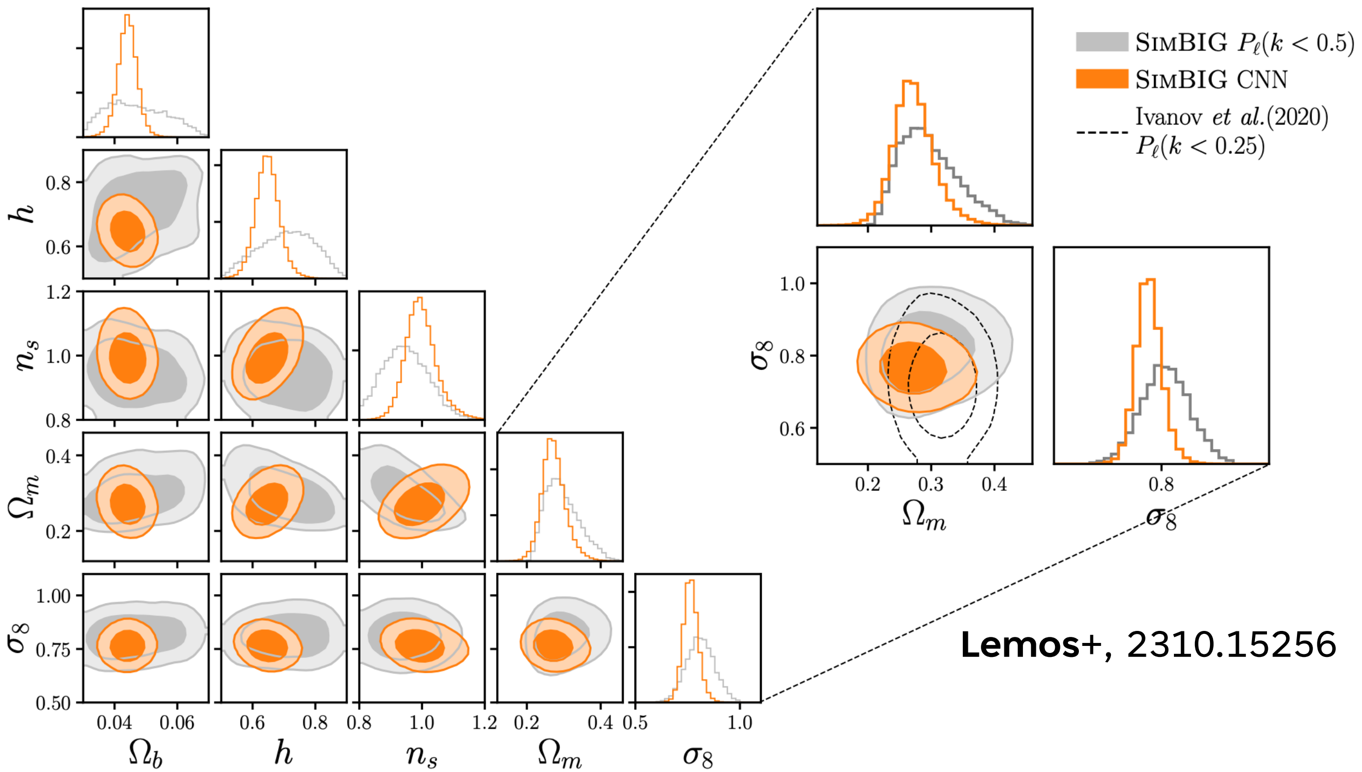
Field-level inference



Leclercq+, 2103.04158

The whole field contains much more information than some summary like a power spectrum!

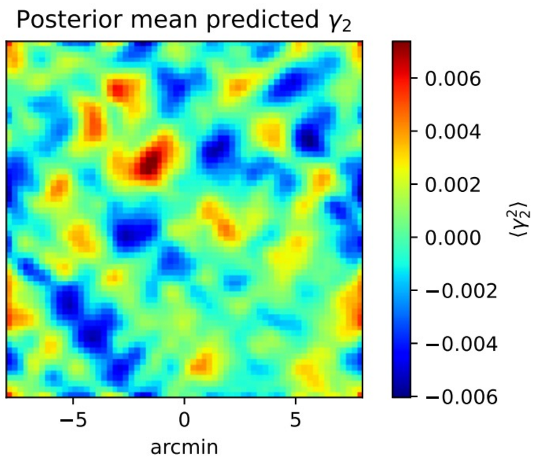
$$\langle \delta_{\rm m}(\mathbf{k})\delta_{\rm m}(\mathbf{k}')\rangle = (2\pi)^3\delta_{\rm D}^3(\mathbf{k}+\mathbf{k}')P_{\rm mm}(k)$$



