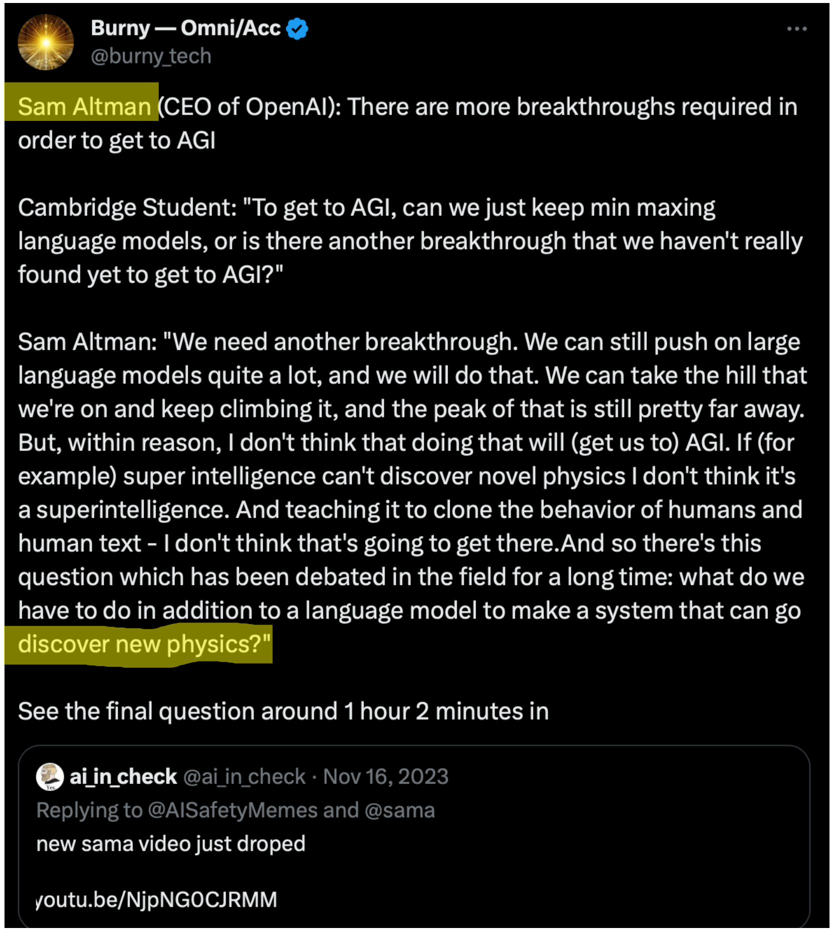
Cosmology and AI

Benjamin D. Wandelt

“Discovering new physics” is the new Turing test of Artificial General Intelligence (AGI)



Burny — Omni/Acc
@burny_tech

Sam Altman (CEO of OpenAI): There are more breakthroughs required in order to get to AGI

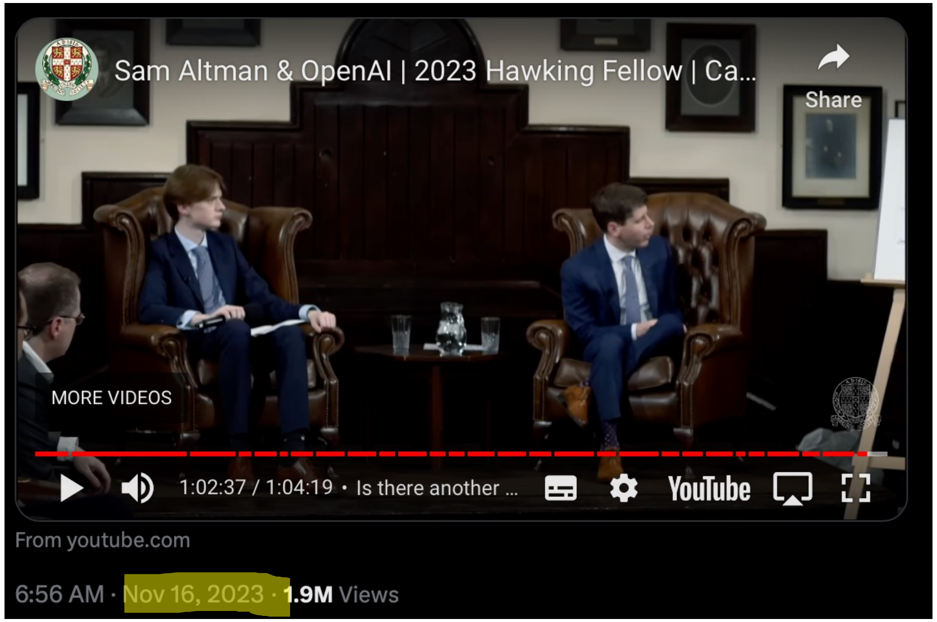
Cambridge Student: "To get to AGI, can we just keep min maxing language models, or is there another breakthrough that we haven't really found yet to get to AGI?"

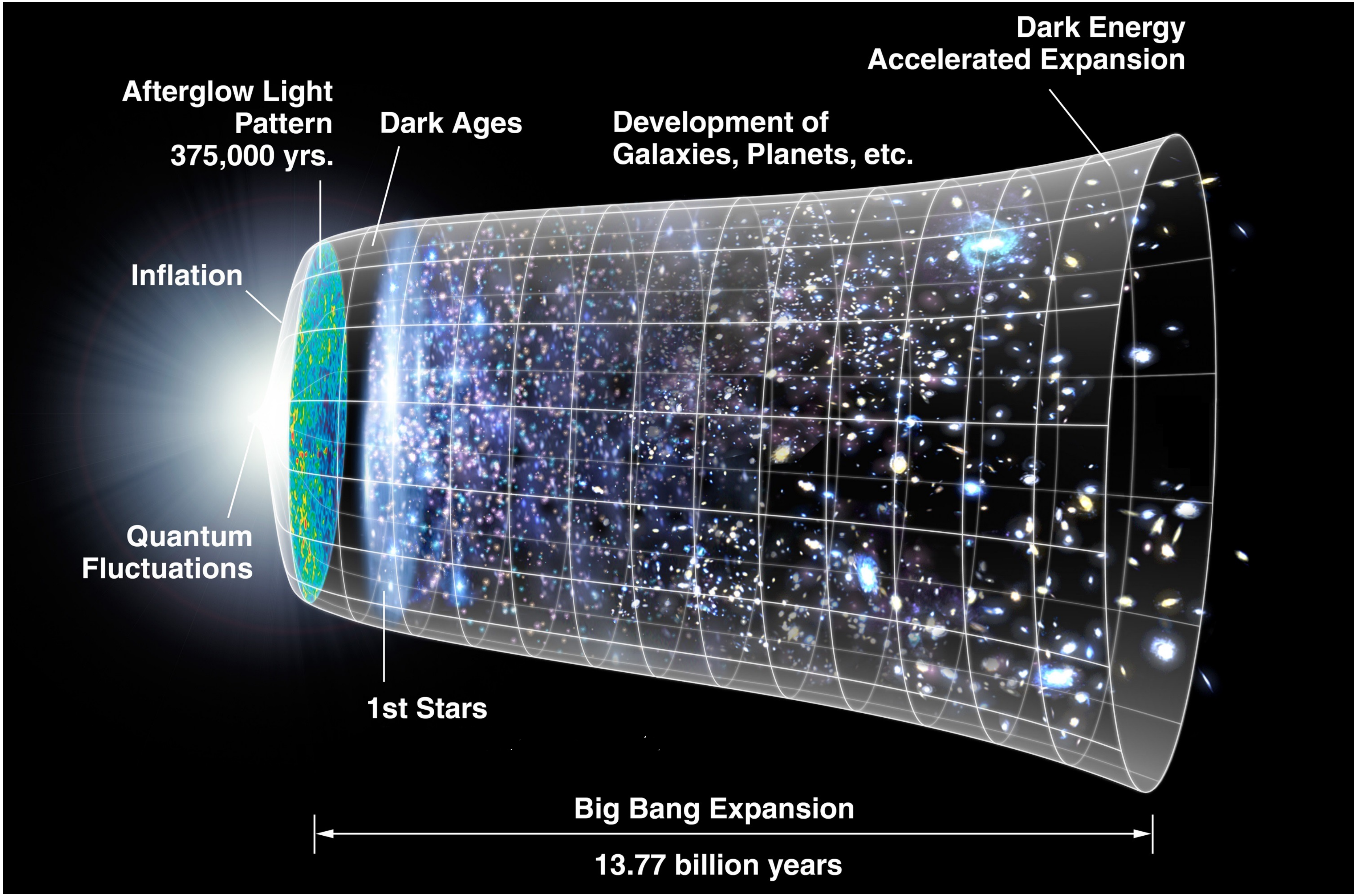
Sam Altman: "We need another breakthrough. We can still push on large language models quite a lot, and we will do that. We can take the hill that we're on and keep climbing it, and the peak of that is still pretty far away. But, within reason, I don't think that doing that will (get us to) AGI. If (for example) super intelligence can't discover novel physics I don't think it's a superintelligence. And teaching it to clone the behavior of humans and human text - I don't think that's going to get there.And so there's this question which has been debated in the field for a long time: what do we have to do in addition to a language model to make a system that can go discover new physics?"

See the final question around 1 hour 2 minutes in

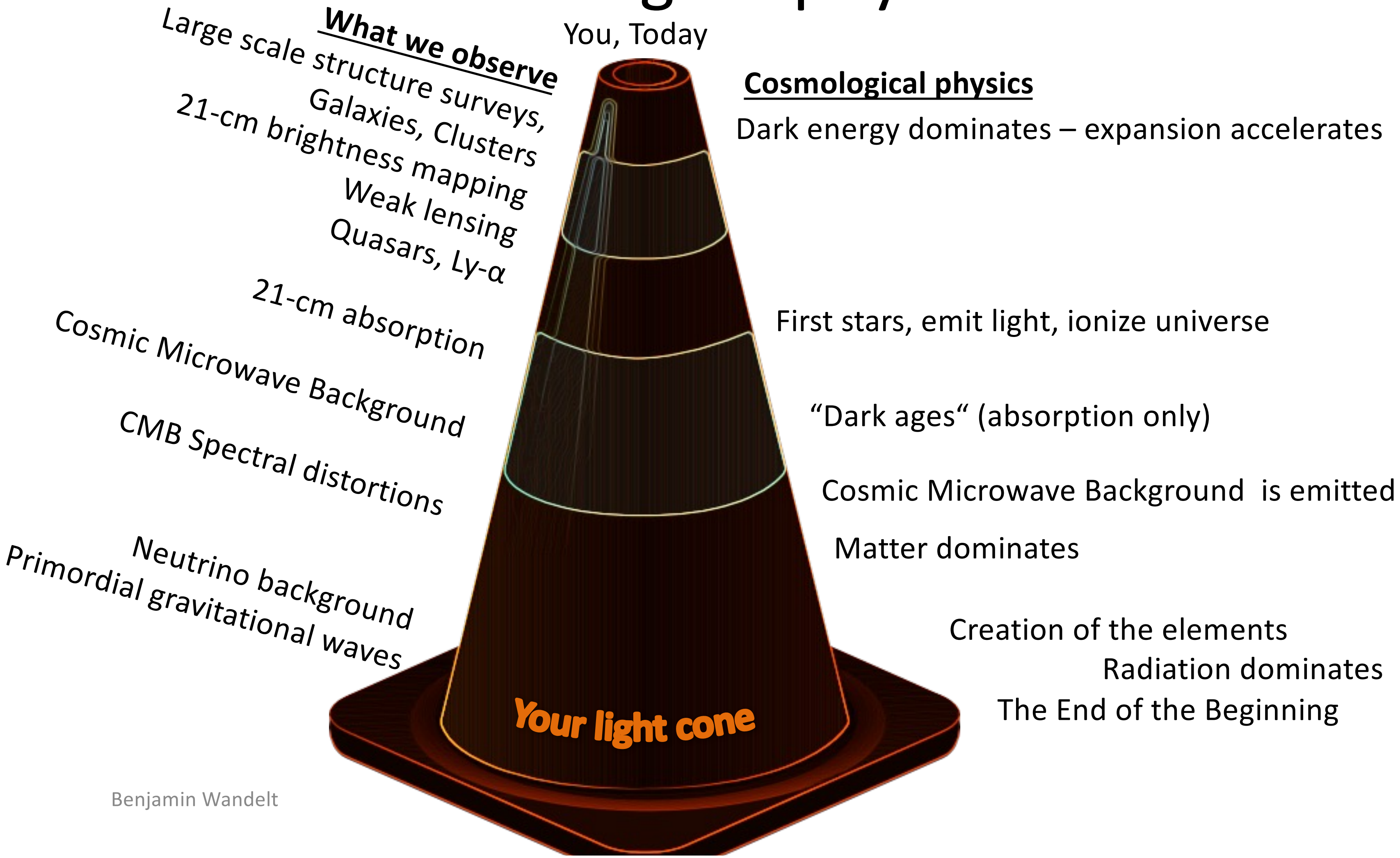
ai_in_check @ai_in_check · Nov 16, 2023
Replying to @AISafetyMemes and @sama
new sama video just droped

youtu.be/NjpNG0CJRMM




Dark Energy
Accelerated Expansion
Afterglow Light
Pattern
375,000 yrs.
Dark Ages
Development of
Galaxies, Planets, etc.
Inflation
Quantum
Fluctuations
1st Stars
Big Bang Expansion
13.77 billion years

Our past lightcone: the playground of cosmological physics



Benjamin Wandelt

What we want to learn

non-Gaussianity
Inflation
$A_s, n_s, r, f_{nl}, \ldots$ **Cosmic Beginning**
Primordial gravitational waves

Neutrino mass
Black holes
$\Omega_m, \Omega_b, m_\nu, \tau, \ldots$ **Cosmic Content**
Dark matter baryons
Star formation

$H_0, \Omega_\Lambda, w_0, w_a, \ldots$ **Cosmic Fate**
Expansion speed
Dark Energy

Statistics of the "initial conditions"

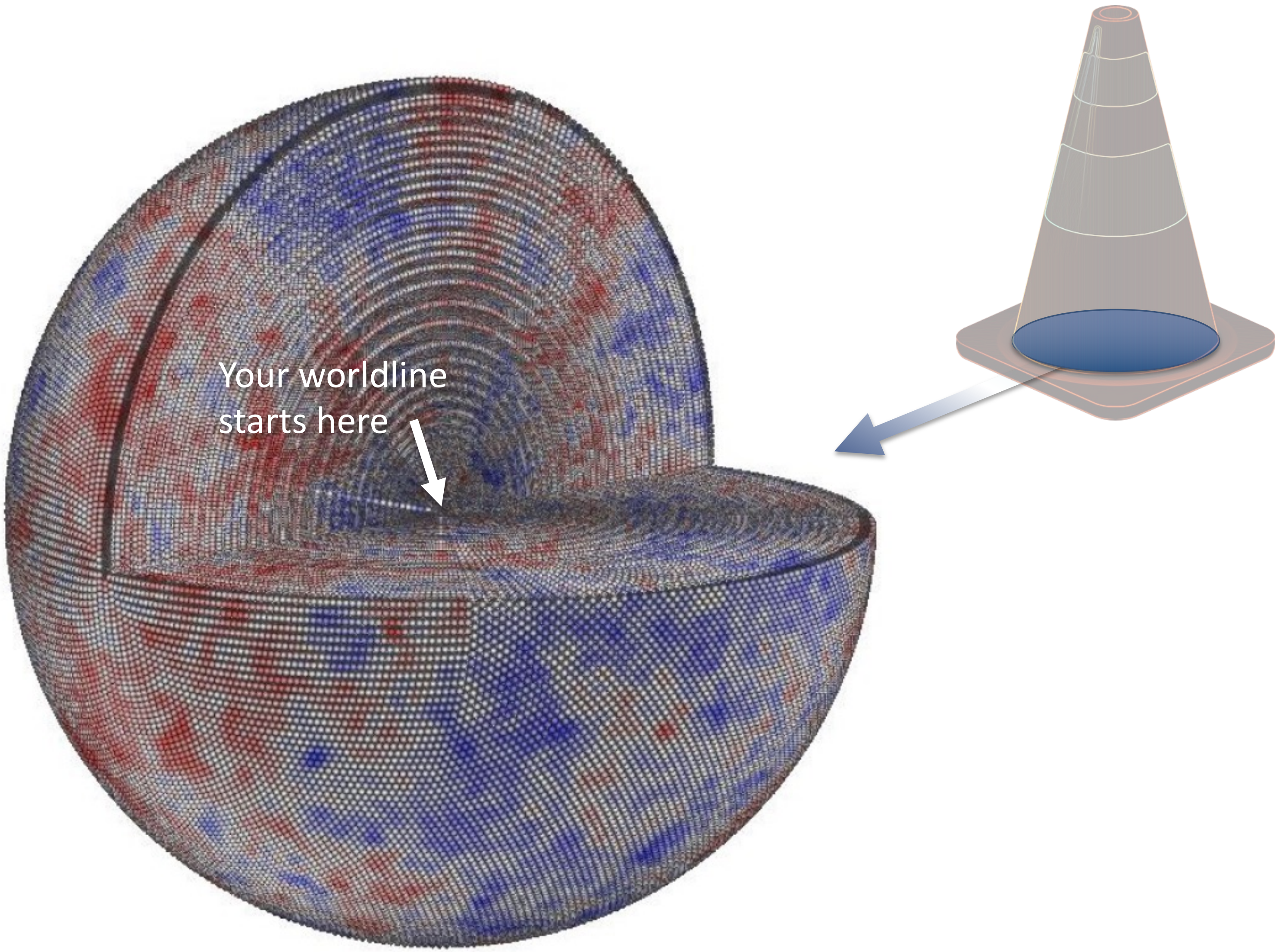
Recipe for our universe

Dynamics

Expansion geometry

The initial conditions of the universe on the base of the past light cone

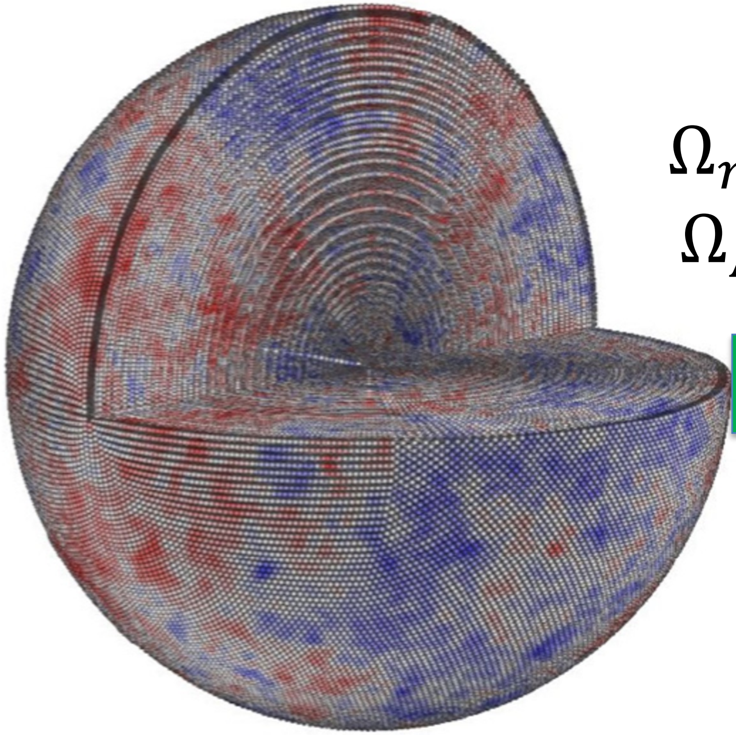
(curvature perturbations)



Benjamin Wandelt

Computational Cosmology

$A_s, n_s, r, f_{nl}, \ldots$

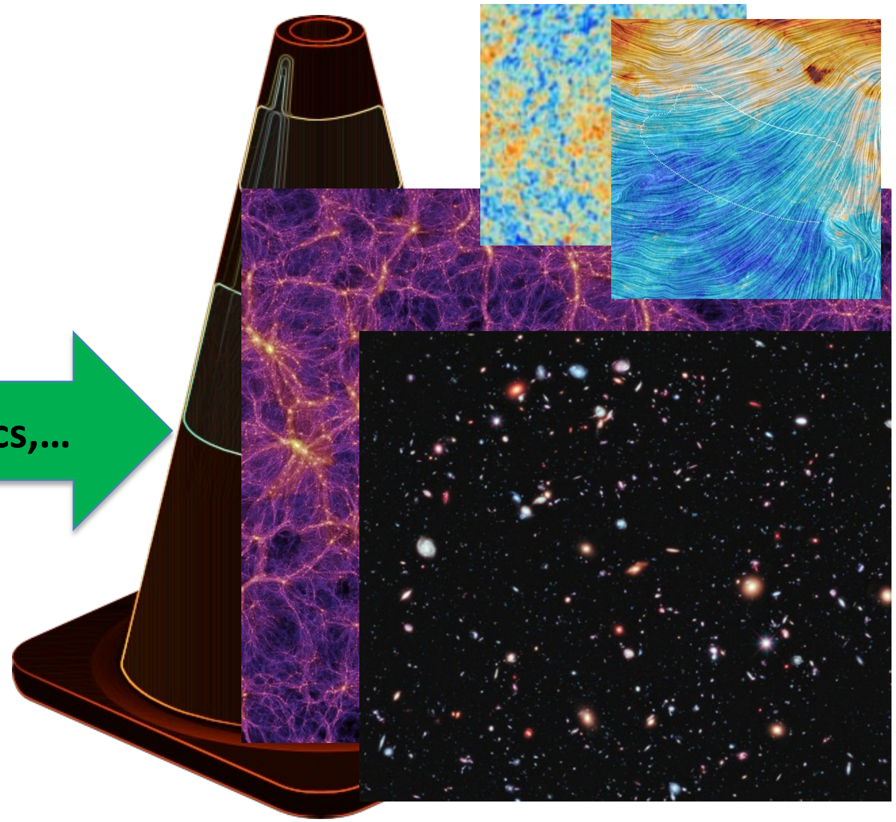


$\Omega_m, \Omega_b, m_\nu, \ldots$

$\Omega_\Lambda, w_0, w_a, \ldots$

Gravity, Hydrodynamics,…

Galaxy formation
and evolution,
Black holes,
Star formation

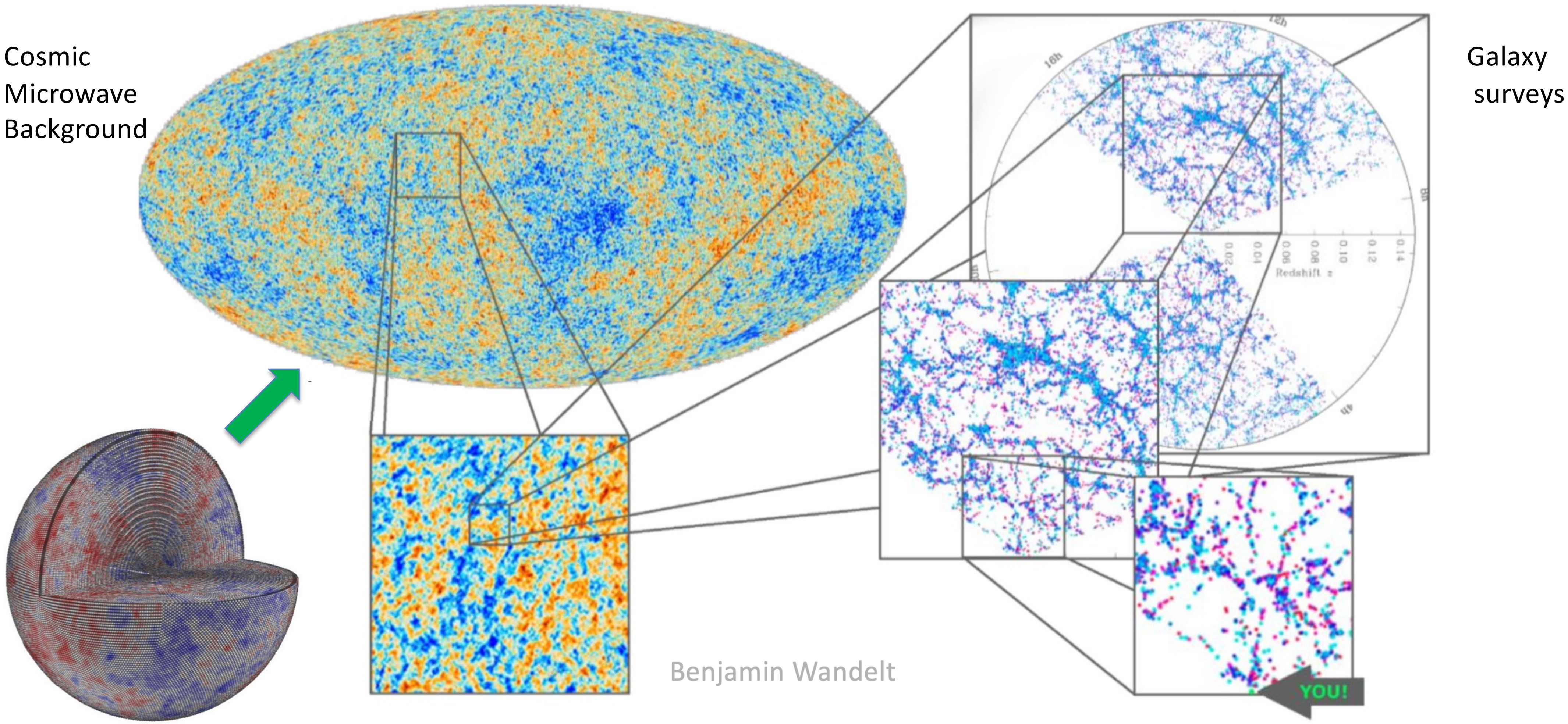


Initial conditions of the universe

The observed universe

Benjamin Wandelt

How do we handle the vast scale hierarchies?