Lemos+, 2310.15256

Reconstructing fields

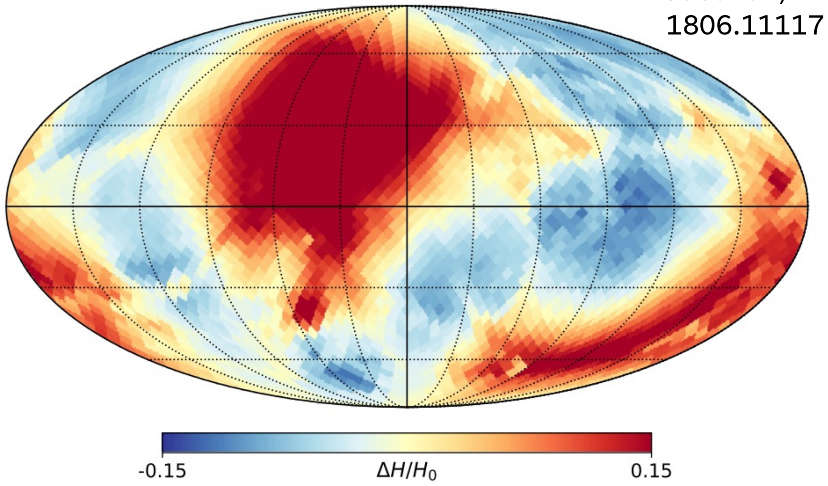
Porqueres+,
2304.04785



Shear

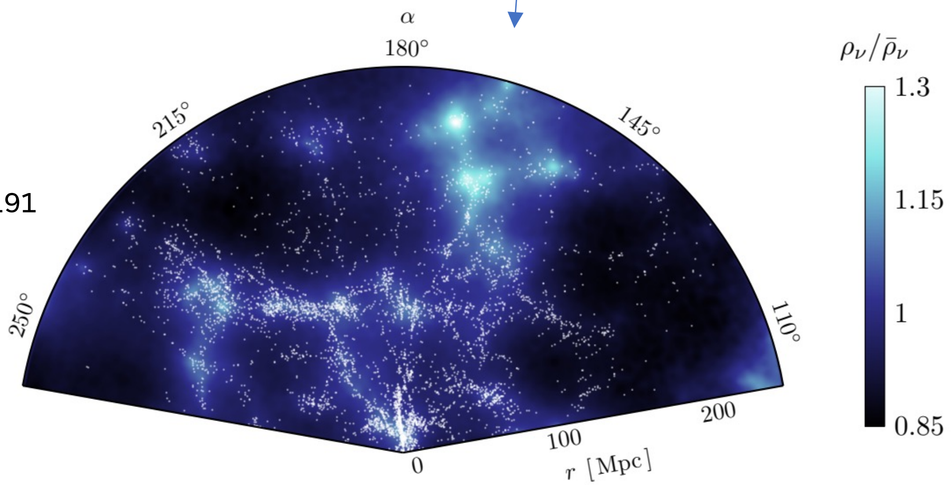
Mass estimation and evolution history, velocity fields, BAO, f_{NL} , 21 cm, weak lensing...

Jasche+,
1806.11117



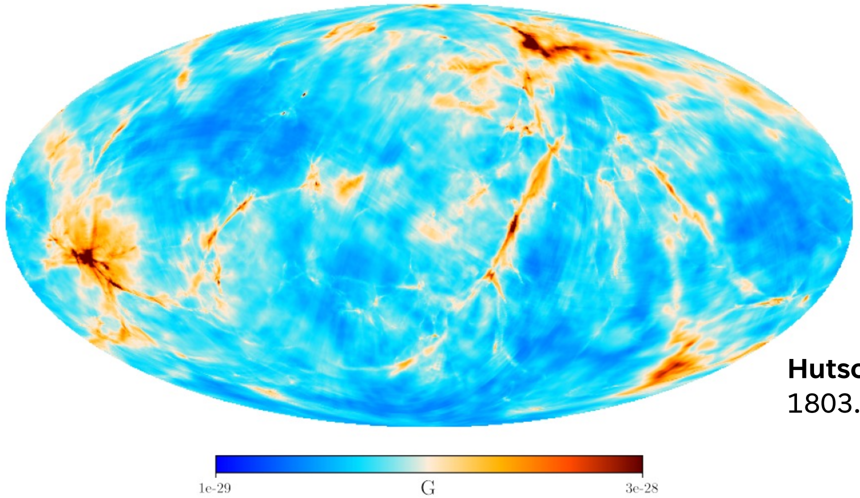
Hubble parameter uncertainties

Elbers+,
2307.03191



Neutrino field

Hutschenreuter+,
1803.02629



Magnetic field

Initial conditions reconstruction

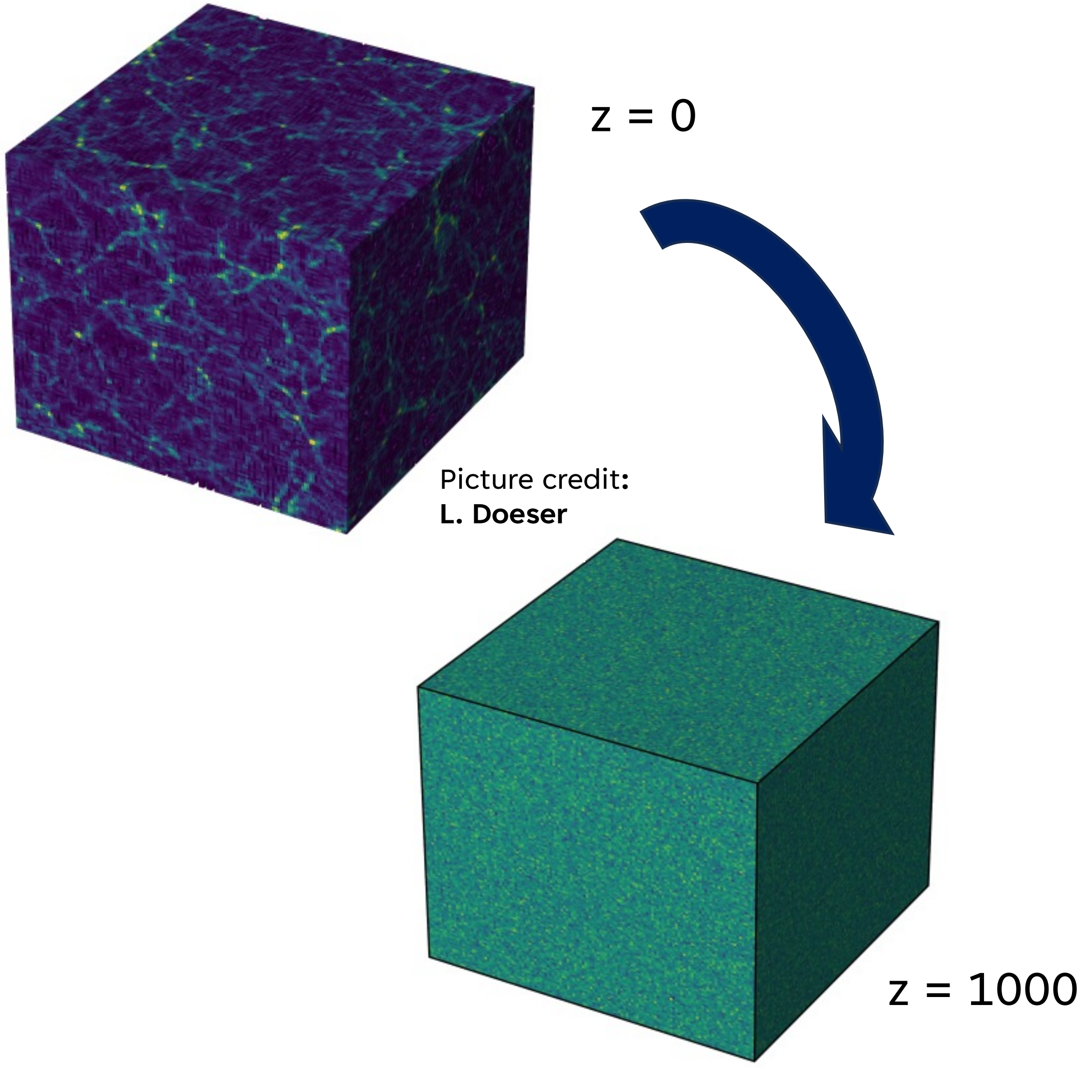
Very early universe had very simple properties!

$\longrightarrow$ feasible way to do field reconstruction is to infer these initial conditions 🙂

$$P(\boldsymbol{\delta}^{\mathrm{IC}} \mid \{N_i^g\}) = \frac{P(\boldsymbol{\delta}^{\mathrm{IC}})P(\{N_i^g\} \mid G_i(\boldsymbol{\delta}^{\mathrm{IC}}))}{P(\{N_i^g\})}$$

$\boldsymbol{\delta}^{\mathrm{IC}}$ - initial density field

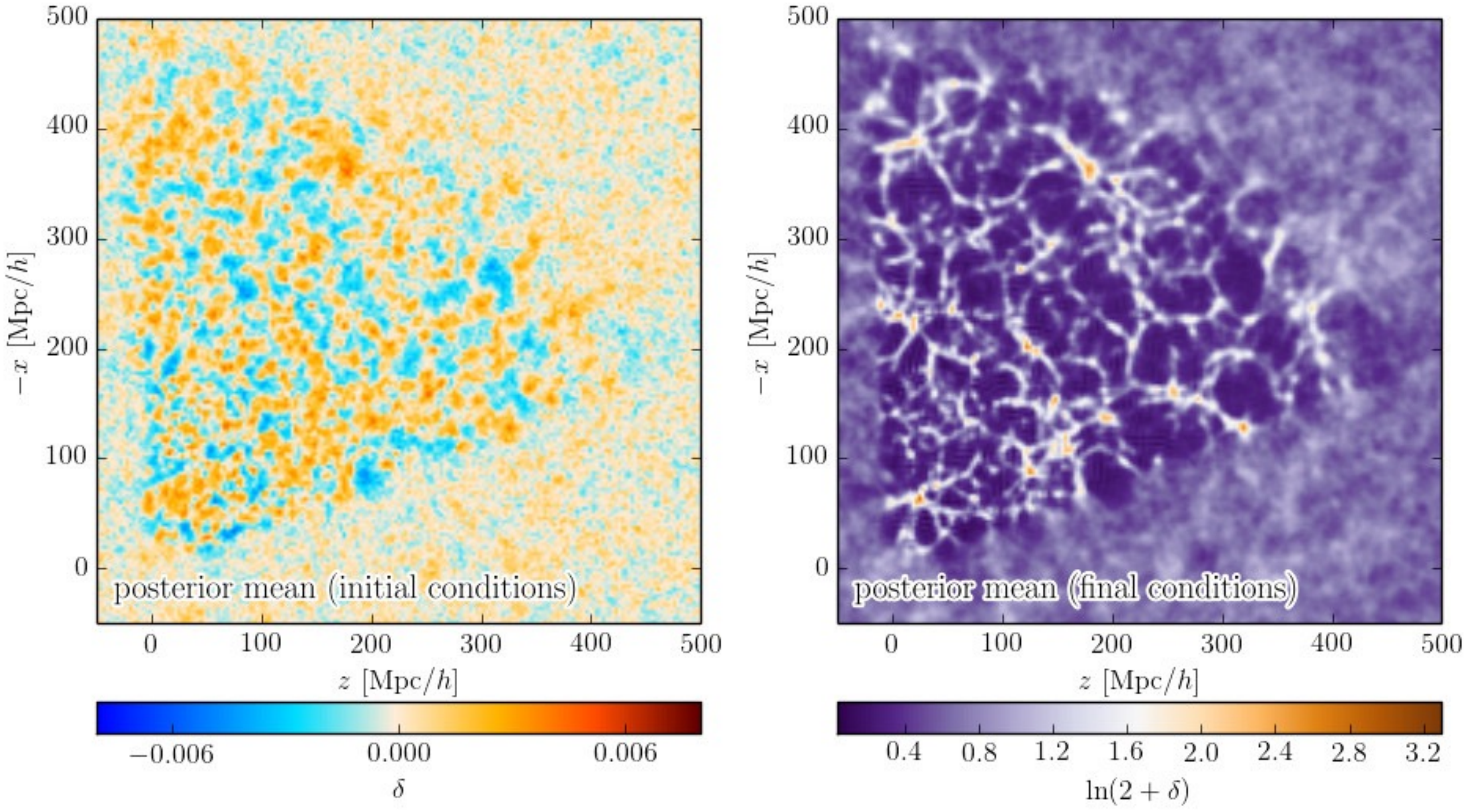
$\{N_i^g\}$ - galaxy catalog data



Gaussian initial conditions

Bayesian Origin Reconstruction from Galaxies (BORG)