The simulation-analysis asymmetry

- Computational simulation models have become very detailed digital twins of the universe.

- But *data analysis* has been based on restrictive analytical/perturbative/semi-numerical approximations.

Benjamin Wandelt

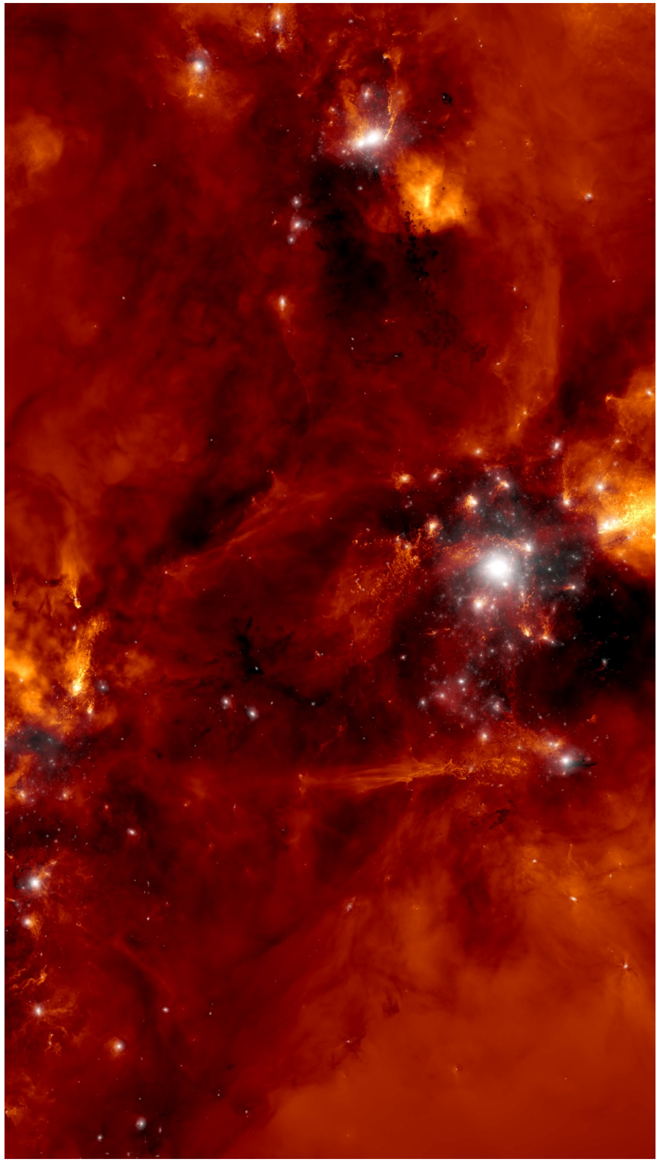


Image credit: Illustris TNG50

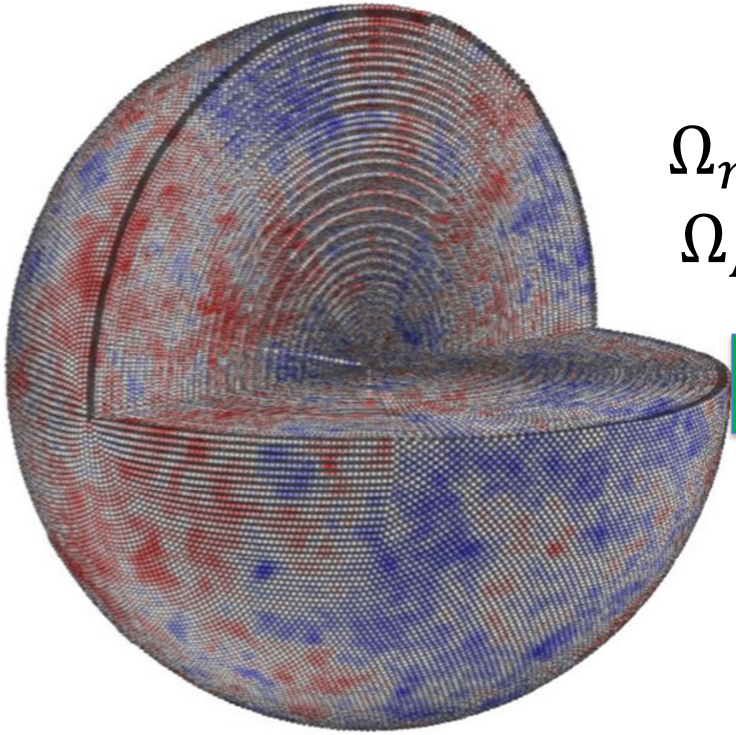
How to deal with complexity on small scales?

Image Credit: Millenium Simulation

How to deal with complexity on small scales?

Computational Cosmology

$A_s, n_s, r, f_{nl}, \ldots$

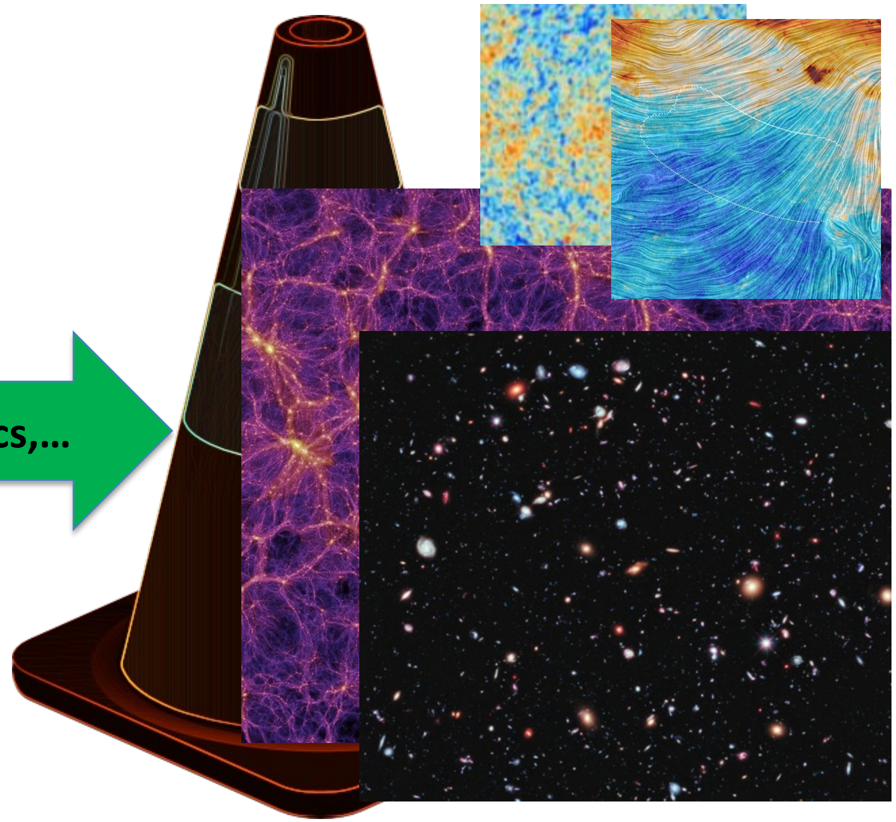


Initial conditions of the universe

$\Omega_m, \Omega_b, m_\nu, \ldots$
$\Omega_\Lambda, w_0, w_a, \ldots$

Gravity, Hydrodynamics,…

Galaxy formation
and evolution,
Black holes,
Star formation

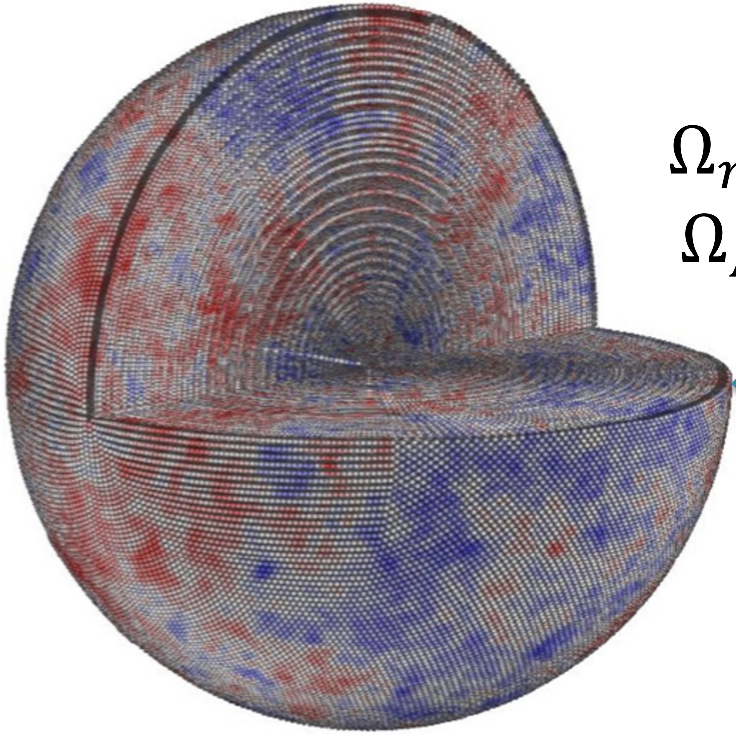


The observed universe

Benjamin Wandelt

Cosmological and Astrophysical Discovery

$A_s, n_s, r, f_{nl}, \ldots$

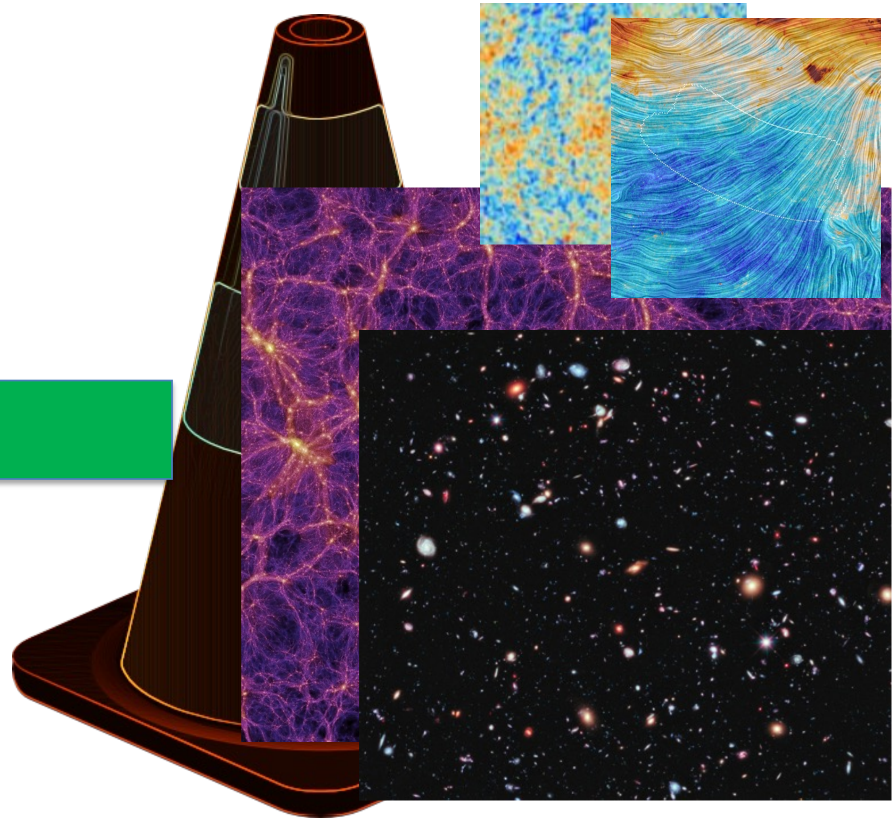


$\Omega_m, \Omega_b, m_\nu, \ldots$

$\Omega_\Lambda, w_0, w_a, \ldots$

Inference

Galaxy formation and evolution, Black holes, Star formation



Initial conditions of the universe

The observed universe

Benjamin Wandelt

The promise of AI for cosmology

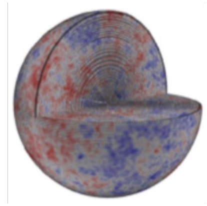
- Can we use AI to bridge the scale hierarchies?
- Can we use AI to have the same fidelity in data analysis as in modeling?

Machine learning, Artificial Intelligence, and Bayesian Analysis open a new way to connect theory and data towards more a more symmetrical analysis

This is made possible by breakthroughs in "Implicit Inference" methods and ML-accelerated generation/simulation models

In this approach cosmological and astrophysical physics connnect with data at a much more fine-grained level than in the past, unlocking large discovery potential.

The Simons Collaboration on "Learning the Universe"



- Recognizes the opportunity to cross scales gaps using recent ML breakthroughs
- International collaboration that brings together experts on
  - Star formation,
  - ISM,
  - Black hole accretion,
  - Galaxy formation and evolution,
  - Cosmology
  - ML, and
  - Bayesian Inference
- Goal: Prove the principle on current data sets and develop methods for the next generation

- Director: Greg Bryan
- PIs: Simone Ferraro, Lars Hernquist, Shirley Ho, Jens Jasche, Guilhem Lavaux, Eve Ostriker, Laurence Perreault-Levasseur, Aarti Singh, Rachel Somerville, Volker Springel, Ben Wandelt

SIMONS FOUNDATION

We have no lack of data

Can we analyze data if all we can do is simulate it?

Yes!

A major shift over the last 5 years.

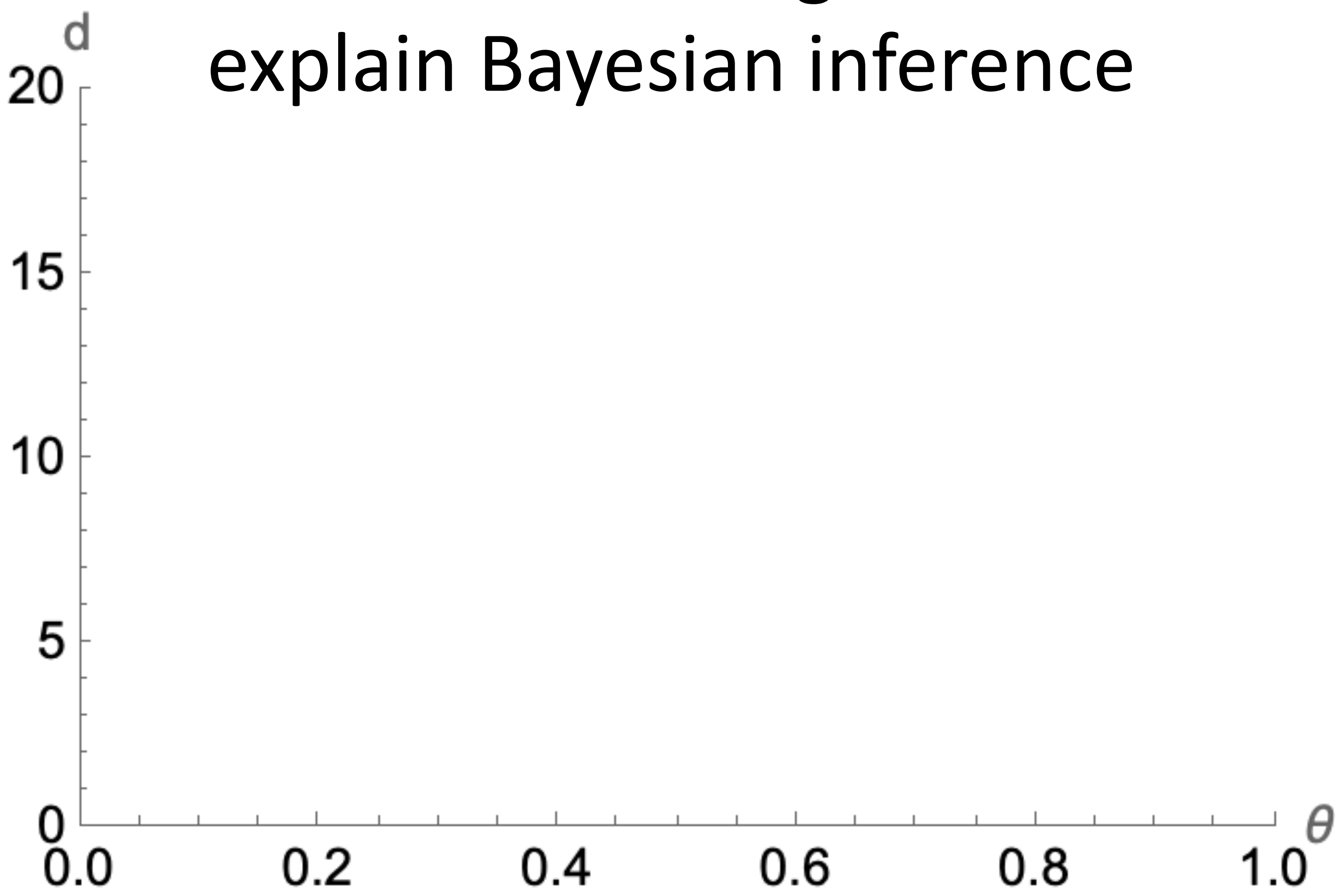
Likelihood is represented *implicitly* through simulations

$$d \hookleftarrow p(d|\theta)$$

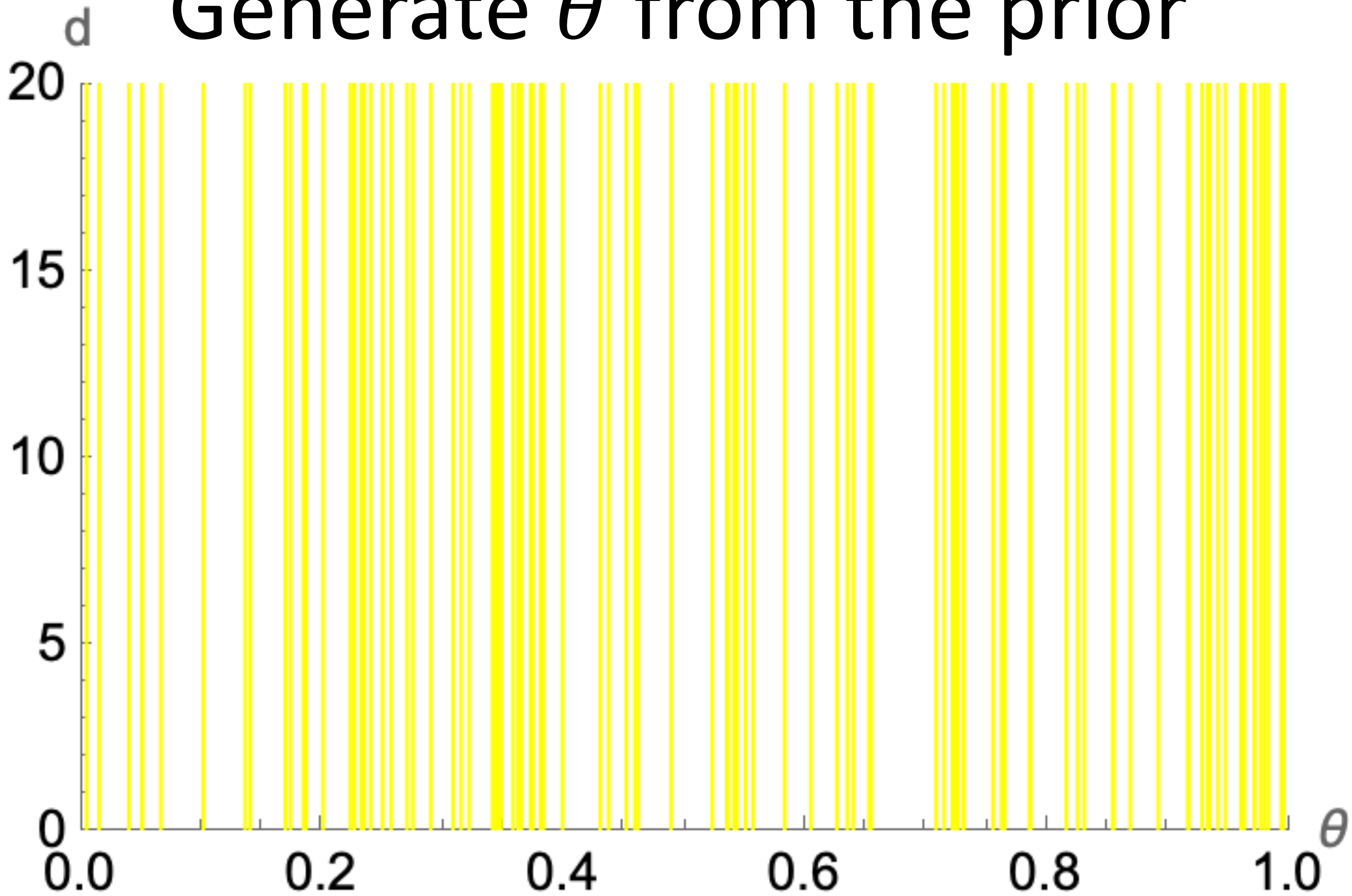
Let's do a simple example.

Challenge: Keep a running count of the number of likelihood and prior evaluations!