Jasche+, 1203.3639, 1806.11117

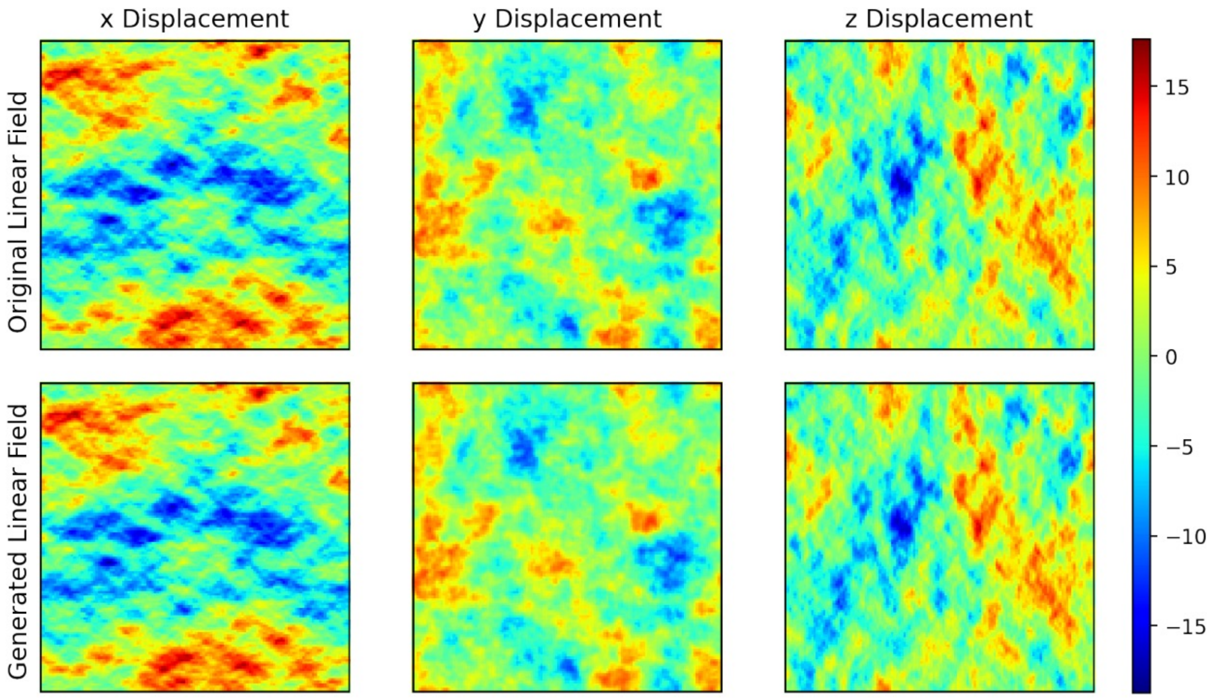


Samples produced with Hamiltonian Monte Carlo

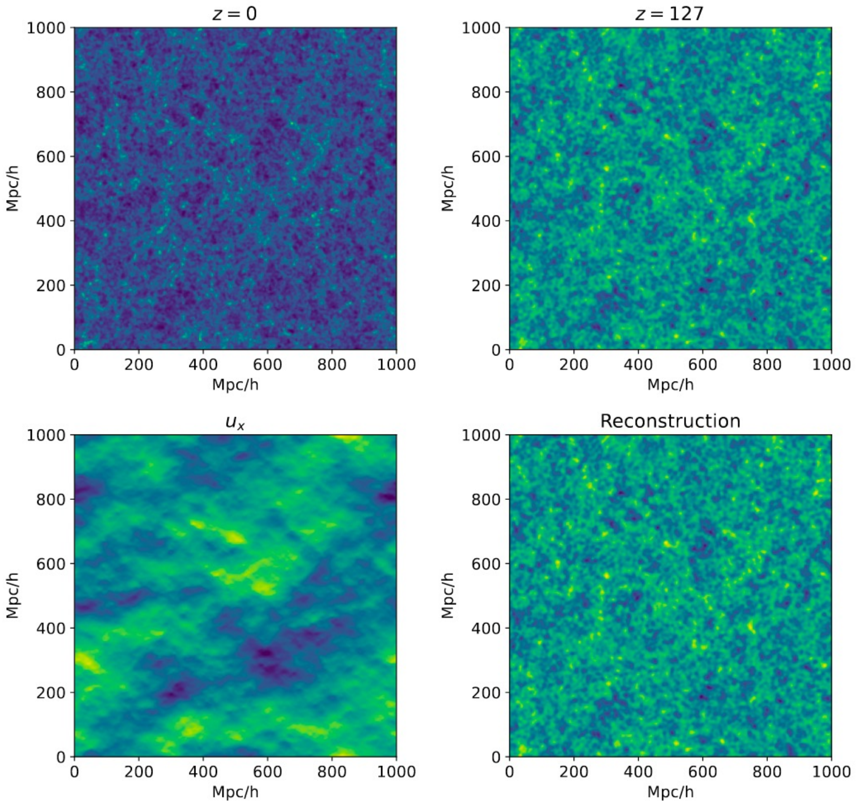
- Most developed method so far
- Explicit likelihood
- Requires gradients from the simulator
- Takes hours to produce a single sample
- Not amortized

ML approaches

- **Point estimates**: train a neural net to give single deterministic prediction (MAP estimation)



Jindal+, 2303.13056



Flöss+, 2305.07018

Diffusion models

- Train a network to approximate the **score:** $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}|\boldsymbol{y})$
- Generate samples via reverse-diffusion process.

24 hrs of training on 4 80GB NVIDIA A100 GPU's

Legin+, 2304.03788

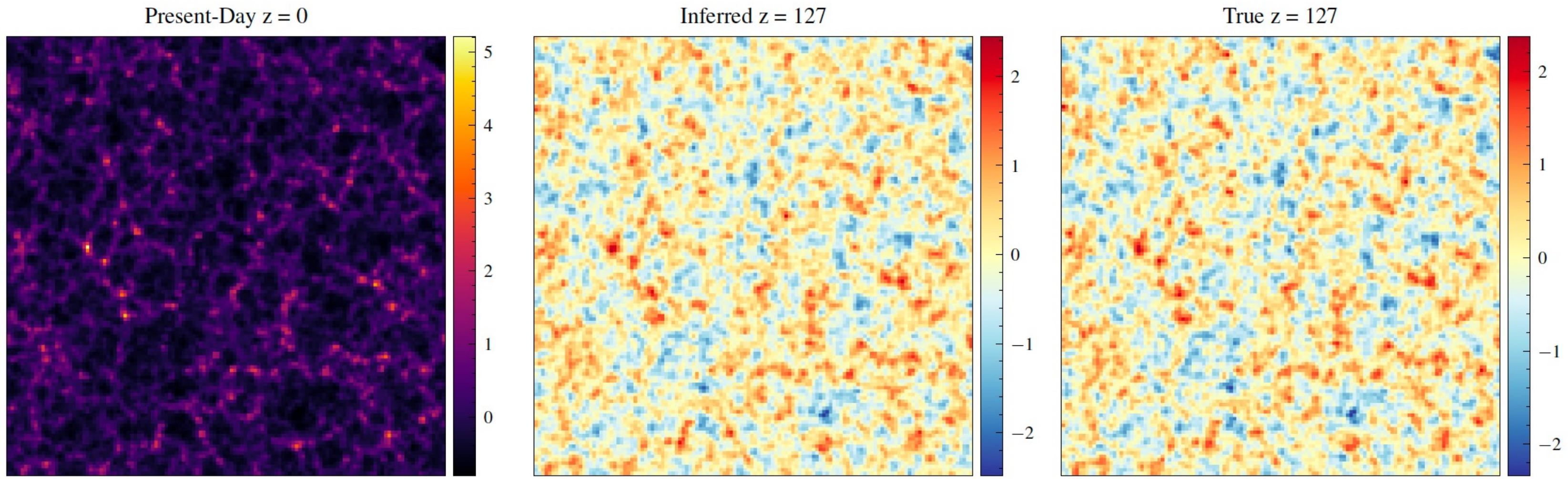
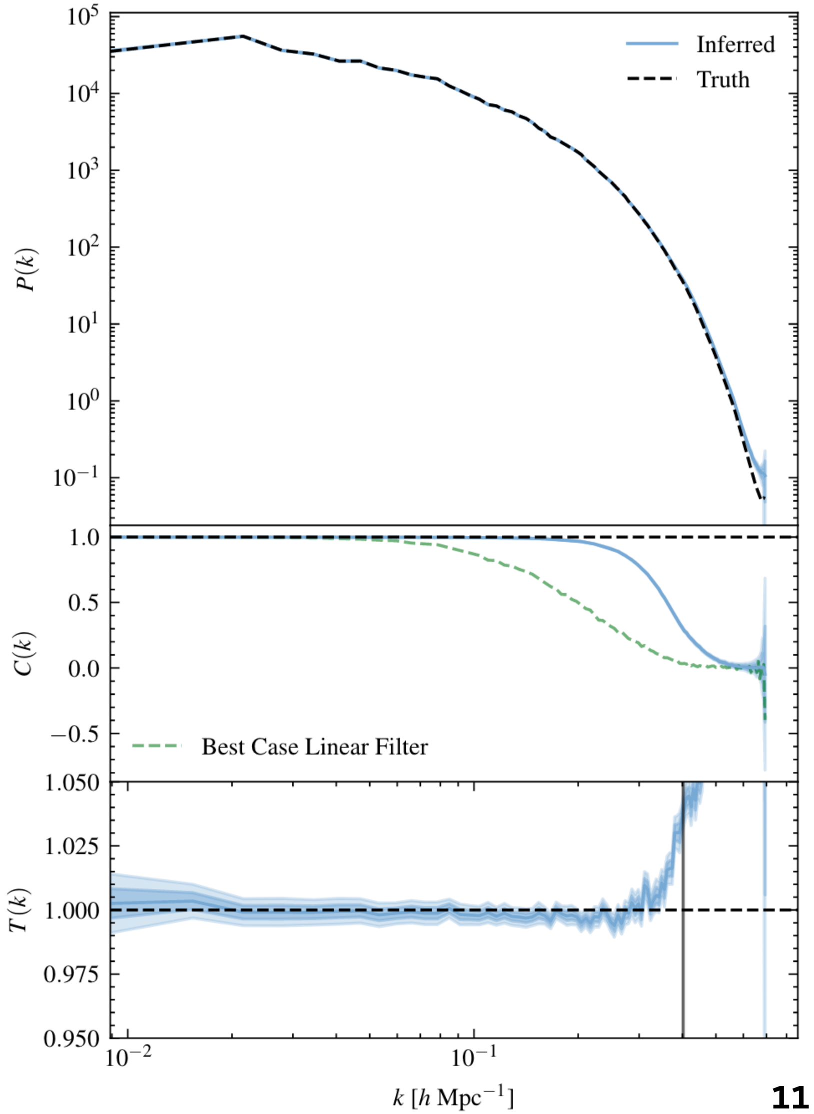
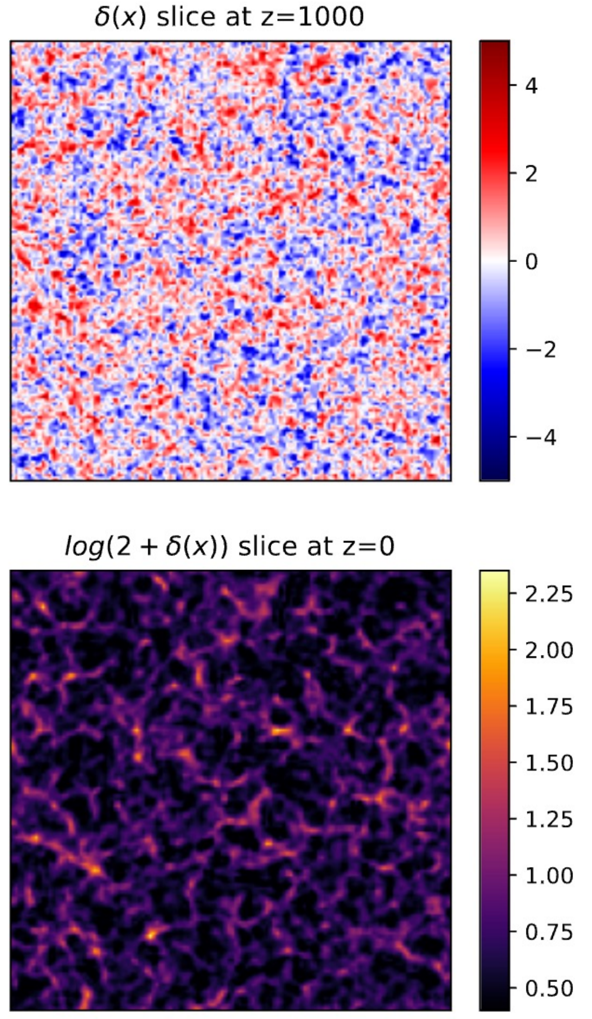