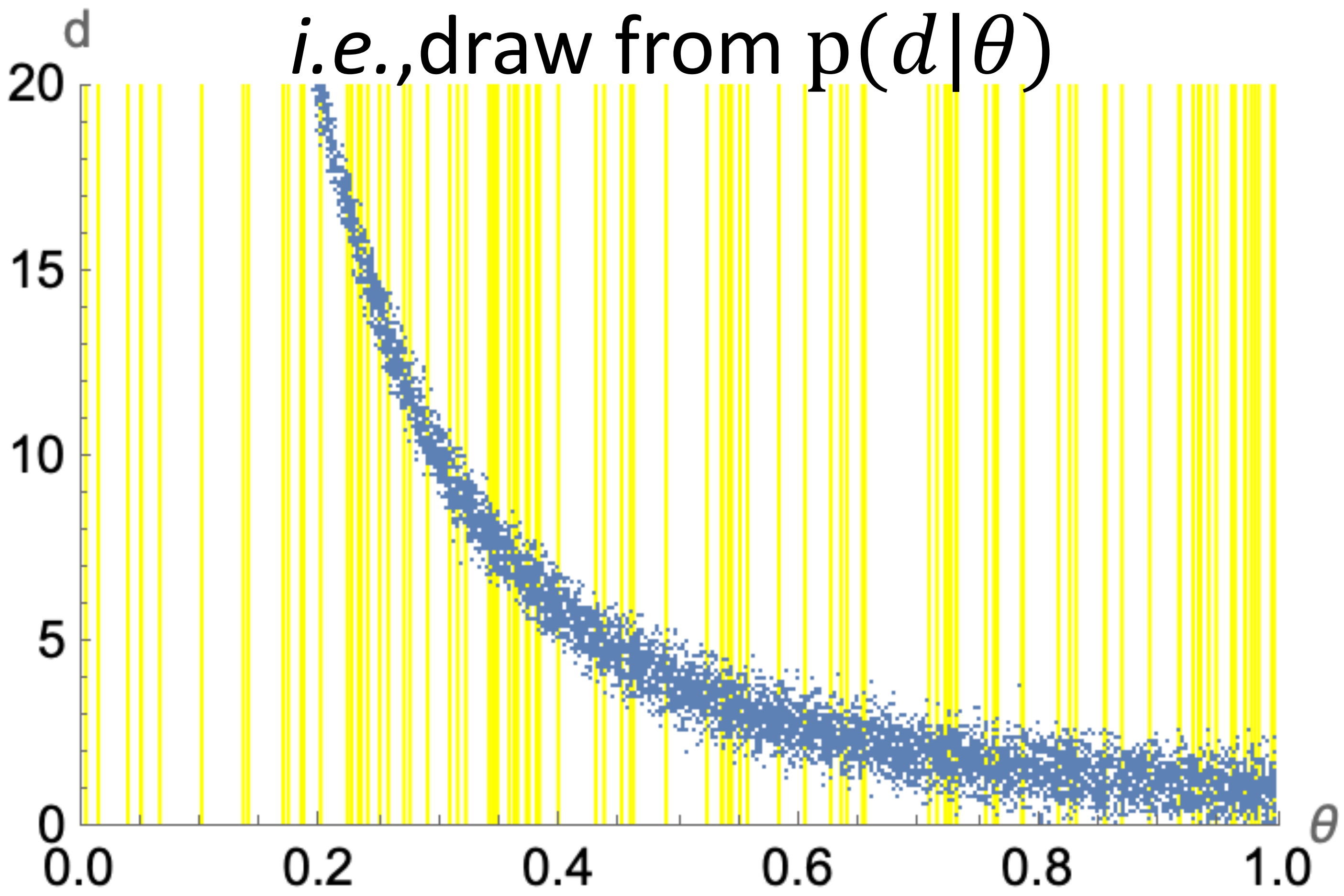
The easiest diagram to explain Bayesian inference



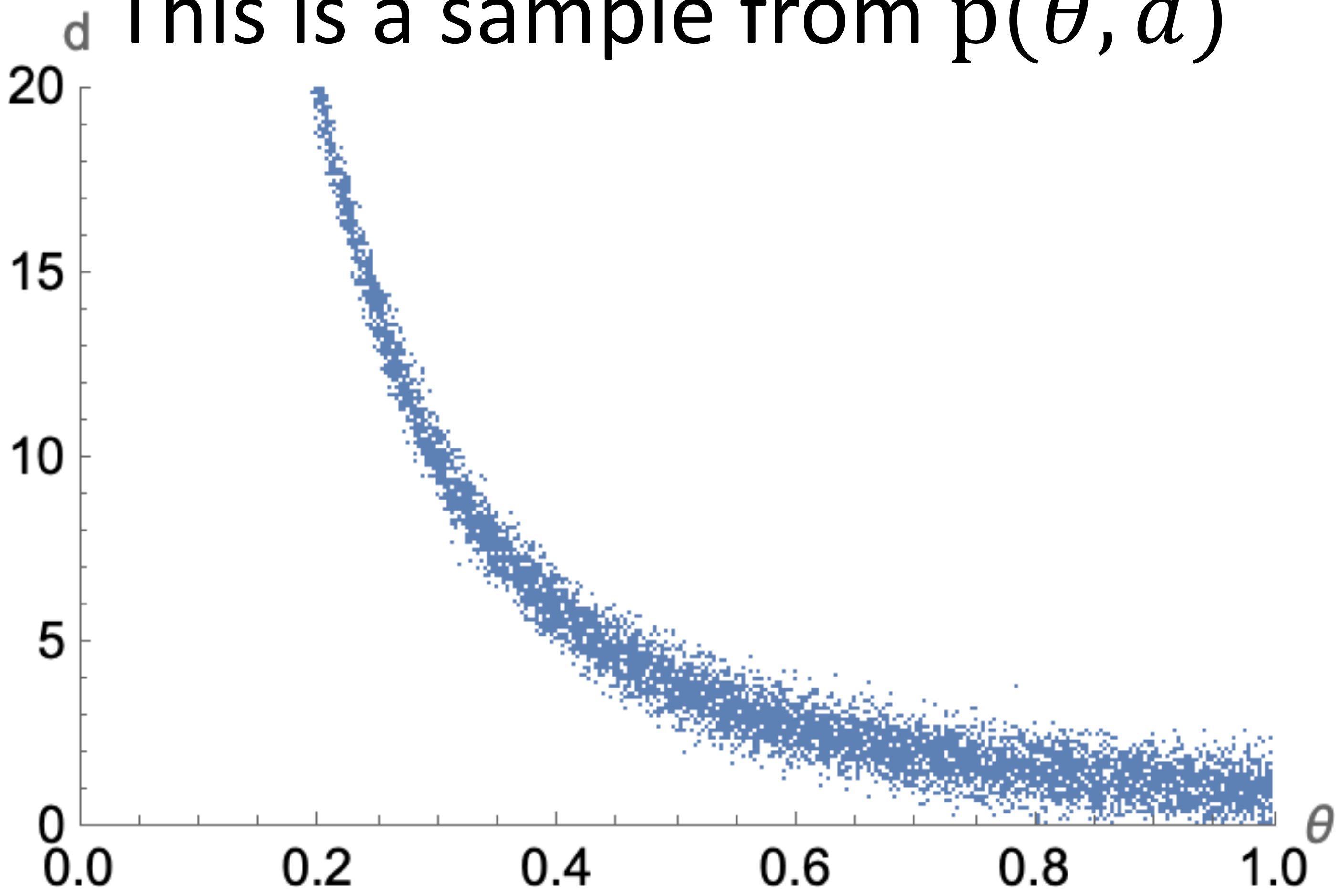
Generate θ from the prior



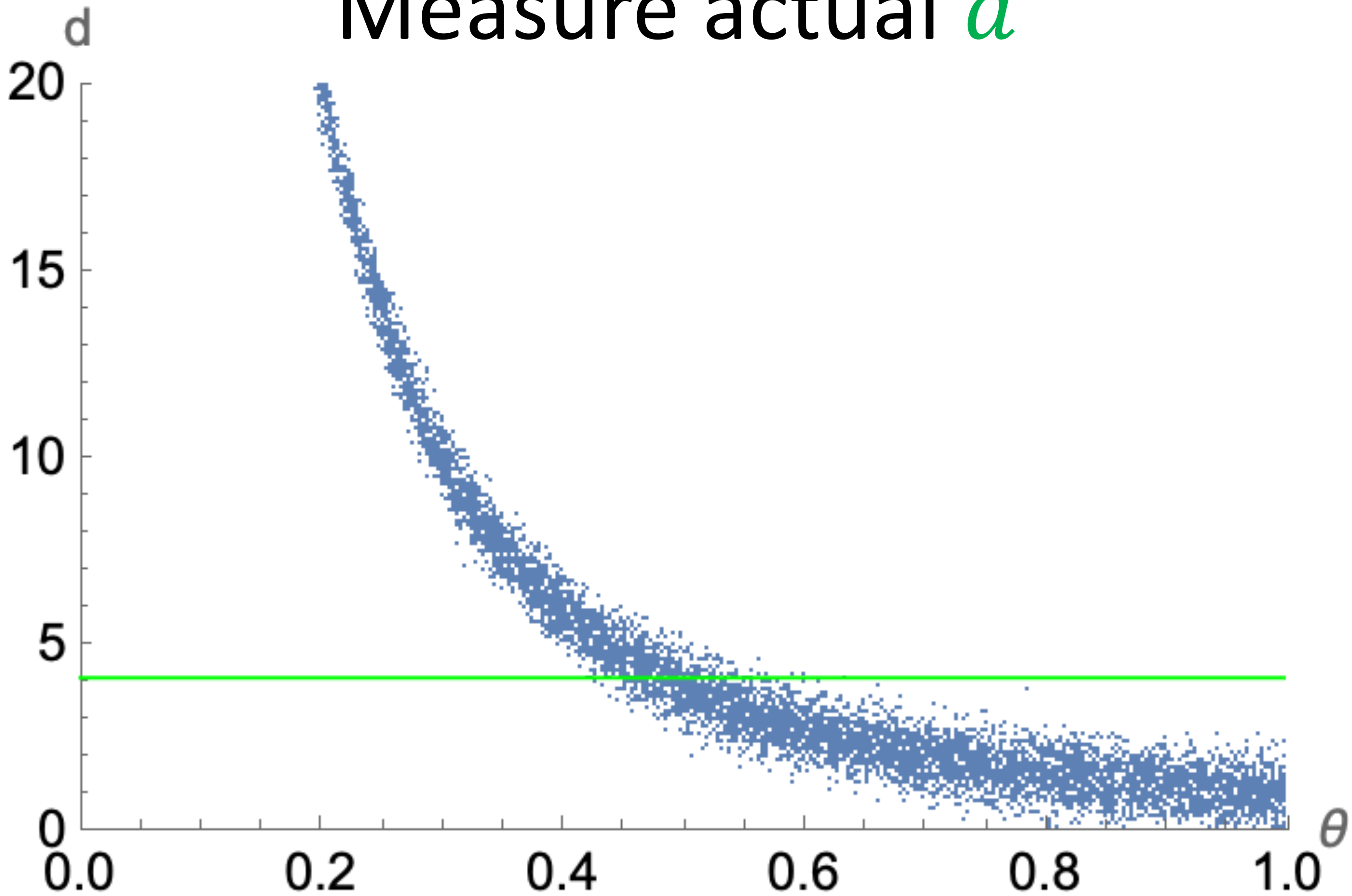
Simulate/generate d given θ
i.e., draw from $\mathrm{p}(d|\theta)$



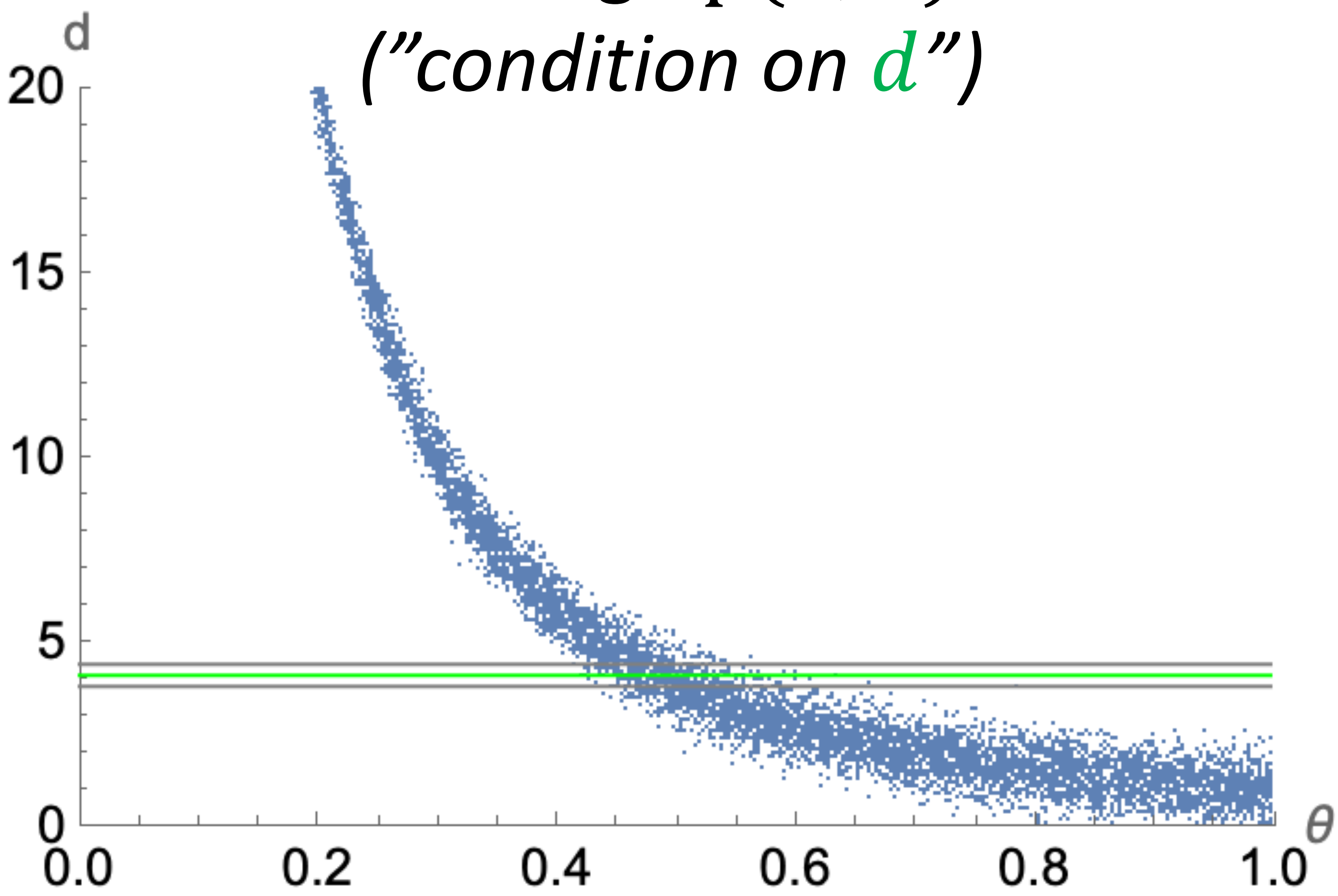
This is a sample from $p(\theta, d)$



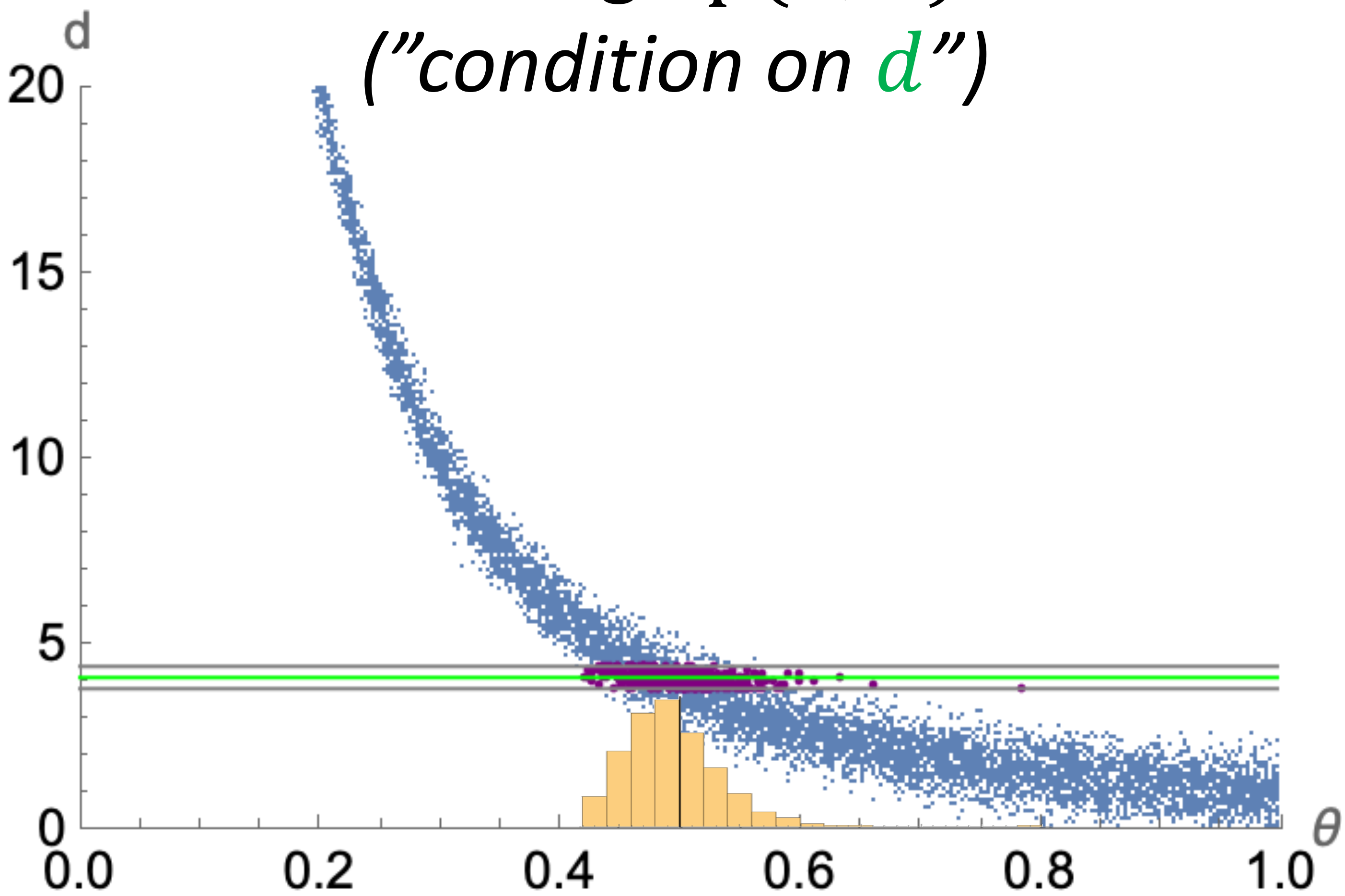
Measure actual d



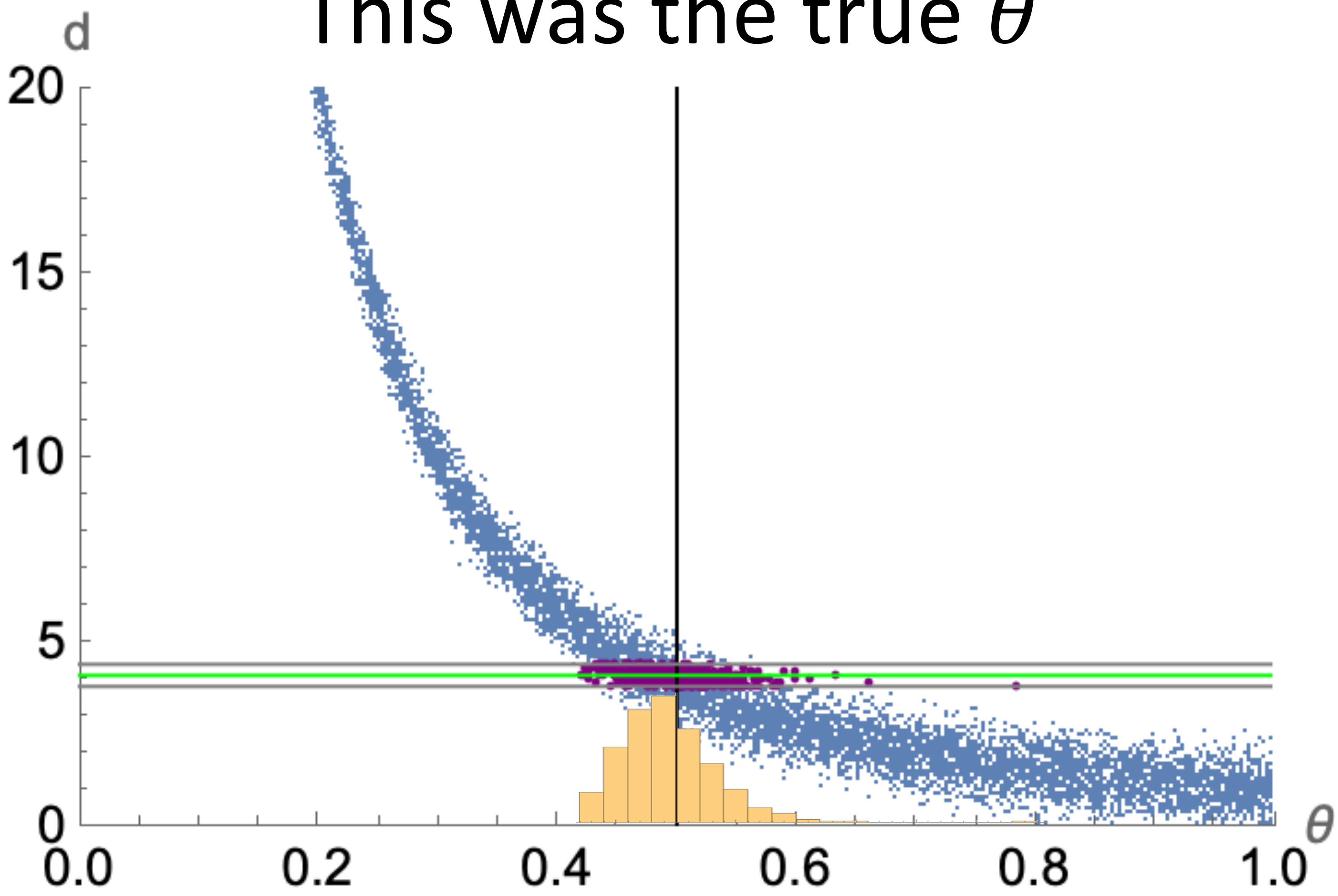
Slice through $\mathrm{p}(\theta, d)$ at d
("condition on d")



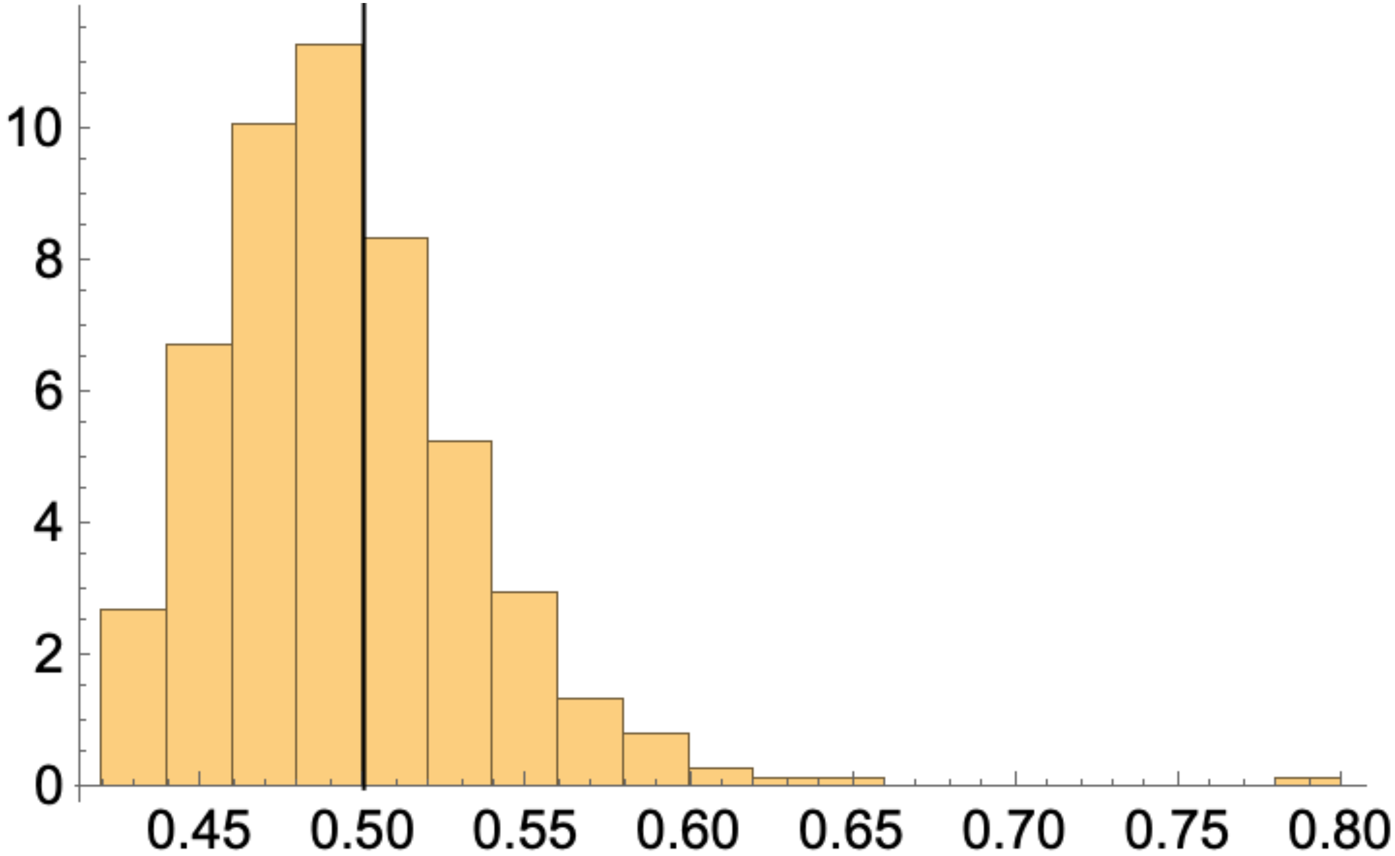
Slice through $\mathrm{p}(\theta, d)$ at d
("condition on d")



This was the true θ



Posterior $\mathrm{p}(\theta|d)$



How many likelihood evaluations?