Figure 2. Left: The density field at redshift $z = 0$ for the fiducial Planck cosmology. Center: Initial conditions sampled from the posterior $p(\boldsymbol{x}|\boldsymbol{y})$. Right: The true initial conditions. All three density fields span a $1000 \times 1000 \times 125 (h^{-1}\mathrm{Mpc})^3$ region averaged over the third axis. This example demonstrates the capability of score-based generative models to sample highly detailed initial conditions consistent with the ground truth. See Figure 3 for quantification of uncertainty.



Our setting



Training data: **2000** 128^3 $(1\ \mathrm{Gpc/h})^3$ Simbelmyne or Quijote simulations

128^3 resolution: ~**million**-dimensional parameter space!

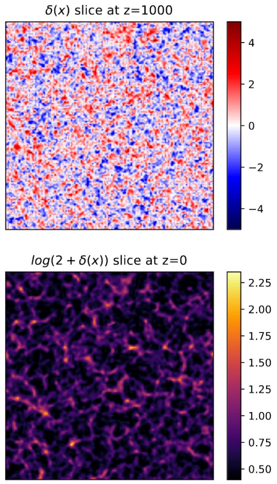
- Want to explore the full posterior, not only get point estimates
- Want to keep things as simple as possible
- Want to be able to do things sequentially: need to estimate the likelihood:

$$p(\boldsymbol{z}|\boldsymbol{x}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{z}-\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{x}))^T \boldsymbol{Q}_{\boldsymbol{\theta}}^L(\boldsymbol{z}-\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{x})) - \frac{1}{2}\boldsymbol{z}^T\boldsymbol{Q}^P\boldsymbol{z}\right\}$$

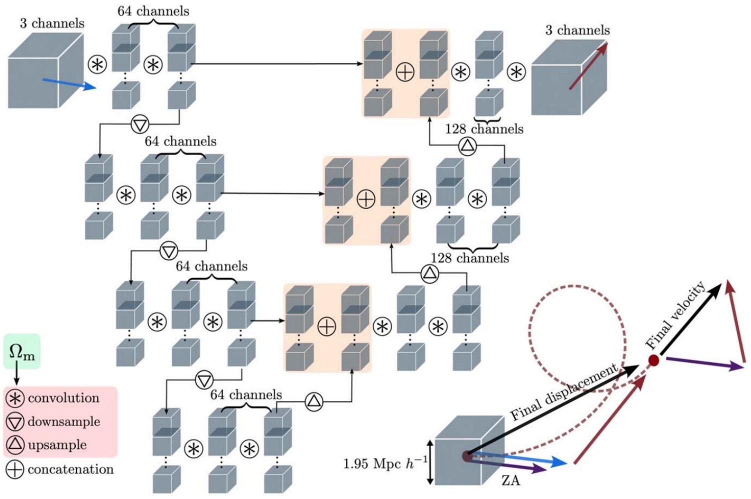
- Tried different approaches: autoregressive modelling, sliced score matching...

Our approach

Apply a U-Net to estimate μ_{MLE}



Training data



map2map U-Net

Jamieson+, 2206.04594

Train the model with a simple MAP loss function

$$\log \mathcal{L}(\vec{\mu}, \boldsymbol{Q}) = \mathbb{E}_{\vec{x}\sim p(\vec{x})}\left[\frac{1}{2}(\vec{x}_{(i)} - \vec{\mu})^T \boldsymbol{Q}(\vec{x}_{(i)} - \vec{\mu}) - \frac{1}{2}\mathrm{tr}(\log \boldsymbol{Q})\right]$$

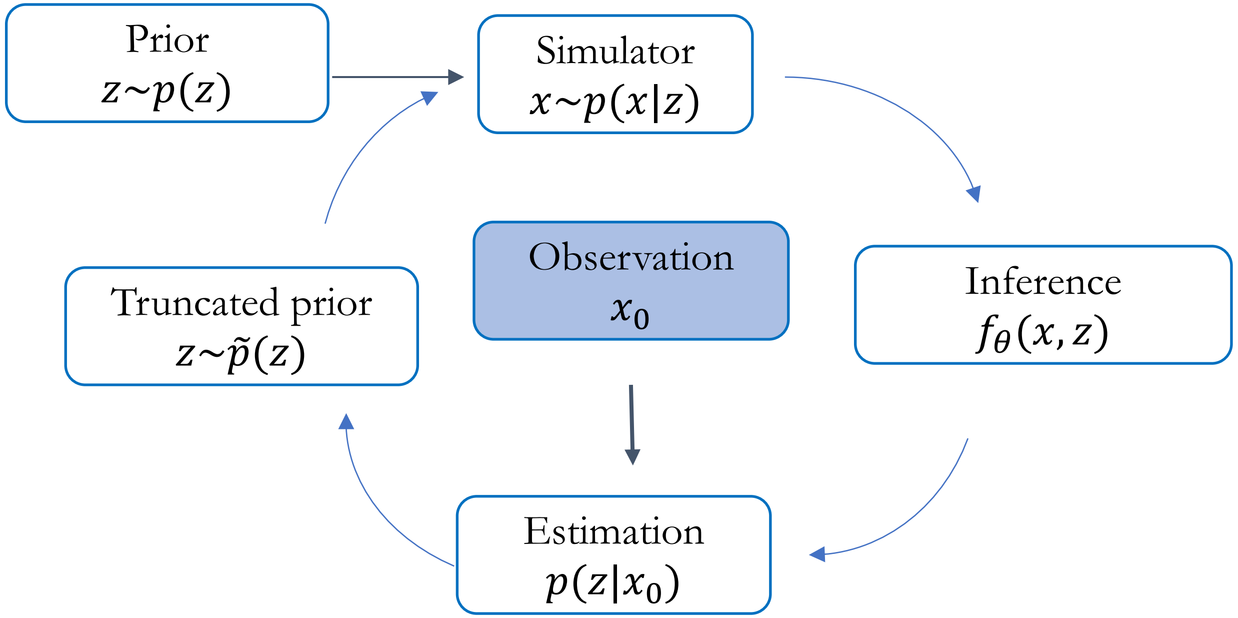
Q matrix diagonal in Fourier space, depends only on |***k***|

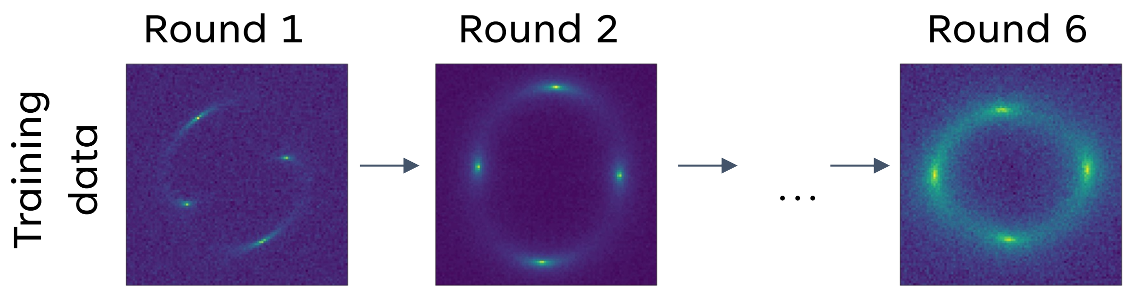
Learn correlation matrix and embedding network simultaneously

Super-fast sampling from a Gaussian

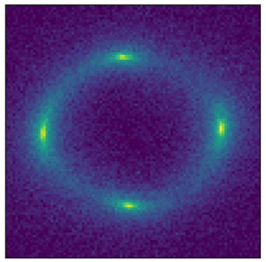
Sequential inference

- Our parameter space is too vast to explore
- Want to 'zoom in' into it and obtain precise results with a low number of simulations