How many prior evaluations?

This is ***Implicit* Inference**

- When likelihood and/or prior are not *explicitly* specified but *implicit* in…
  - simulations, generative models, labelled data.

- Various forms known as
  - Likelihood-free inference
  - Simulation-based inference
  - Approximate Bayesian Computation (ABC)

Machine learning takes us the rest of the way

- Many problems that we considered impossible now **solved**
  - Automated finding of informative data summary statistics
    - computing informative summaries for intractable models (*e.g.,* IMNN, FI)
  - Posteriors/likelihoods/priors for intractable models
    - **Implicit Inference** (likelihood-free, or simulation-based): LRE, DELFI
    - Routinely used to compute Bayesian posteriors (*e.g.,* Moment Networks)
    - Posterior samples for huge non-linear inverse problems (*e.g.,* Initial Conditions)
  - Bayesian Evidence for intractable models
    - Evidence Networks

IMNN: Charnock, Lavaux & Wandelt arXiv:1802.03537; LRE: Cranmer, Pavez & Louppe 1506.02169; Miller et al. 2107.01214
DELFI:Papamakarios, Murray + coauthors: 1705.07057, 1805.07226; Alsing & Wandelt 1712.00012;
Alsing, Feeney & Wandelt 1801.01497, arXiv:1903.01473; MN & EN: Jeffrey & Wandelt arXiv:2011.05991, 2305.11241;
FI: Coulton & Wandelt 2305.08994, ICs: Legin et al., 2304.03788

Simplest example

- What do you train a ML model f(x) to compute when you train it with (x, y) pairs to predict y from x, minimizing squared error?

$$L = \sum_i (f(x_i) - y_i)^2$$

$$\approx \int (f(x) - y)^2 p(x, y) dx dy$$

minimize $\rightarrow \quad \hat{f} = \int y \; p(y|x) dy$

MOMENT AND POSTERIOR MARGINAL NETWORKS

Main idea: construct $\mathcal{F}(d)$,$\mathcal{G}(d)$ to go directly from data to posterior.

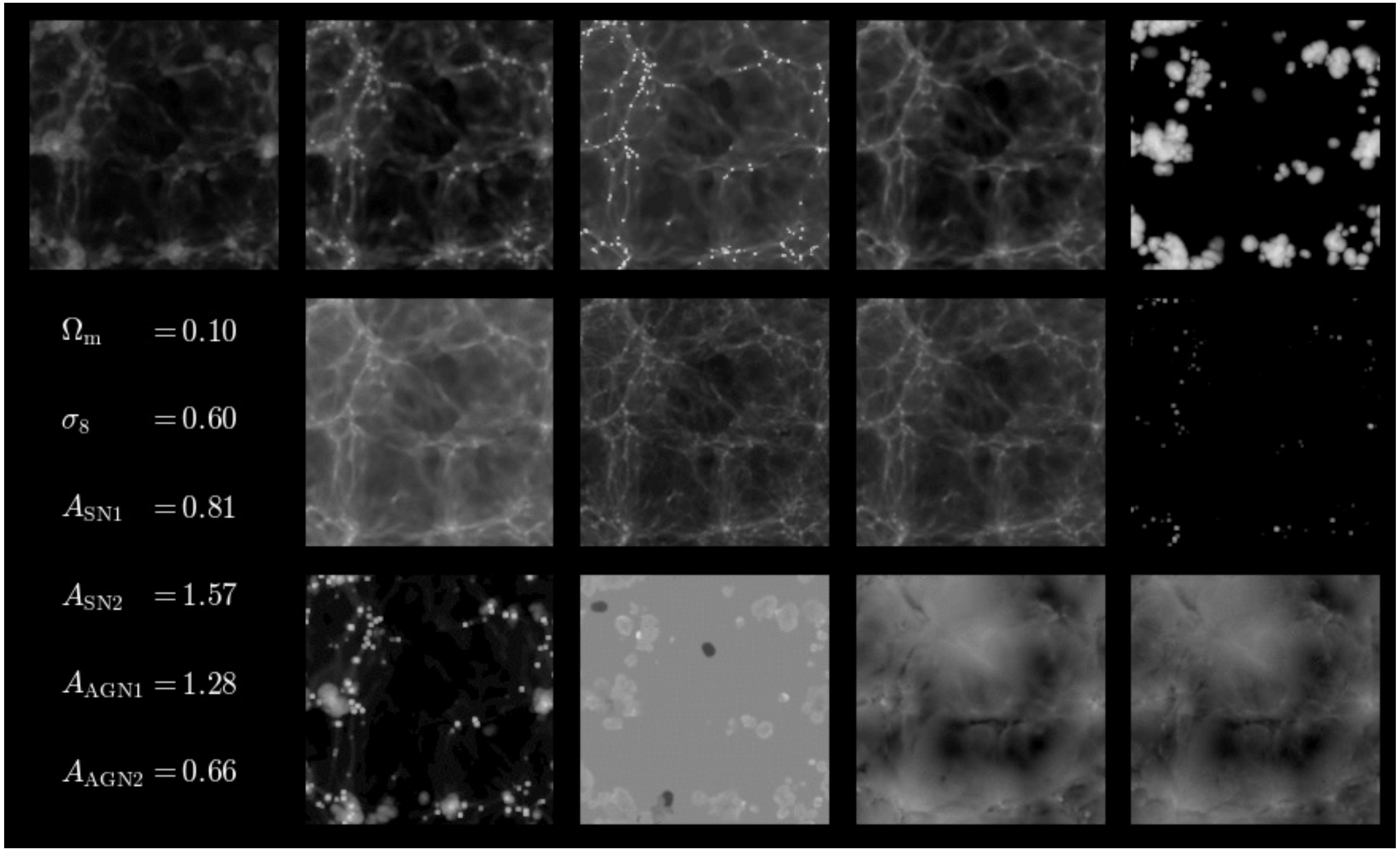
- **Moment networks**: obtain posterior moments directly from data by training NNs to solve

$$\langle\theta\rangle_{p(\theta|d)} = \underset{\mathcal{F}(d)}{\arg\min} \int ||\theta - \mathcal{F}(d)||_2^2 \; p(d,\theta) ddd\theta$$

$$\mathrm{Var}[\theta]_{p(\theta|d)} = \underset{\mathcal{G}(d)}{\arg\min} \int || \, ||\theta - \langle\theta\rangle_{p(\theta|d)}||_2^2 - \mathcal{G}(d)||_2^2 \; p(d,\theta) ddd\theta$$

(Jeffrey & Wandelt arXiv:2011.05991, presented at NeurIPS 2020)

Example: CAMELS hydrosimulations



Paco Villaescusa-Navarro,
Shy Genel,
Daniel Angles-Alcazar, and
the CAMELS collaboration

13 fields from

1000 IllustrisTNG sims
1000 SIMBA sims
and
2000 matched Nbody sims

arXiv:2109.10915

https://camels-multifield-dataset.readthedocs.io

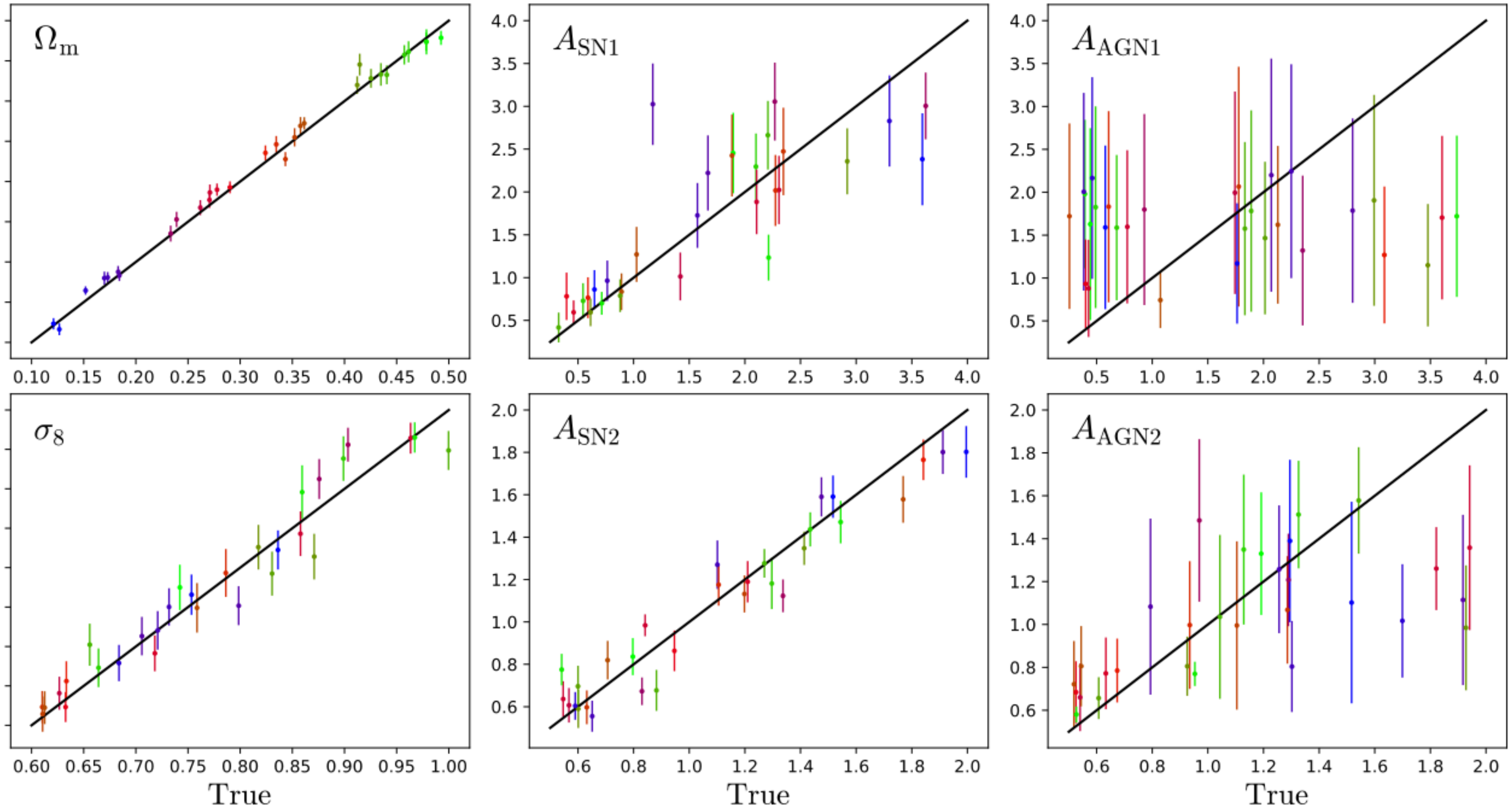
SBI: COSMOLOGY FROM SMALL-SCALE HYDRO

Villaescusa-Navarro, arXiv:2109.09747

Computing posterior means & variances from **gas temperature**

$$\mathcal{L} = \sum_{i=1}^{6} \log\left(\sum_{j\in\text{batch}} (\theta_{i,j}-\mu_{i,j})^2\right) + \sum_{i=1}^{6} \log\left(\sum_{j\in\text{batch}} \left((\theta_{i,j}-\mu_{i,j})^2 - \sigma_{i,j}^2\right)^2\right)$$

Posterior means & variances computed by **moment network** minimizing $\mathcal{L}$



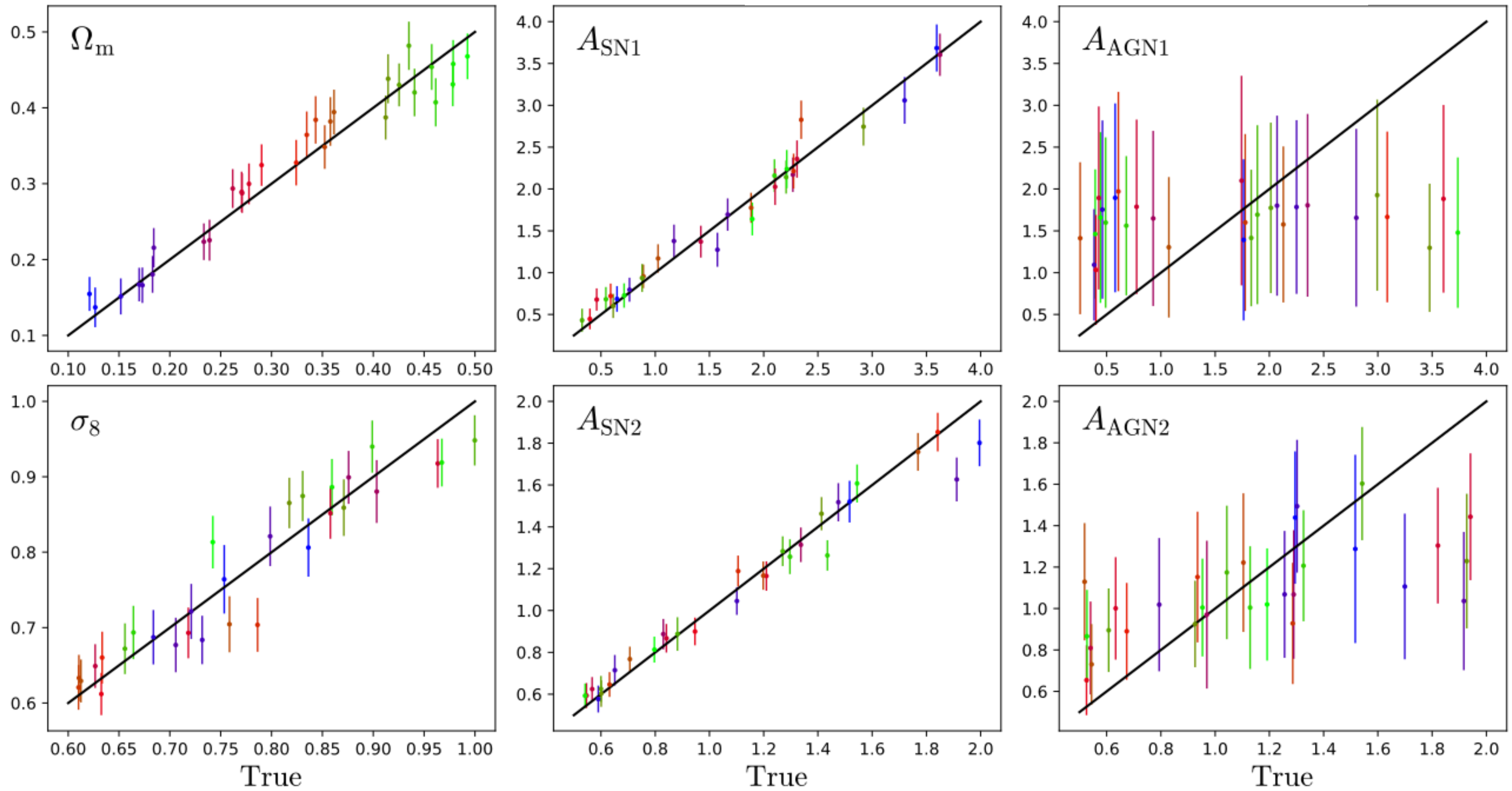
SBI: Cosmology from Small-Scale Hydro

Villaescusa-Navarro, arXiv:2109.09747

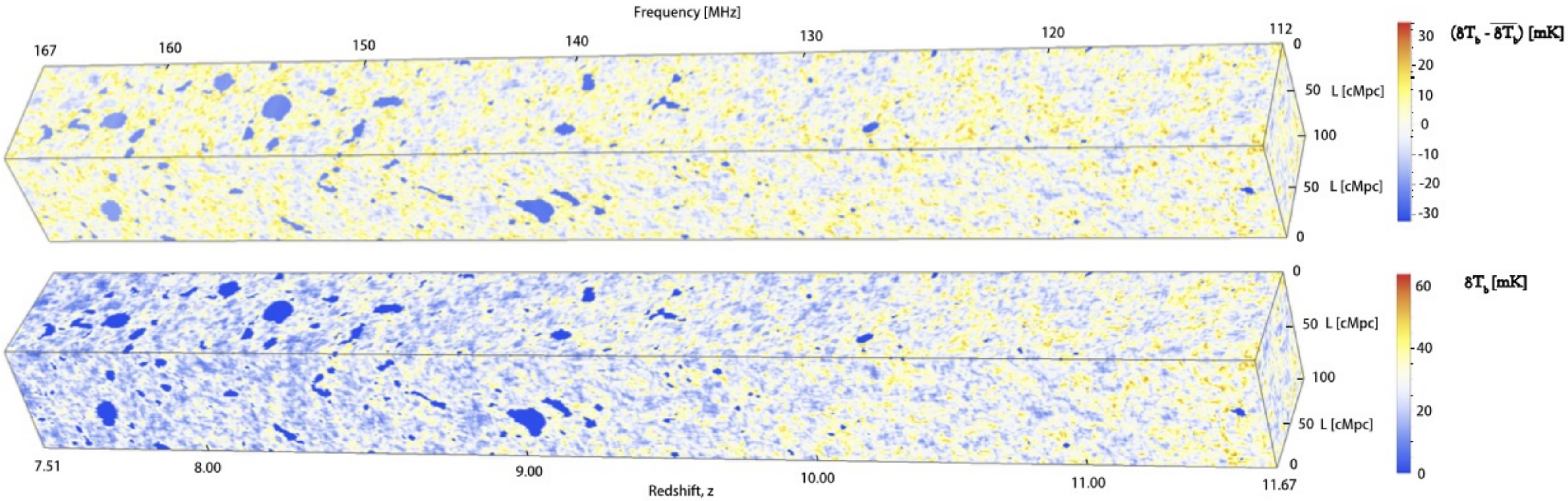
Computing posterior means & variances from **gas metallicity**

$$\mathcal{L} = \sum_{i=1}^{6} \log\left(\sum_{j\in\text{batch}} (\theta_{i,j} - \mu_{i,j})^2\right) + \sum_{i=1}^{6} \log\left(\sum_{j\in\text{batch}} \left((\theta_{i,j} - \mu_{i,j})^2 - \sigma_{i,j}^2\right)^2\right)$$

Posterior means & variances computed by **moment network** minimizing $\mathcal{L}$



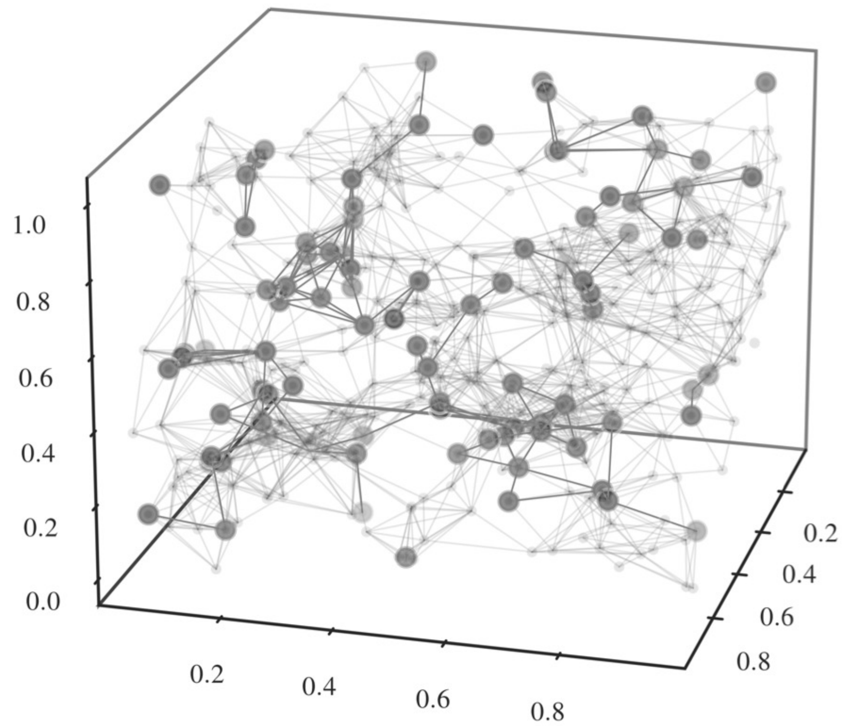
Examples: Implicit Inference to Infer Reionization Parameters from 3D 21cm Light Cones