Target observation



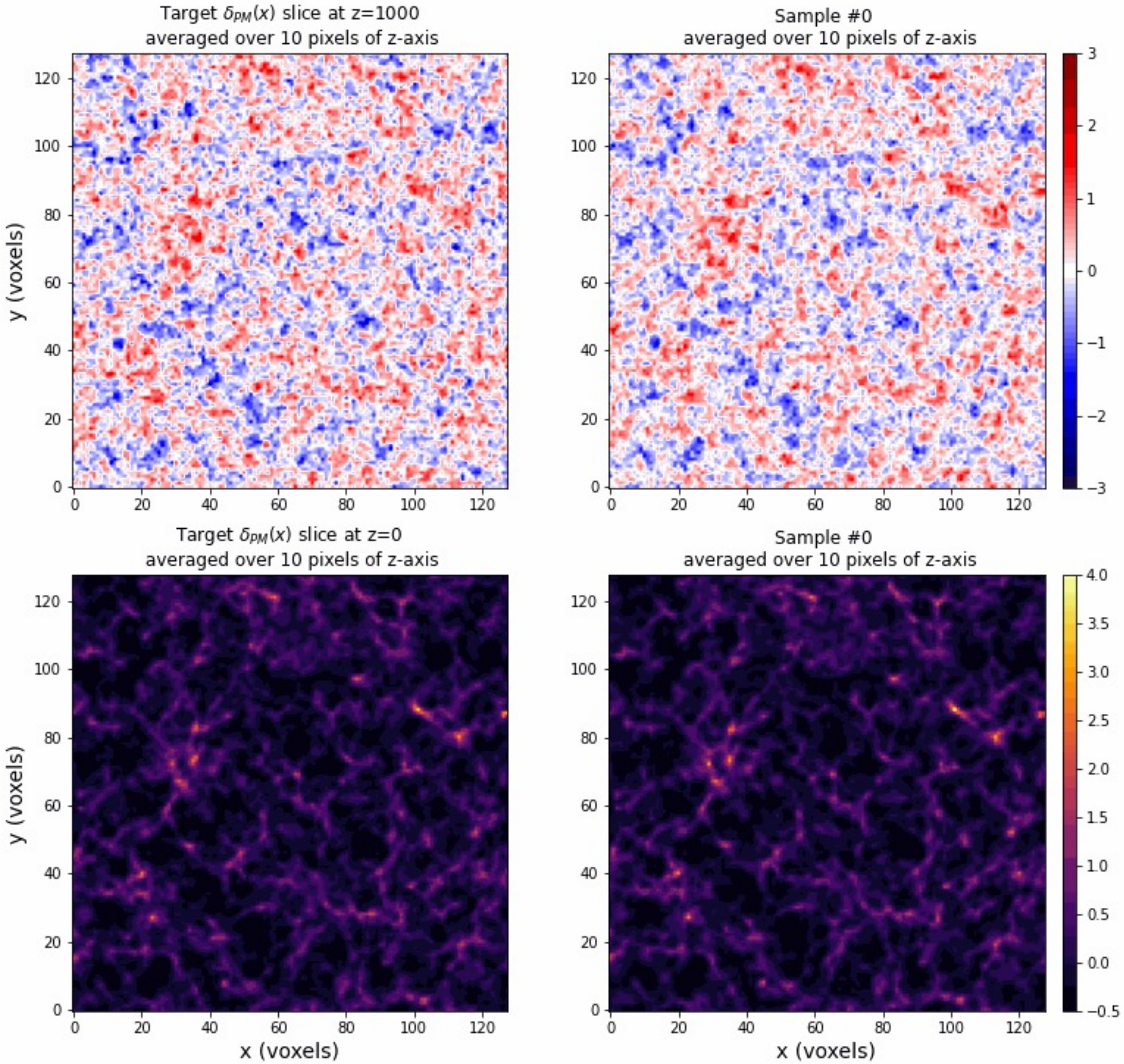
Learning with
high-precision

Slide credit:
N. Anau Montel

Results

Target image

Posterior draws



1 hour of training

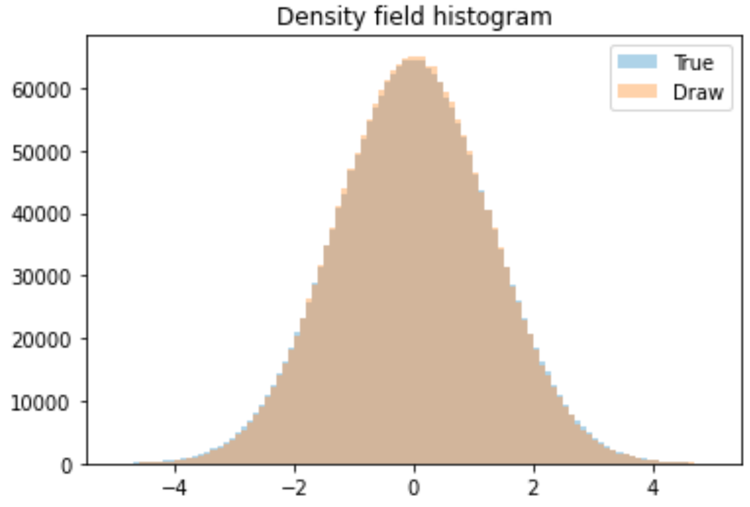
1 GPU

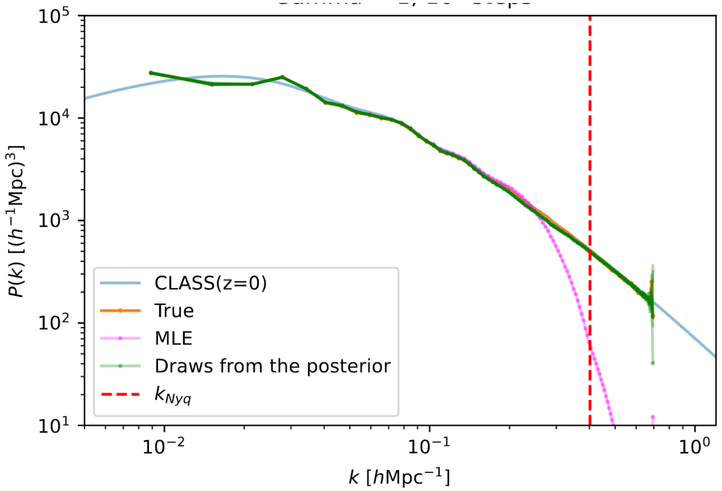
NVIDIA 40GB A100

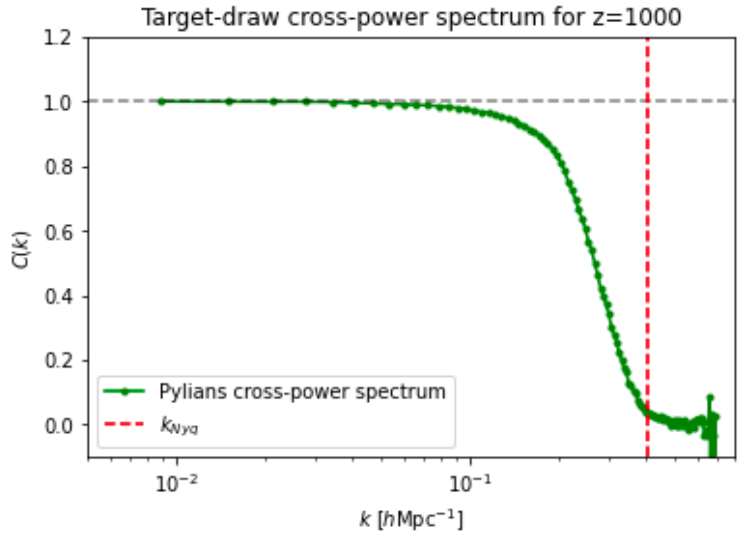
$(1\ \mathrm{Gpc}/h)^3$ volume

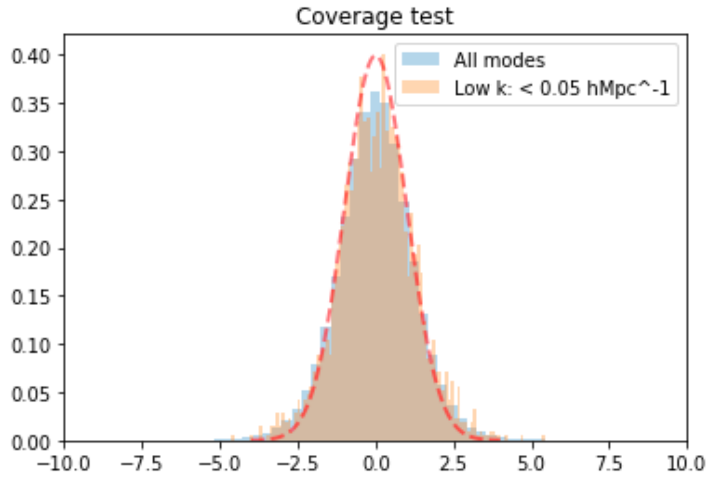
128^3 voxels

Summary statistics comparison









1-2% agreement in the power spectrum

Coverage test shows that samples follow the correct distribution

Summary

- Reconstruction of cosmological initial conditions is an important problem that allows to analise LSS data in the fullest way
- Methods like BORG achieve the goal, but are very slow, require gradients, run for months on supercomputers...
- Most ML approaches do point estimates, not explore the full posterior
- Our approach does this full exploration, is easy to train and produces samples in a super-fast way; it allows to turn any point estimator into a sampler
- The method is flexible and allows to do inference in an effective sequential way, with many applications in the future!