Zhao, Mao, Cheng, Wandelt arXiv:2105.03344

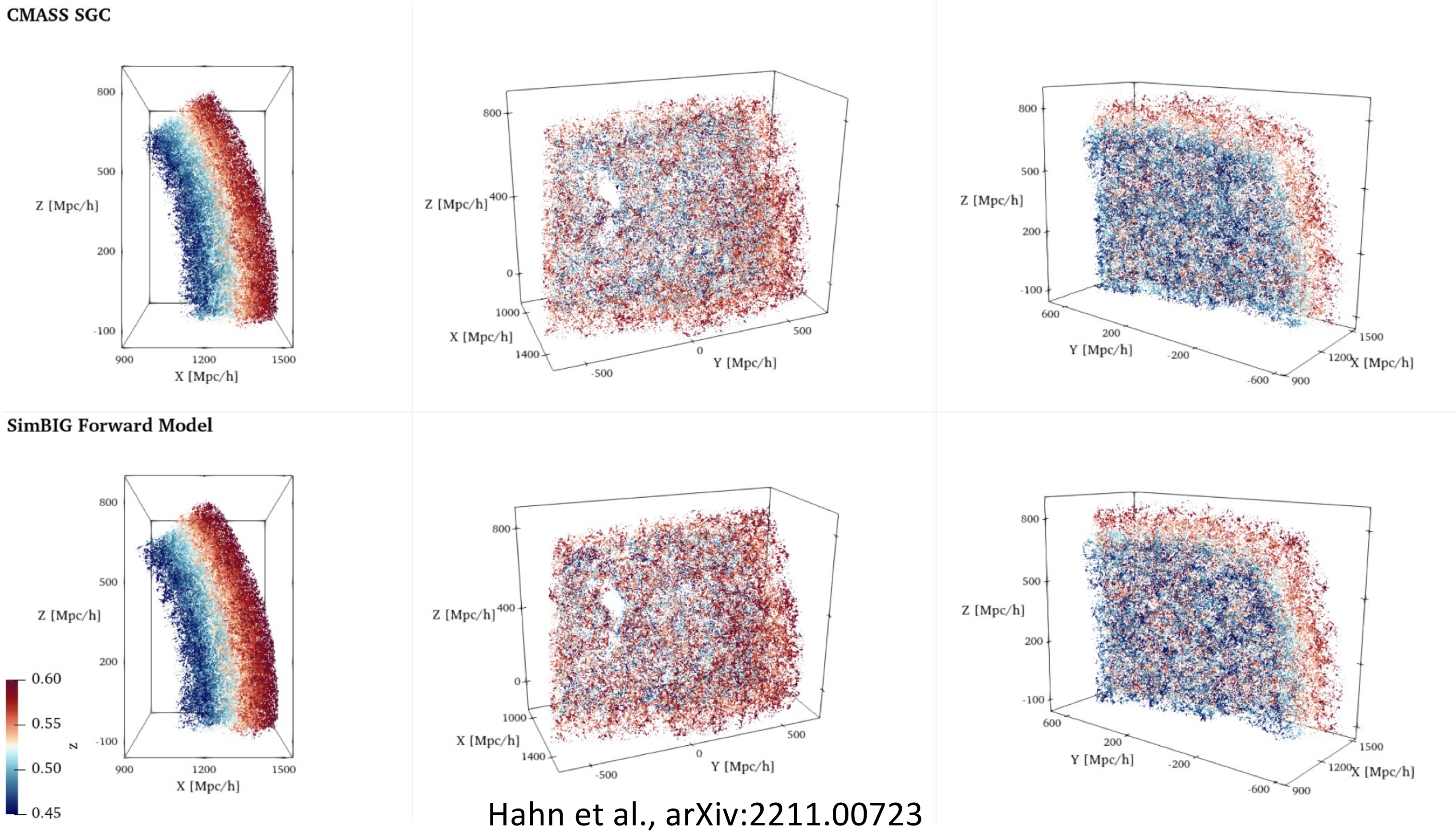
Example: optimal cosmological inference from graphs

Uses message passing graphs to encode *rotational and translational symmetries* and infer cosmological parameters



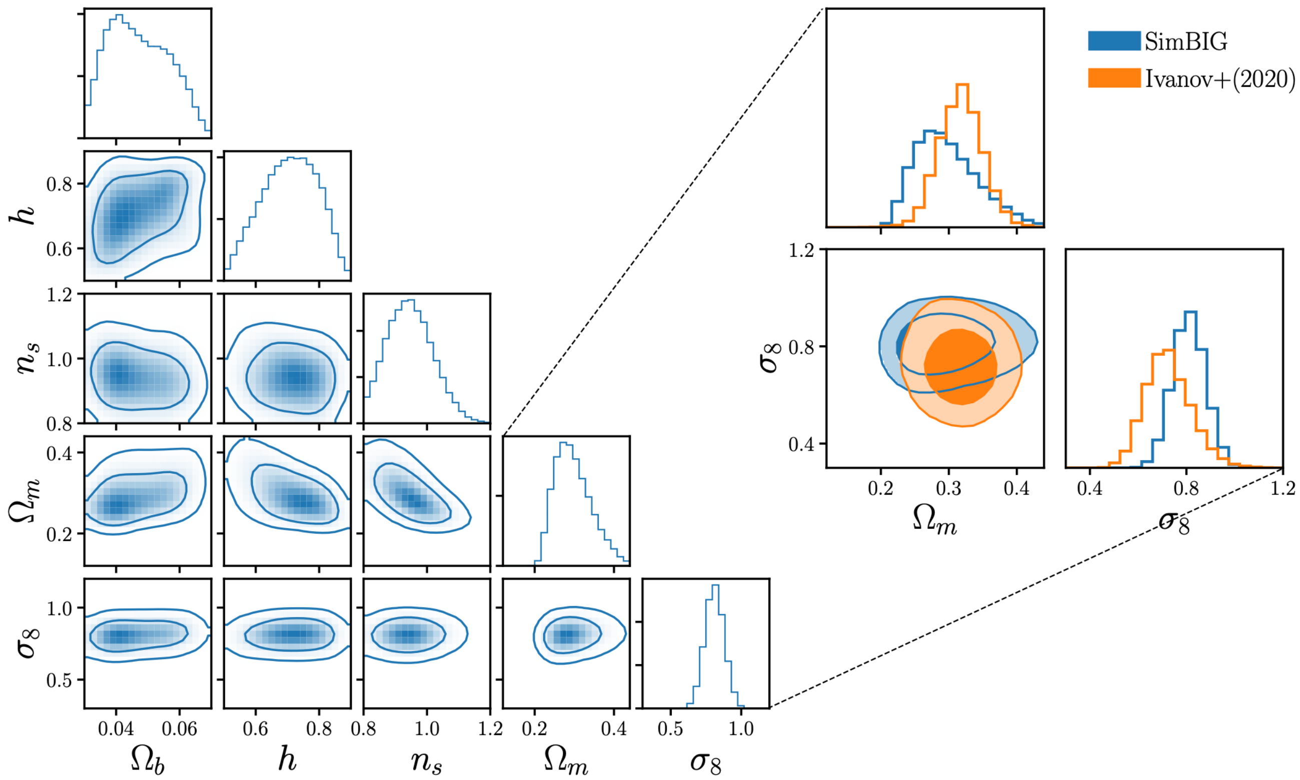
Charnock et al., arXiv: 1802:03537; Makinen et al., arXiv:2207.05202

Analysis of galaxy surveys: now have proof of principle on actual data with SIMBIG



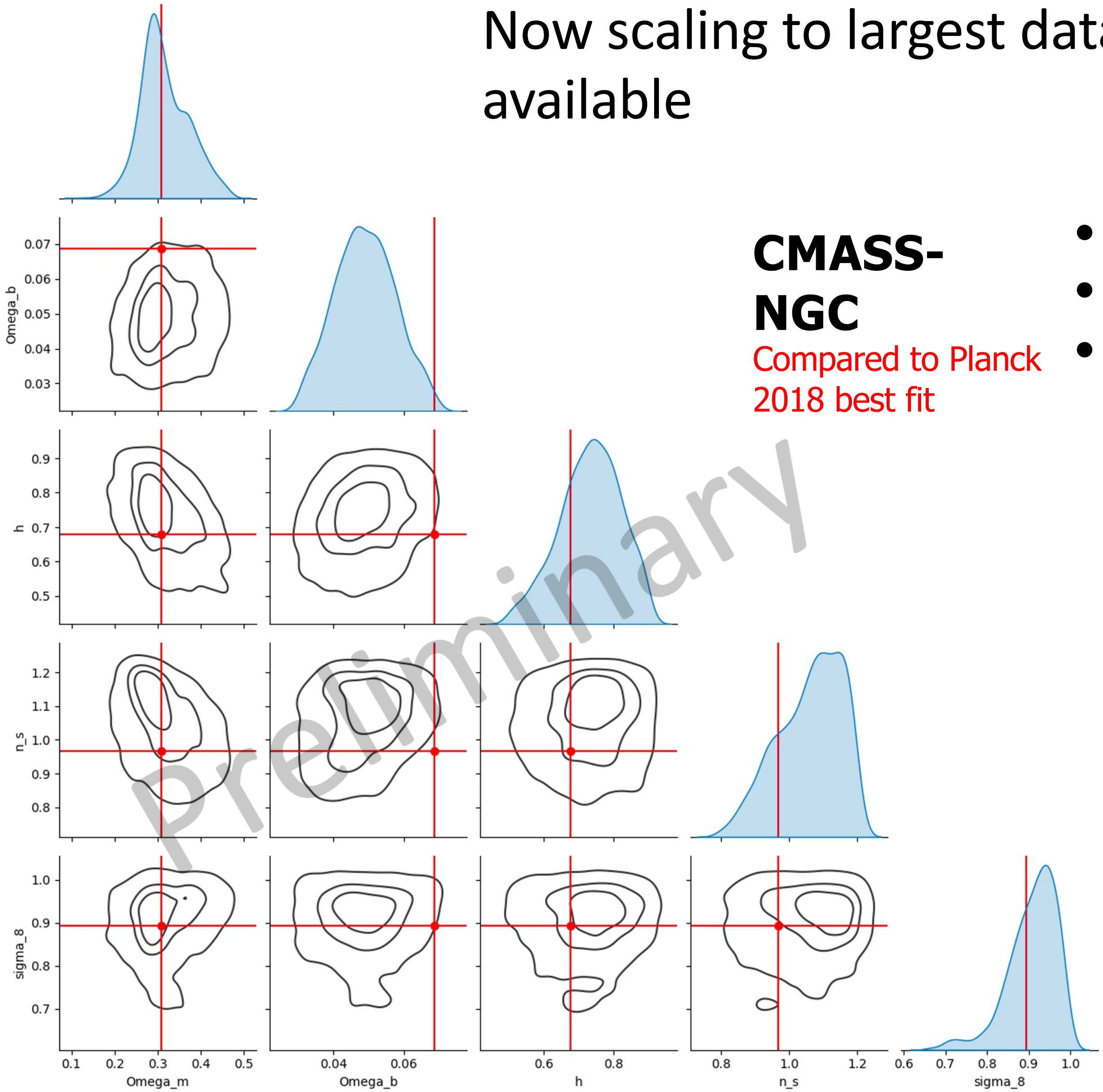
Hahn et al., arXiv:2211.00723

Analysis of galaxy surveys: now have proof of principle on actual data with SIMBIG



Hahn et al., arXiv:2211.00723

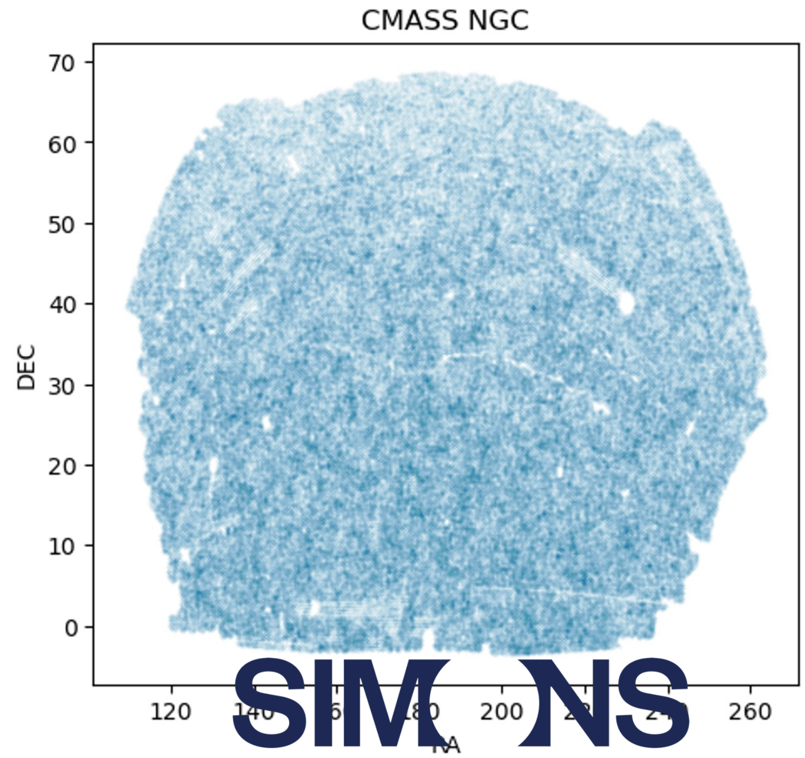
Now scaling to largest data sets currently available



CMASS-NGC

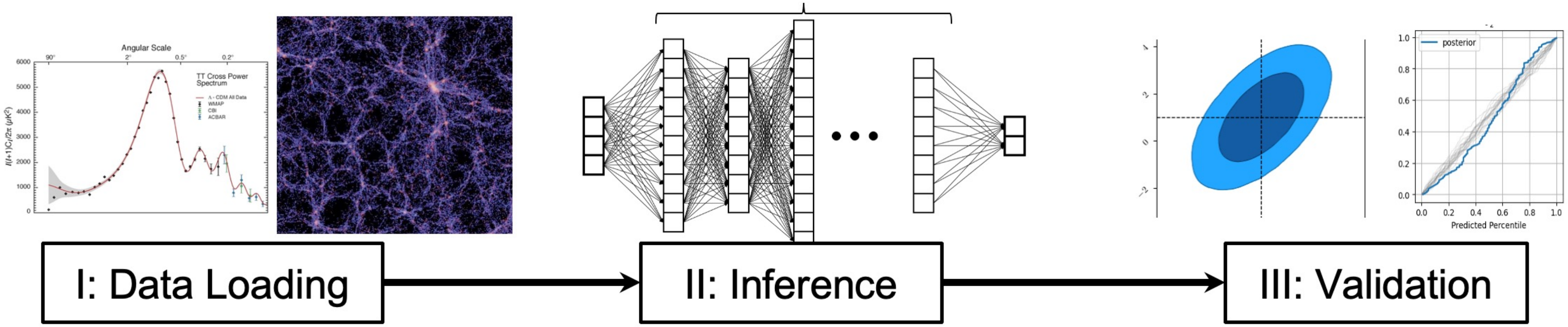
Compared to Planck 2018 best fit

- 2000 pmwd simulations,
- 5 HOD settings per simulation
- Cut at k <= 0.2 $(Mpc/h)^{-1}$



SIMONS FOUNDATION

Learning the Universe Implicit Likelihood Inference (LtU-ILI)



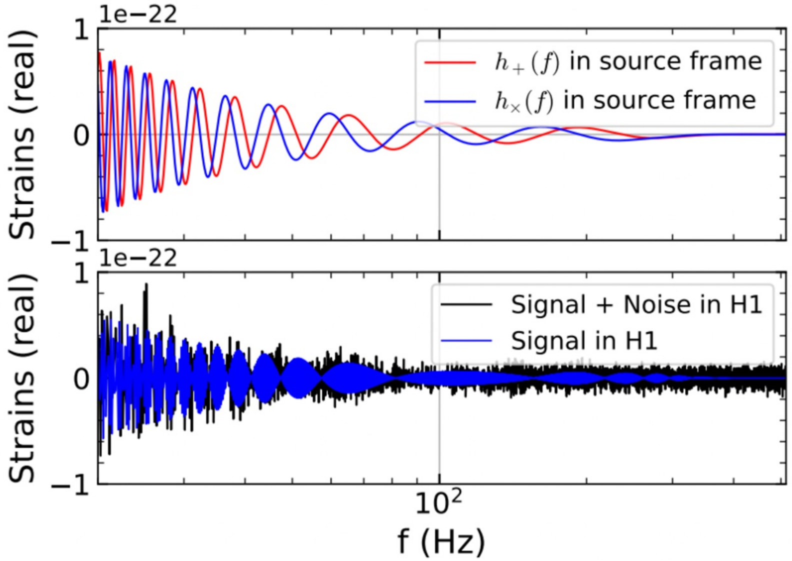
An all-in-one framework for implicit inference in astronomy and cosmology.

- **A lightweight, user-focused design** for both exploratory analysis and production testing.
- Automation of: setup, preprocessing, **model ensembling,** training/testing, and **validation metrics.**
- **All-inclusive of different methods,** including: Neural Posterior/Likelihood/Ratio Estimation and sequential learning
- Easy integration with modern embedding networks such as **CNNs and graph neural networks**
- Combines multiple backends (**sbi**, **pydelfi**, **lampe**) for exhaustive apples-to-apples comparisons
- **Jupyter** and **command-line** interface

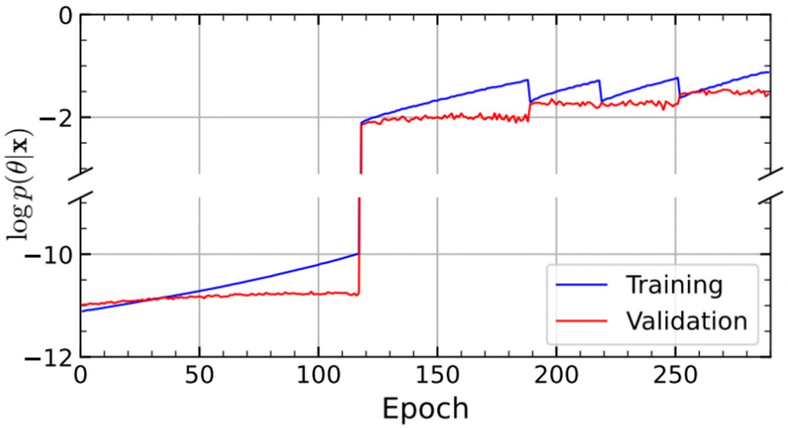
[arXiv:2402.05137](https://arxiv.org/abs/2402.05137)

Matt Ho, Simon Ding, Nicolas Chartier, Chirag Modi, Pablo Lemos, Deaglan Bartlett, Shivam Pandey, Lucia Perez, Guilhem Lavaux, Ludvig Doeser, Lucas Makinen, Carolina Cuesta, Axel Lapel, Hadi Sotoudeh

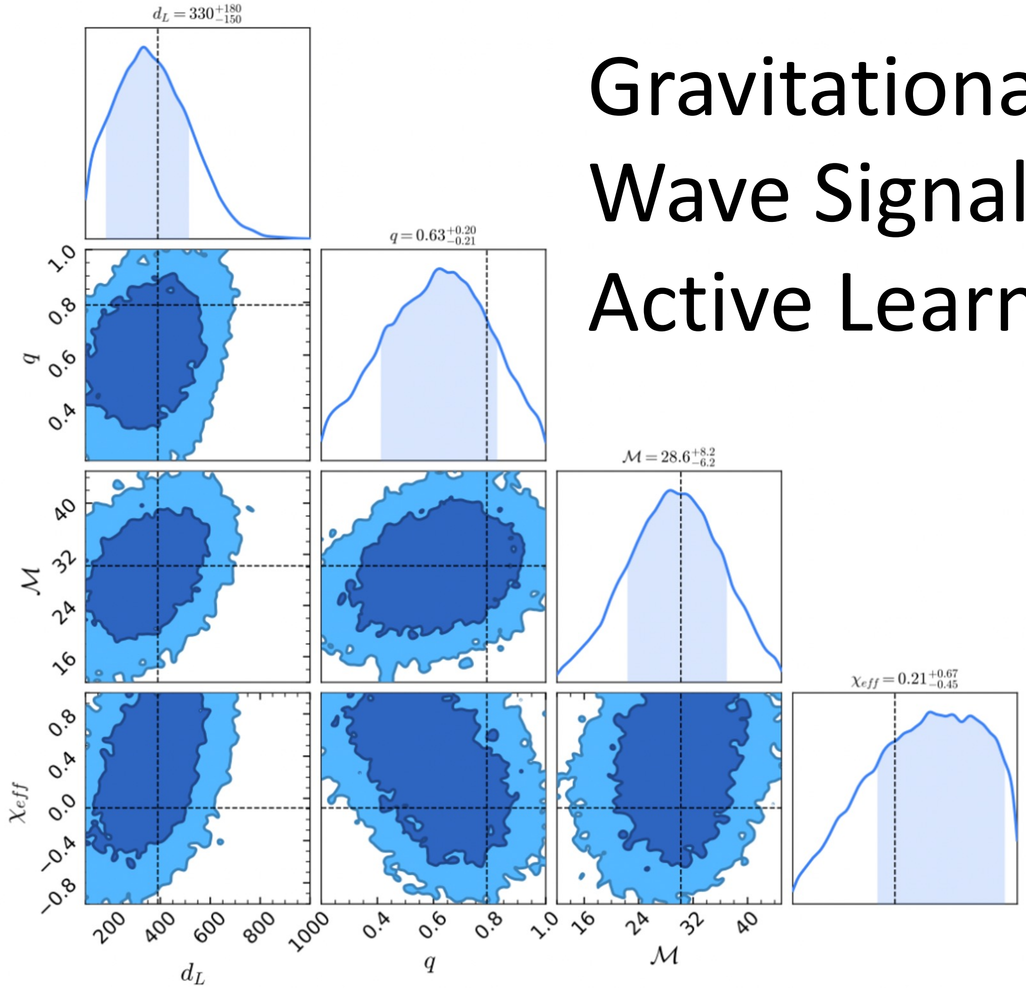
Gravitational Wave Signals with Active Learning



(a) Example of GW signal in source and detector frame.



(b) Multi-round training and validation posterior probability.



(c) Posterior corner plot for SNPE inference.

arXiv:2402.05137

Best part: A full inference pipeline in 5 lines!

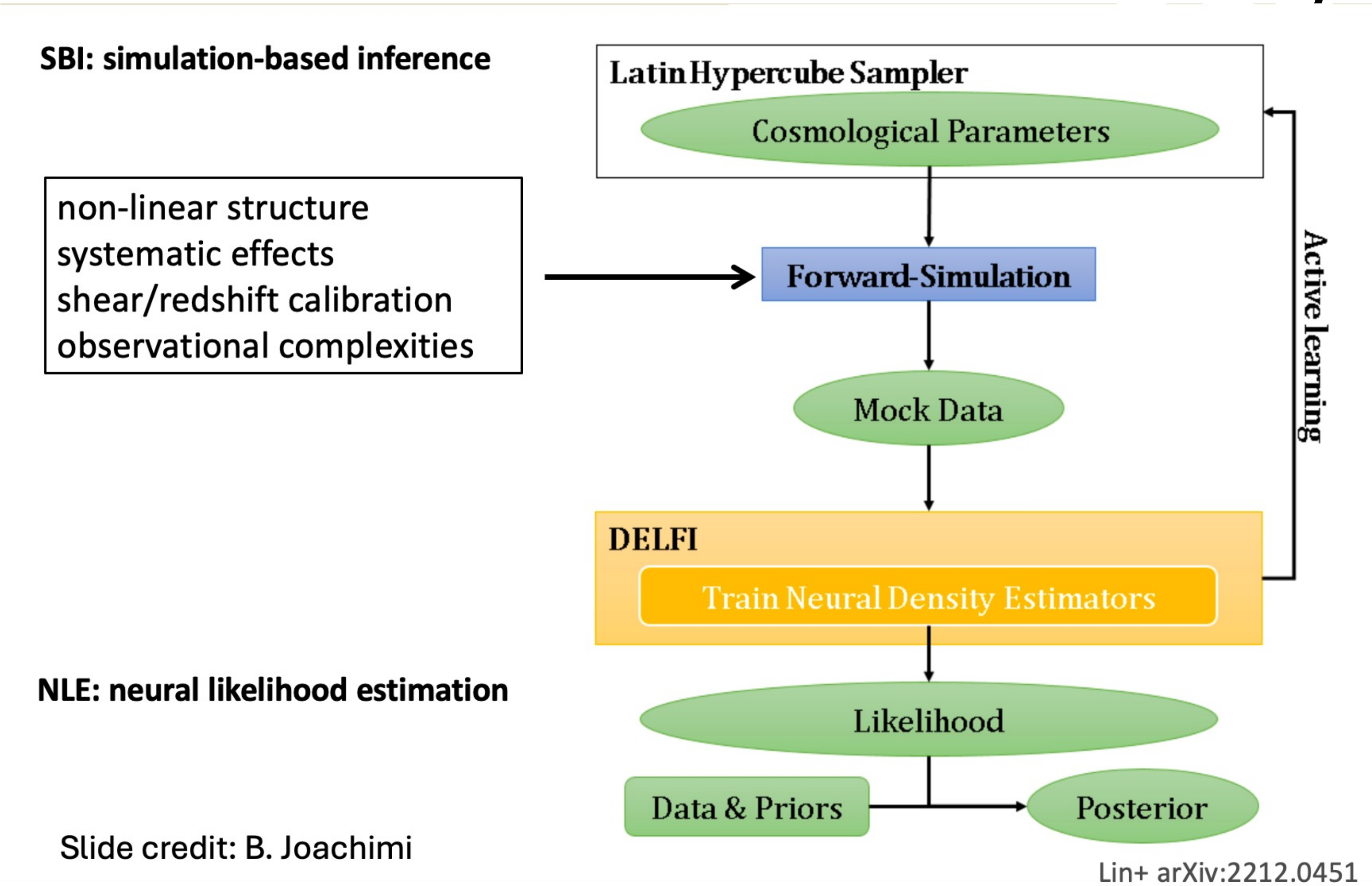
```
...  # Imports

X, Y = load_data()                                          # Load training data and parameters
loader = ili.dataloaders.NumpyLoader(X, Y)                  # Create a data loader

trainer = ili.inference.InferenceRunner.load(
  backend = 'lampe', engine = 'NPE',                        # Choose Neural Posterior Estimation
  prior = ili.utils.Uniform(low=-1, high=1),                # Define a prior
  nets = [ili.utils.load_nde_lampe(model='maf')]            # Define a neural network architecture
)

posterior, _ = trainer(loader)                              # Run training to map data -> parameters

samples = posterior.sample(x[0], (1000,))                   # Generate 1000 samples from the posterior
```

ltu-ili

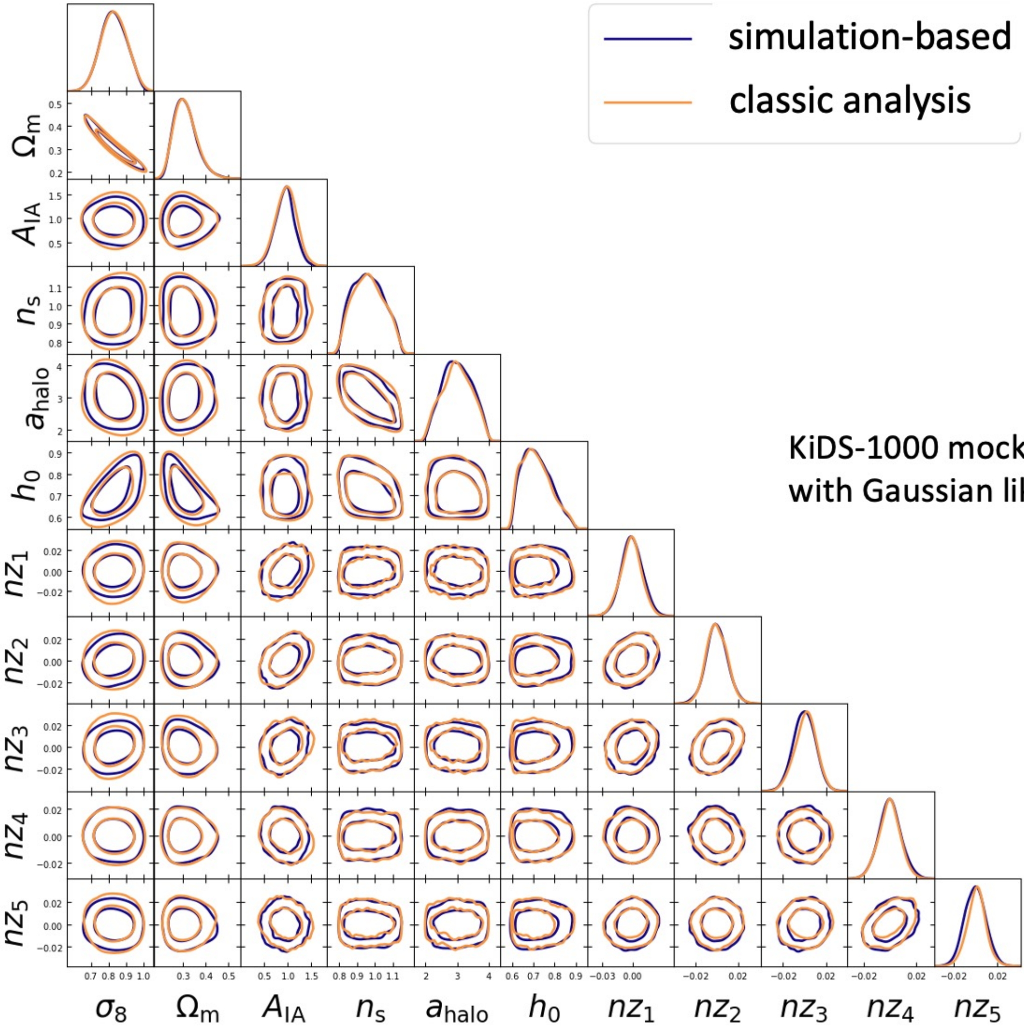
arXiv:2402.05137

https://github.com/maho3/ltu-ili

Real data: KiDS-1000 cosmic shear analysis



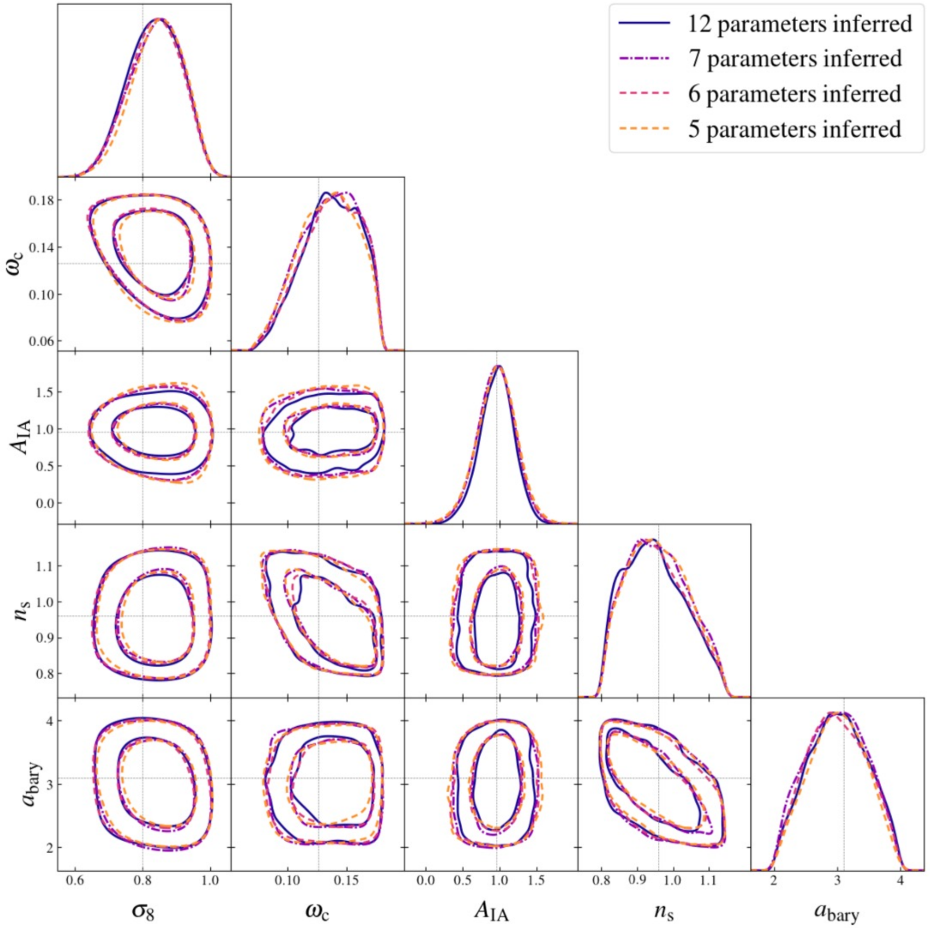
KiDS-1000 cosmic shear analysis



KiDS-1000 mock analysis
with Gaussian likelihood imposed

Lin+ (2023)

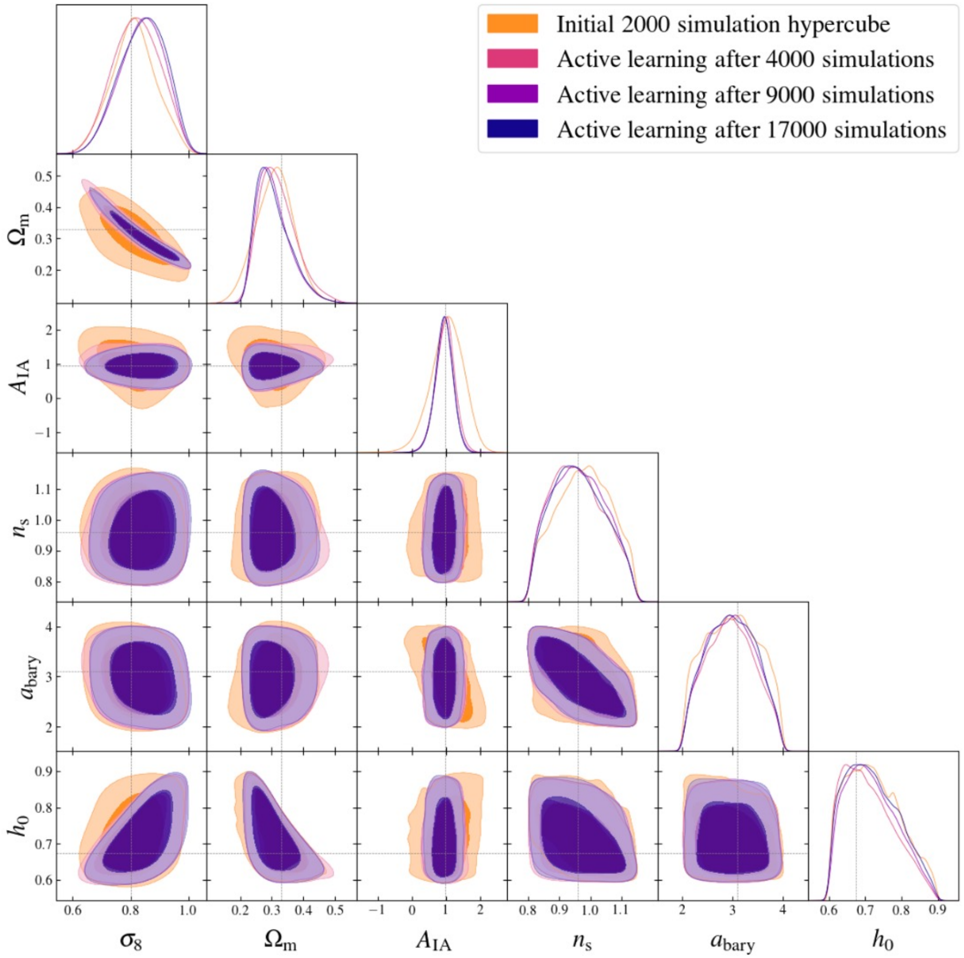
KiDS-1000 cosmic shear analysis: “pre-” marginalization test



0 parameters pre-marginalised
5 parameters pre-marginalised
6 parameters pre-marginalised
7 parameters pre-marginalised

Lin+ (2023)

KiDS-1000 cosmic shear analysis: convergence with number of simulations



Actual data results: see von Wietersheim-Kramsta et al arXiv:2404.15402

Lin+ (2023)

Beyond parameters

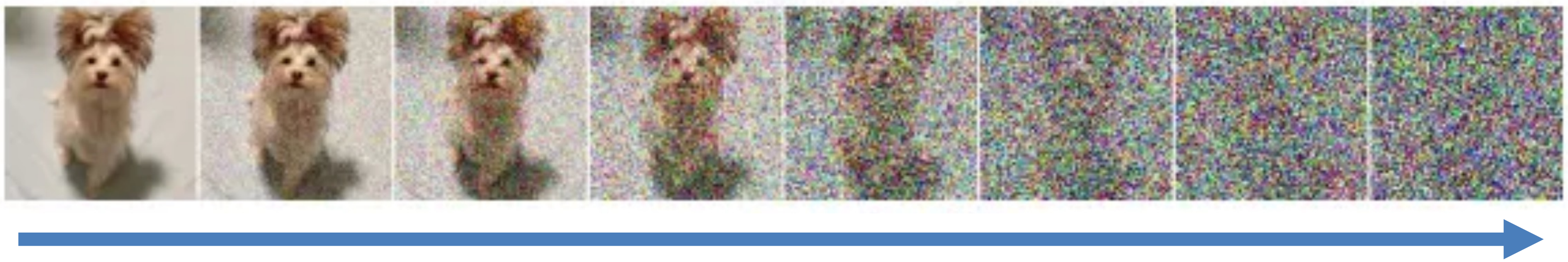
First initial condition reconstructions from fully non-linear simulations with Implicit Inference and conditional score-based diffusion models

R. Legin *et al.*, arxiv:2304.03788

“Score”-based Diffusion: Training

- Consider a random walk of images

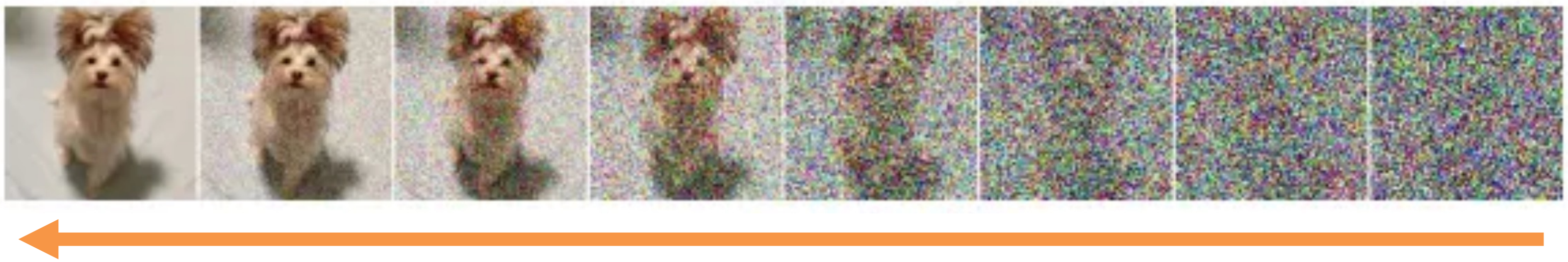
- Initialise with training example as initial condition
- At every step learn a denoiser (the “score”)
- Add Gaussian noise at every step
- Central limit theorem: this has an Gaussian *attractor*



Song et al 2021, +++

Score”-based Diffusion: Generation

- Use trained denoiser to solve a series of inference problems to go from Gaussian noise back to a sample of the initial conditions
- If the number of steps is large enough, each step is a Gaussian inference problem.
- Train a neural network on simulations to learn the posterior mean for each of these steps



Song et al 2021, +++

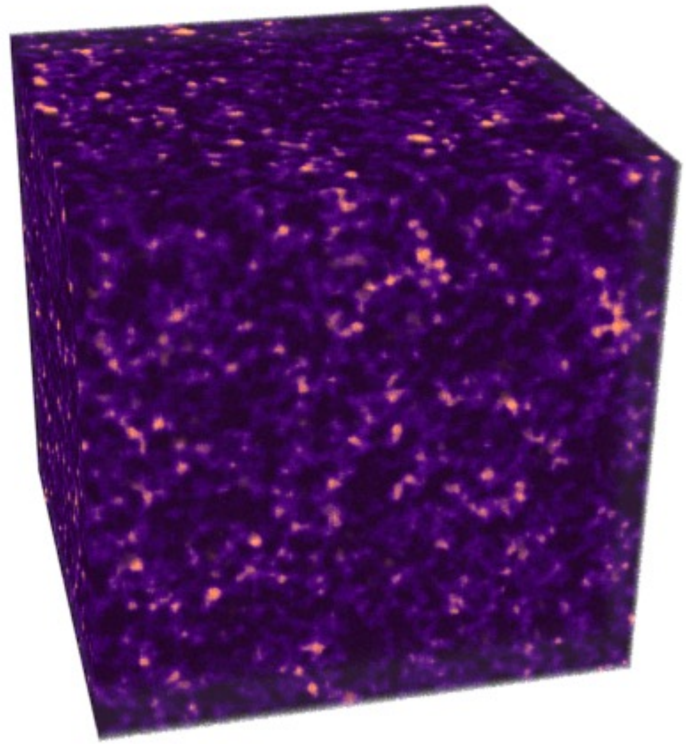
Train full n-body dynamics

- **QUIJOTE:** Largest release of N-body simulation data to date
  - 43,100 full GADGET 3 simulations (1 Gpc)^3, 512^3 or 1024^3 particles
  - ~1 PB of data
- Goal: quantify statistics information content of non-Gaussian non-linear density field about cosmological parameters
- Includes full dark matter snapshots, halo and void catalogues, and many pre-computed statistics.

Villaescusa-Navarro et al, **arXiv:1909.05273**

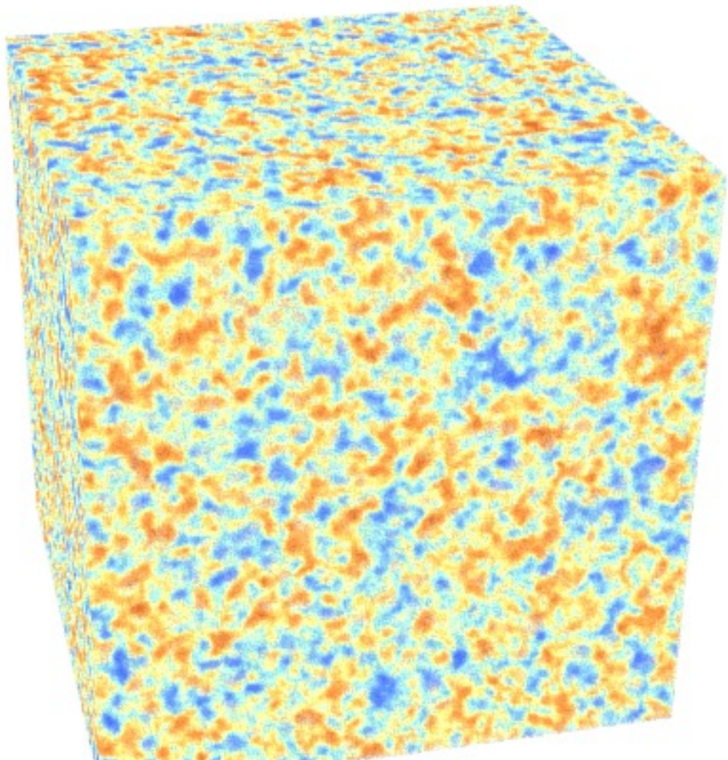
First full-field inference of initial conditions from fully non-linear density field

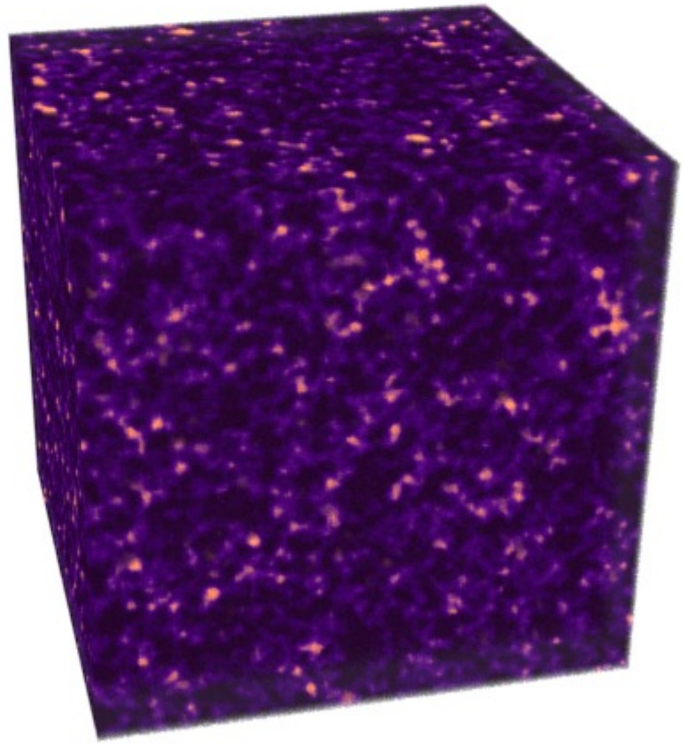
- 1 $(Gpc)^3$ GADGET 1024^3 particle simulation at z=0
- Binned on 128^3 grid
- Resolution 8 Mpc/h



R. Legin *et al.*, arxiv:2304.03788

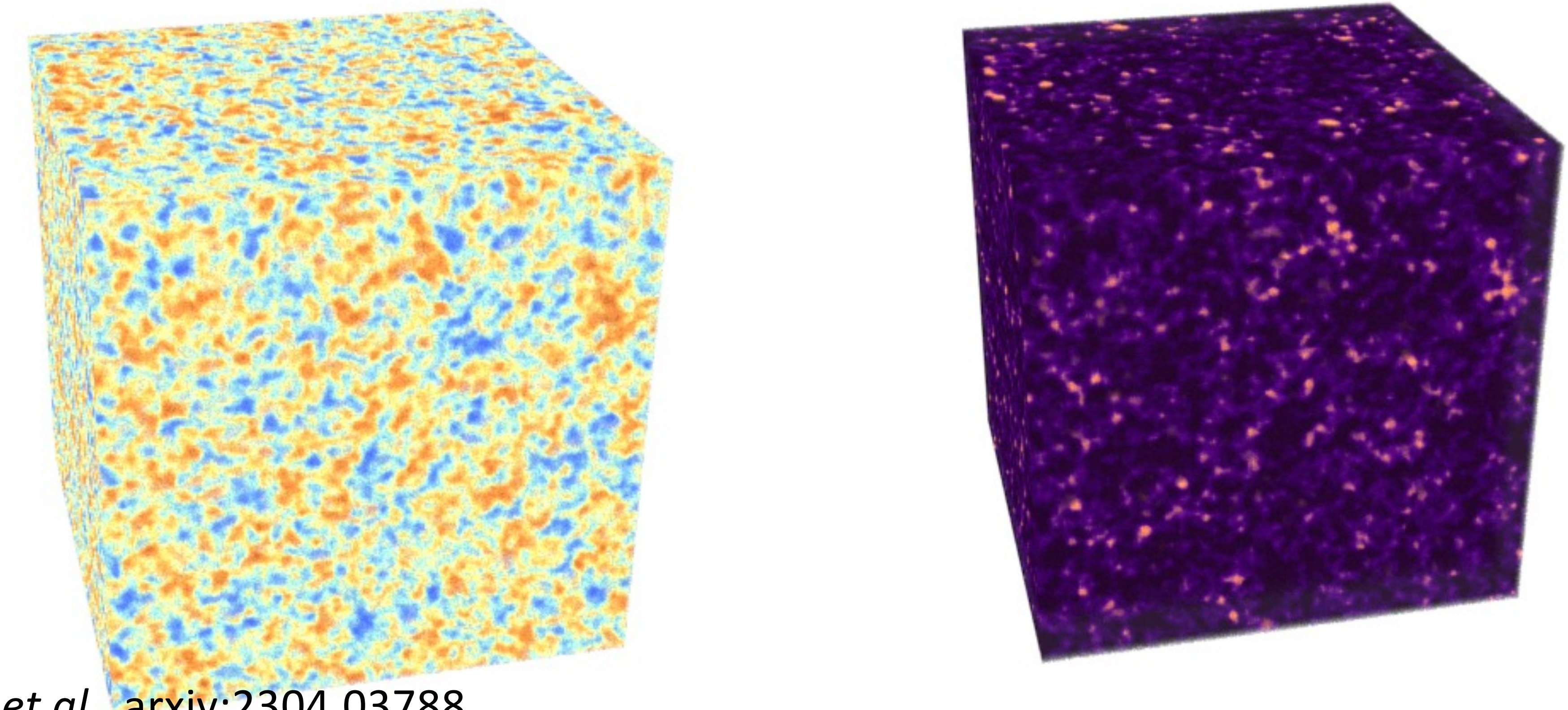
First full-field inference of initial conditions from fully non-linear density field





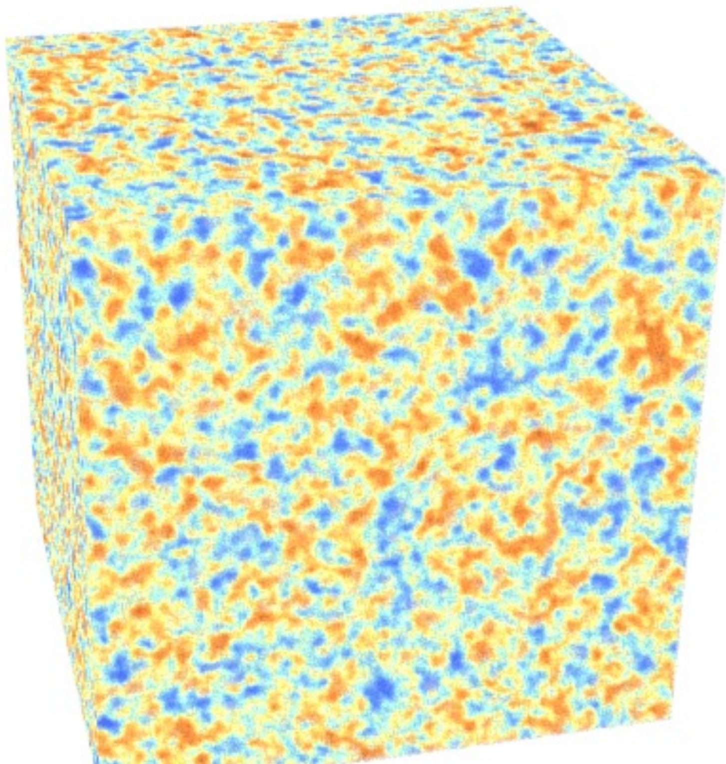
R. Legin *et al.*, arxiv:2304.03788

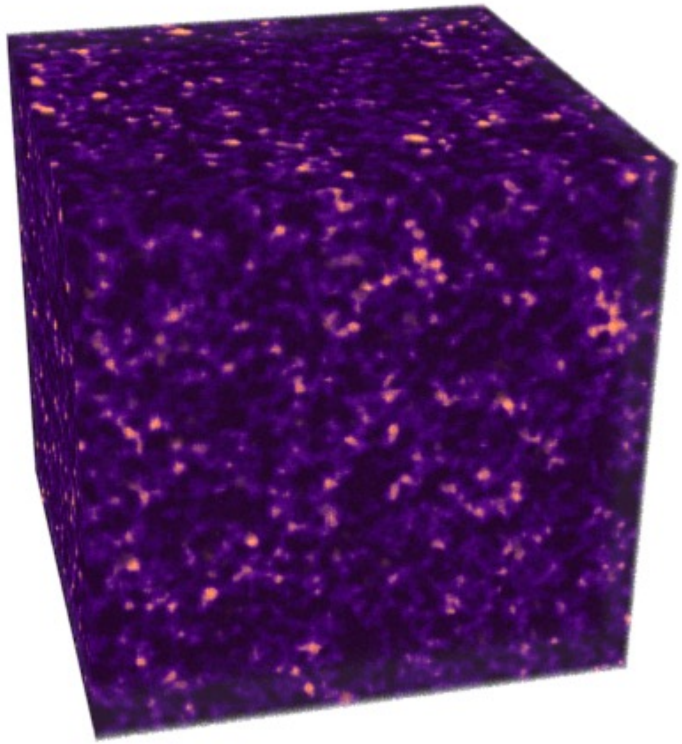
First full-field inference of initial conditions from fully non-linear density field



R. Legin *et al.*, arxiv:2304.03788

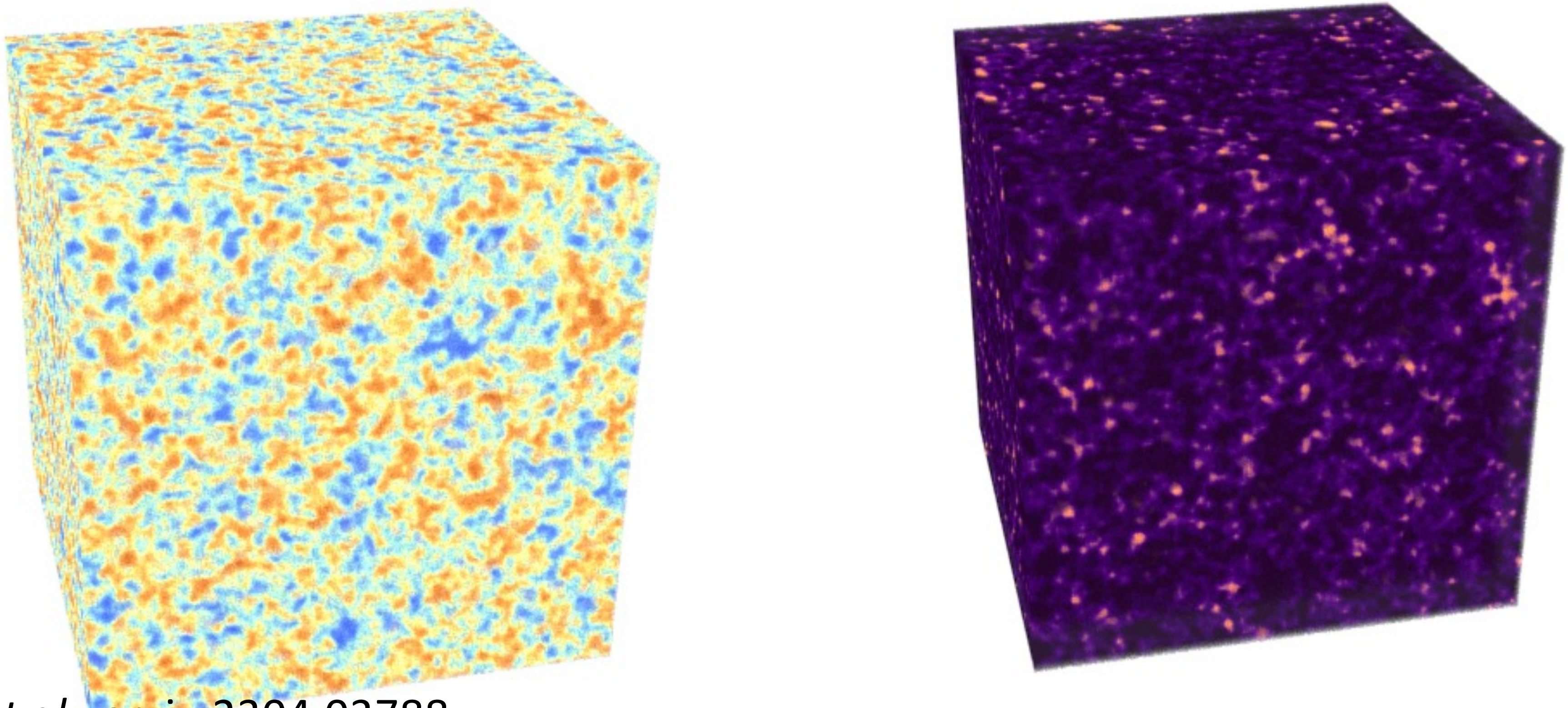
First full-field inference of initial conditions from fully non-linear density field





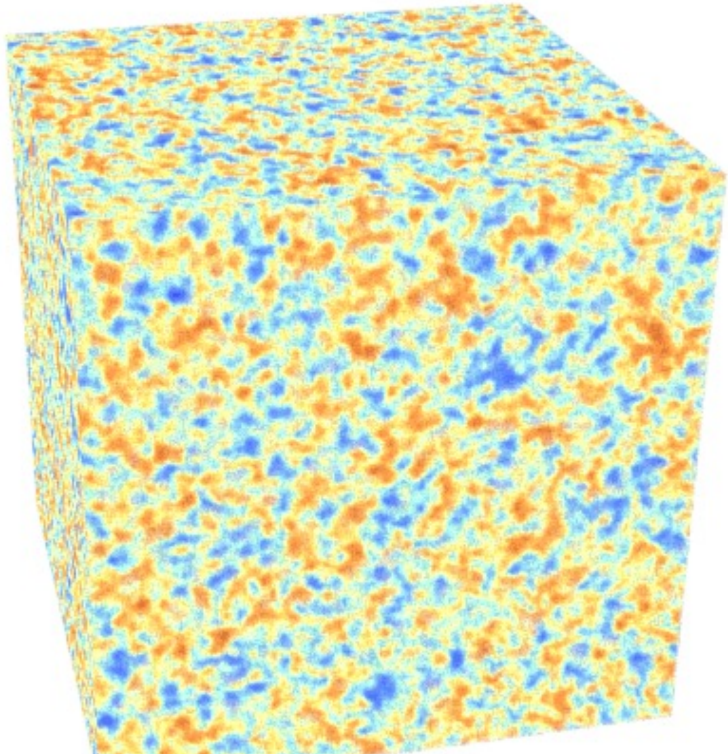
R. Legin *et al.*, arxiv:2304.03788

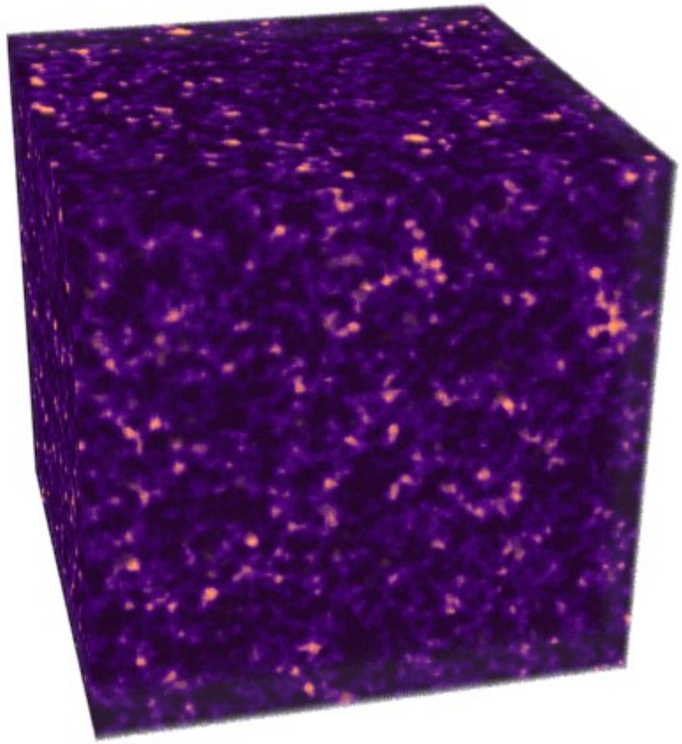
First full-field inference of initial conditions from fully non-linear density field



R. Legin *et al.*, arxiv:2304.03788

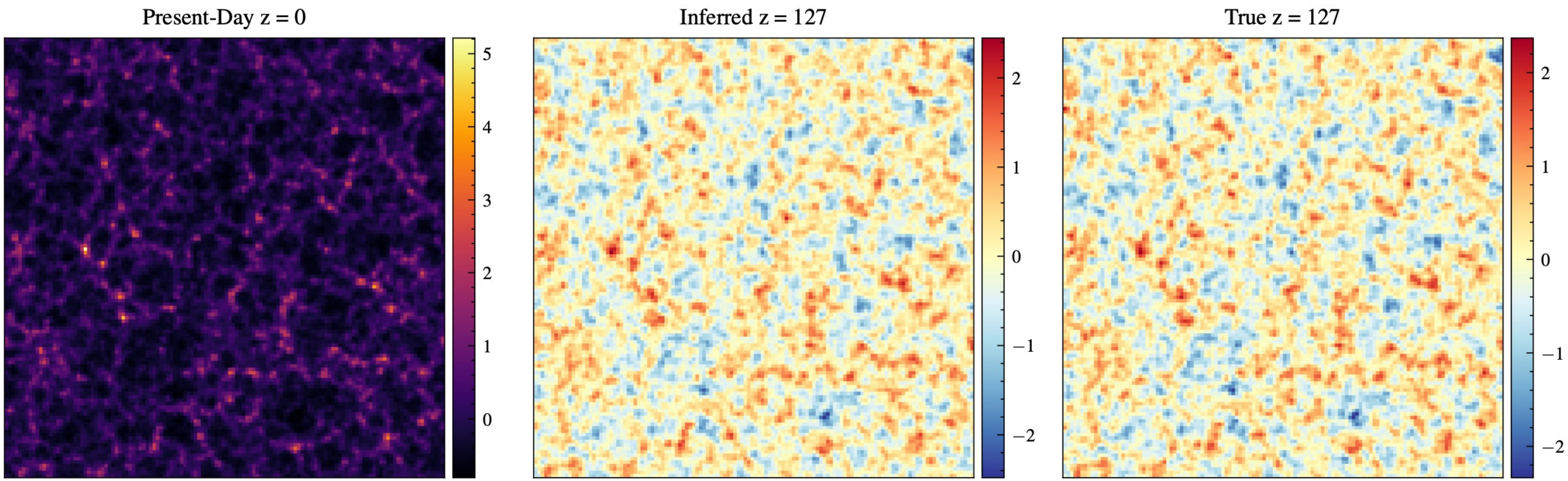
First full-field inference of initial conditions from fully non-linear density field





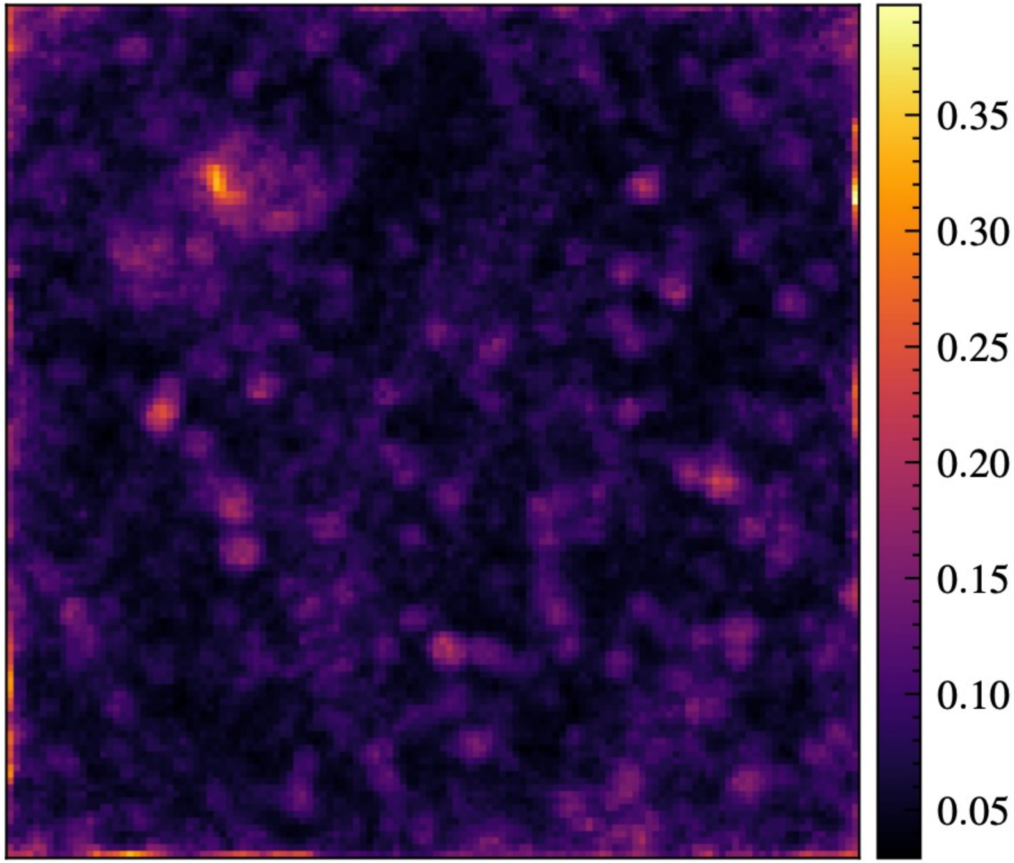
R. Legin *et al.*, arxiv:2304.03788

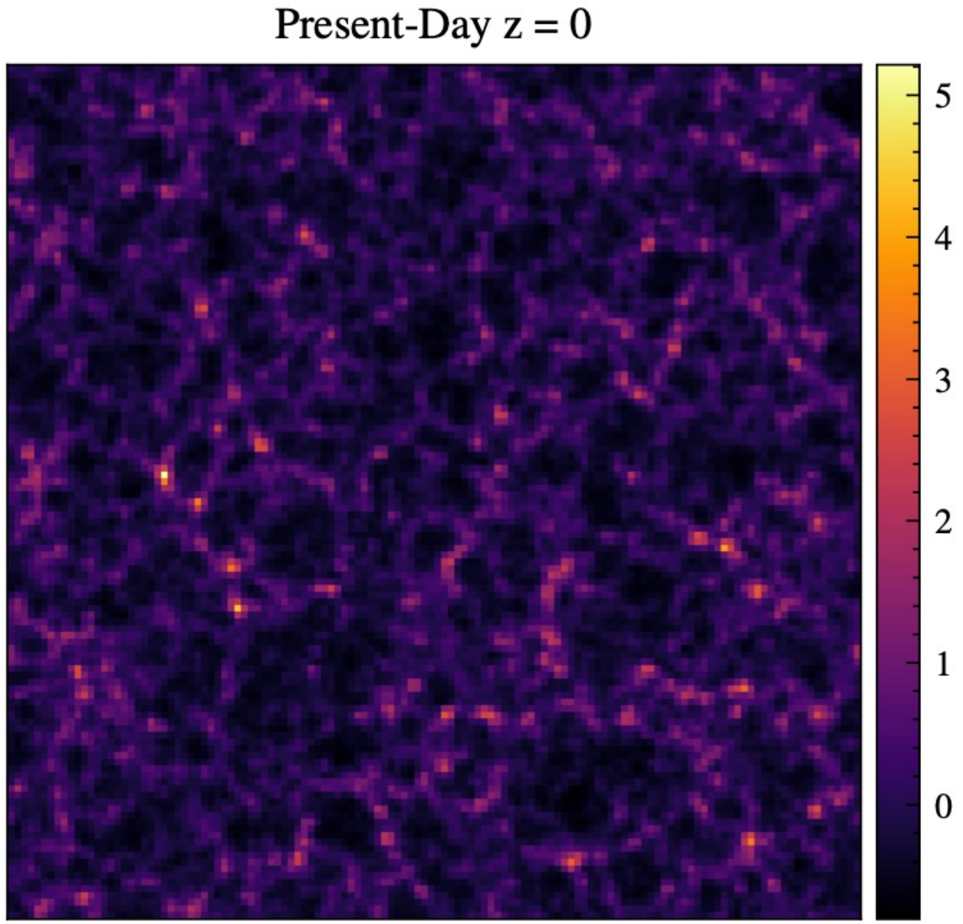
Faithful reconstruction…



R. Legin *et al.*, arxiv:2304.03788

… including uncertainties (posterior variance)





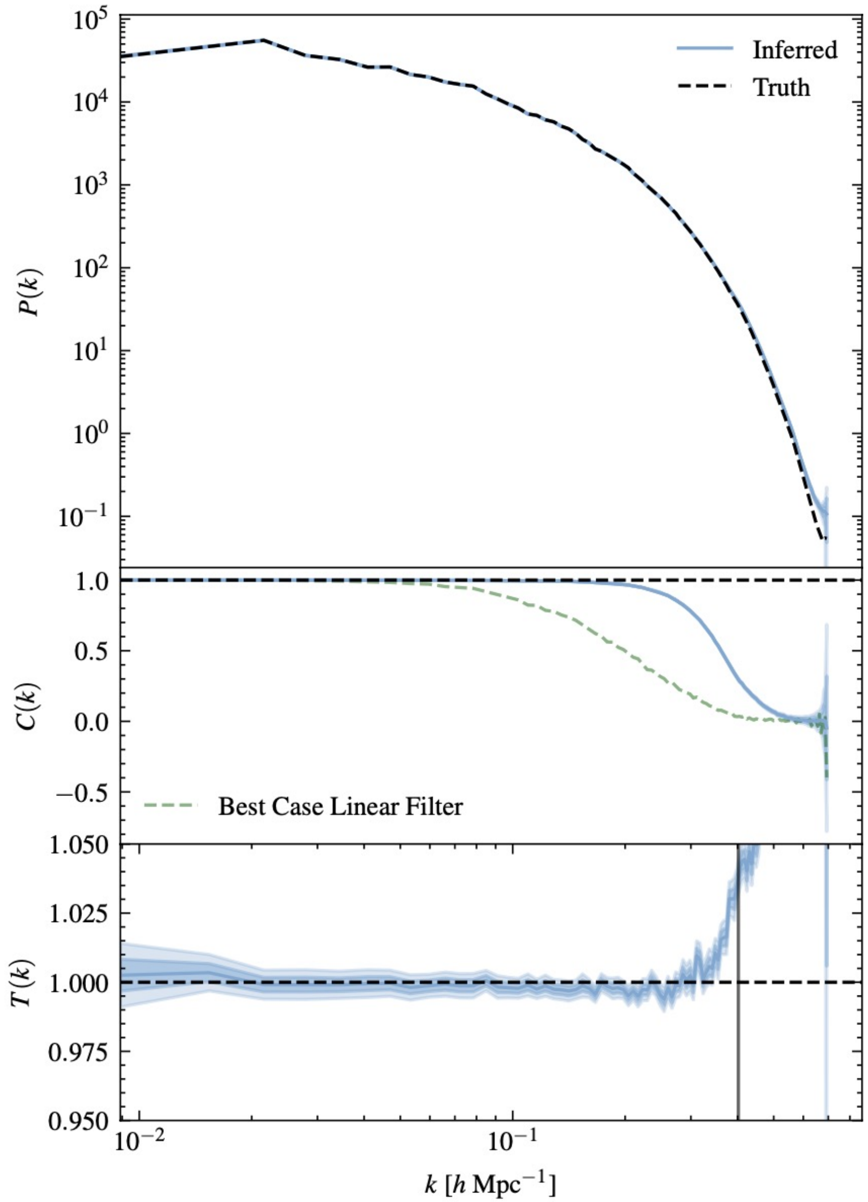
R. Legin *et al.*, arxiv:2304.03788

Accurate reconstructions

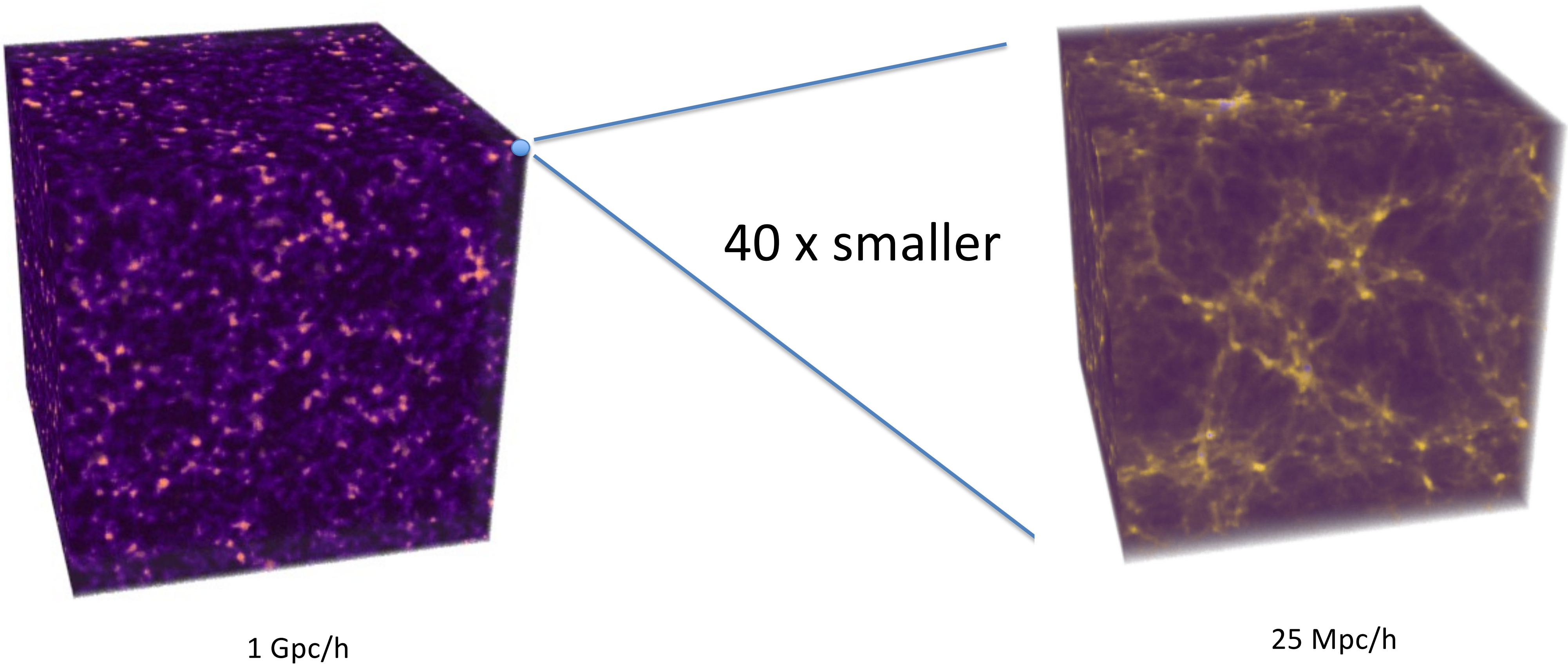
Points to note:

- full non-linear gravity
- No need for differentiability of the computational model

R. Legin *et al.*, arxiv:2304.03788

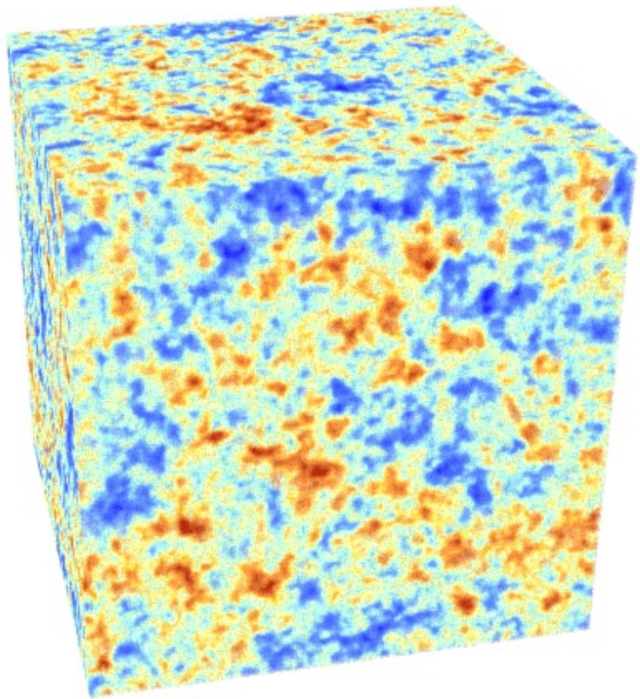


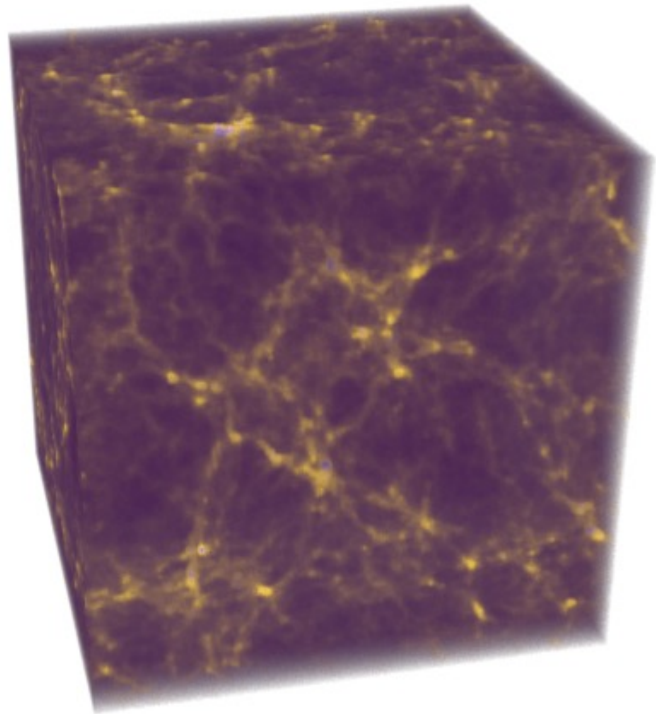
Going even more non-linear



Extremely non-linear regime cosmological ICs

Comparable accuracy as large-scale result!

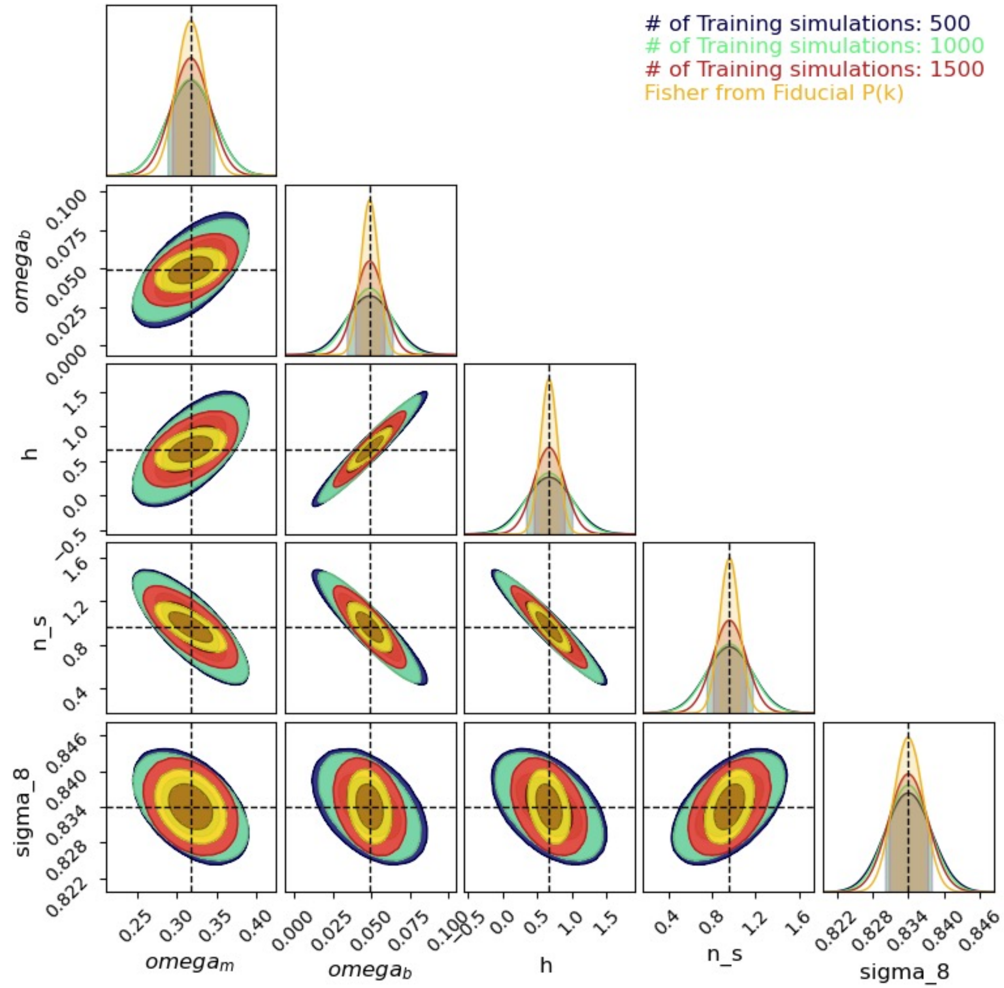




How many simulations are enough?

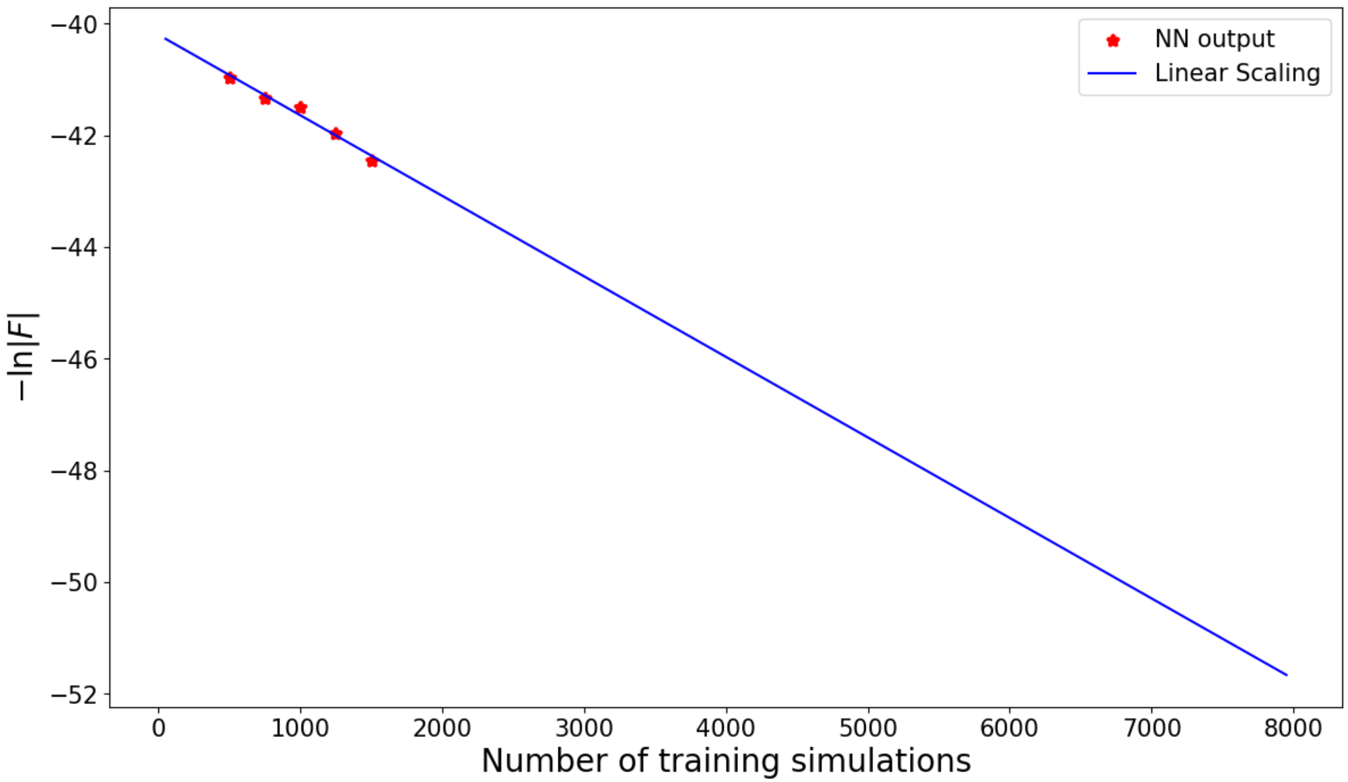
- How do we know if inference is limited by the number of training simulations?

- Combine training simulations with a set of sims in the neighborhood of a fiducial point
- Simple experiment: power spectrum inference from Quijote sims



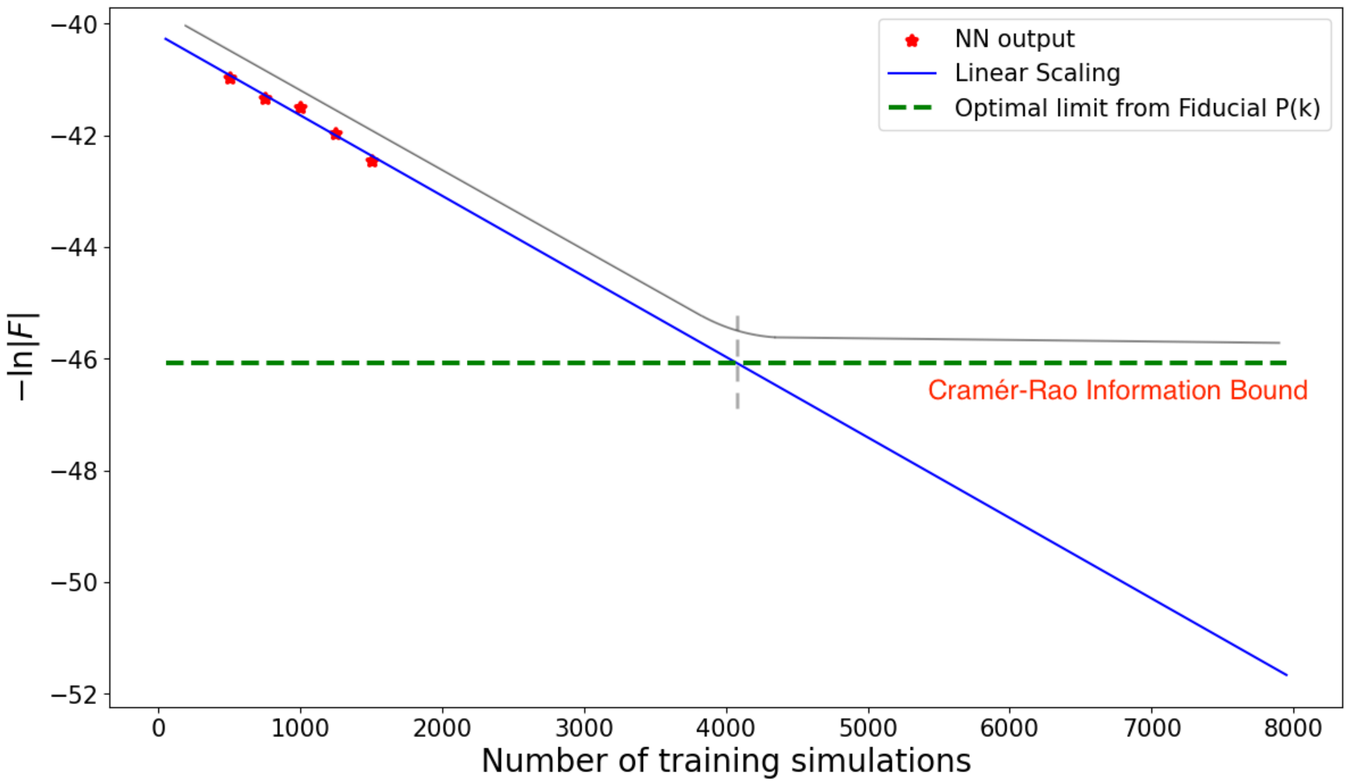
A. Bairagi et al., in prep.

Cosmological neural scaling law



A. Bairagi et al., in prep.

Cosmological neural scaling law



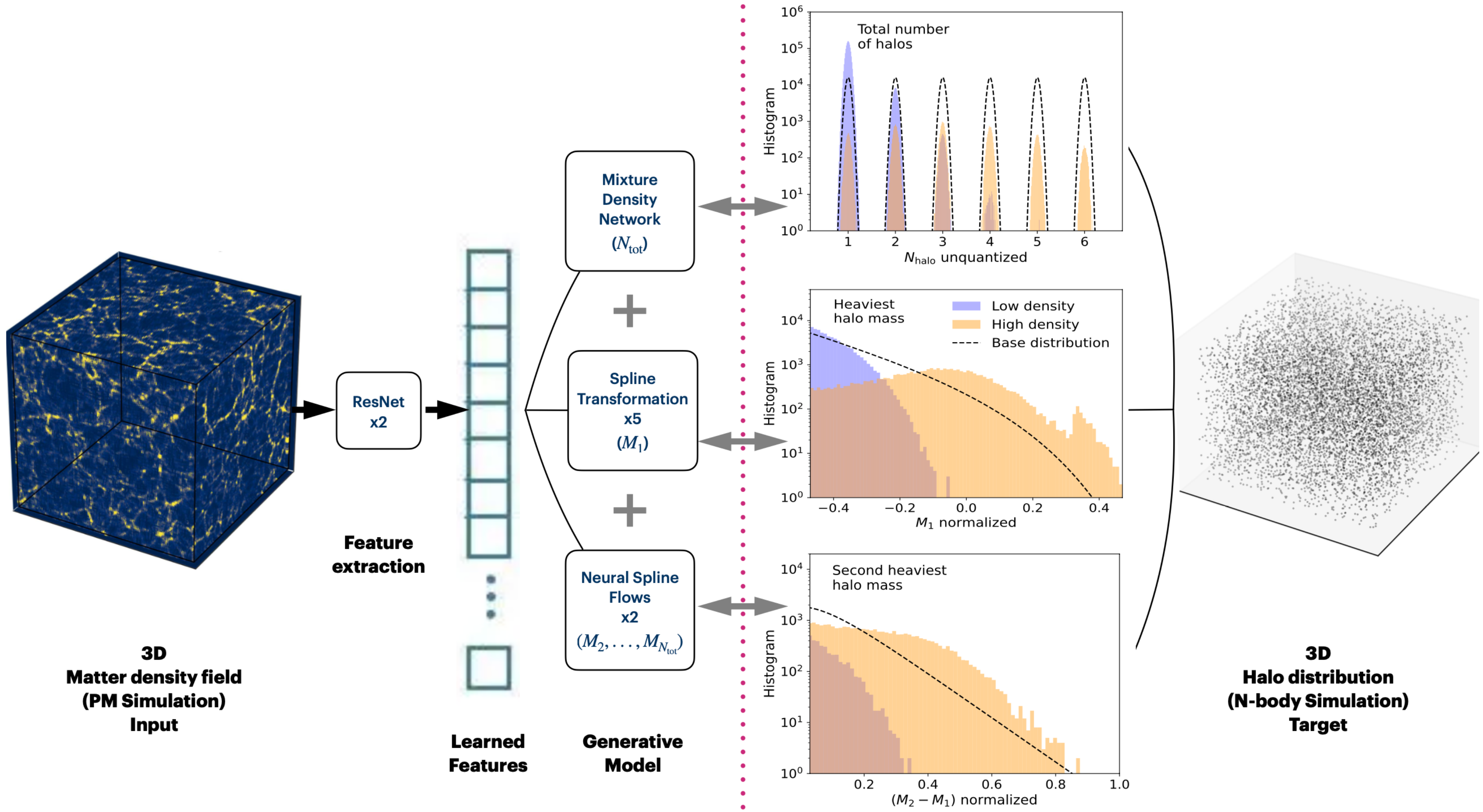
A. Bairagi et al., in prep.

How will we get all the simulations?

- Use/generate simulation corpora (Quijote and Abacus n-body sims, CAMELS hydrosimulations, …)

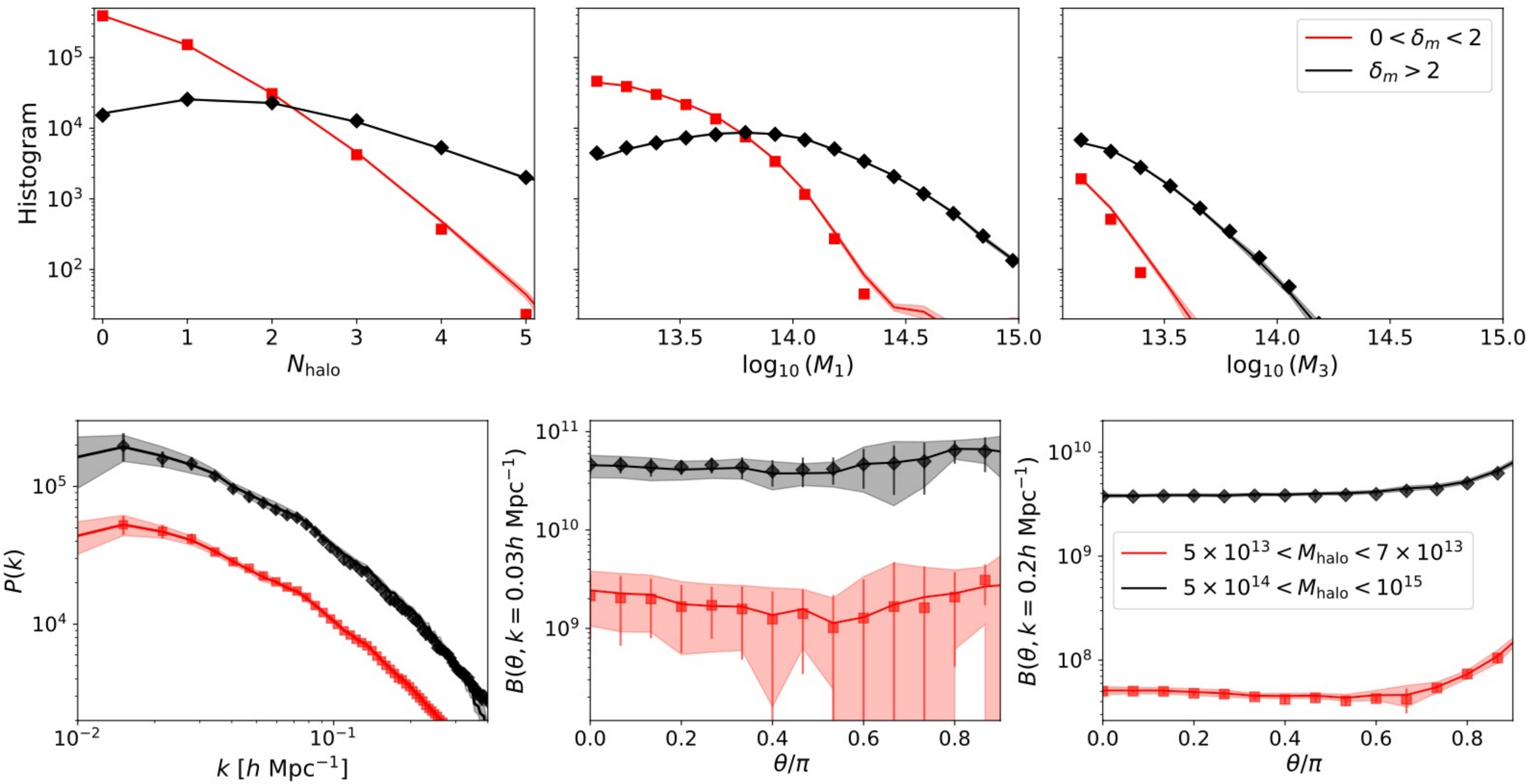
- Emulators! (e.g. Ramanah et al 2019, Jamieson et al 2023,…)

From fast PM to halos with CHARM



Pandey, Modi, Wandelt, Lavaux 2023, NeurIPS

Works like a CHARM!



Pandey, Modi, Wandelt, Lavaux 2023, NeurIPS

From information to insight

Can we use machine learning to discover

…new models (e.g. symbolic searches)?

…the most important degrees of freedom in data?

Evidence-based Model Comparison using Implicit Inference

$$p(\theta|d) = \frac{p(d|\theta)p(\theta)}{p(d)}$$

Actually, $p(d|M_i)$

Jeffrey & Wandelt, arXiv:2305.11241

Bayesian model comparison

$$\frac{p(M_i|d)}{p(M_j|d)} = \boxed{\frac{p(d|M_i)}{p(d|M_j)}} \frac{p(M_i)}{p(M_j)}$$

Bayes factor ***K***

Jeffrey & Wandelt, arXiv:2305.11241

Bayesian model comparison

Even if likelihood and posterior are explicitly given

- Likelihood can be costly to evaluate
- **Evidence can be hard to compute**

$$P(\theta|d,M) = \frac{P(d|\theta,M)P(\theta|M)}{\boxed{P(d|M)}}$$

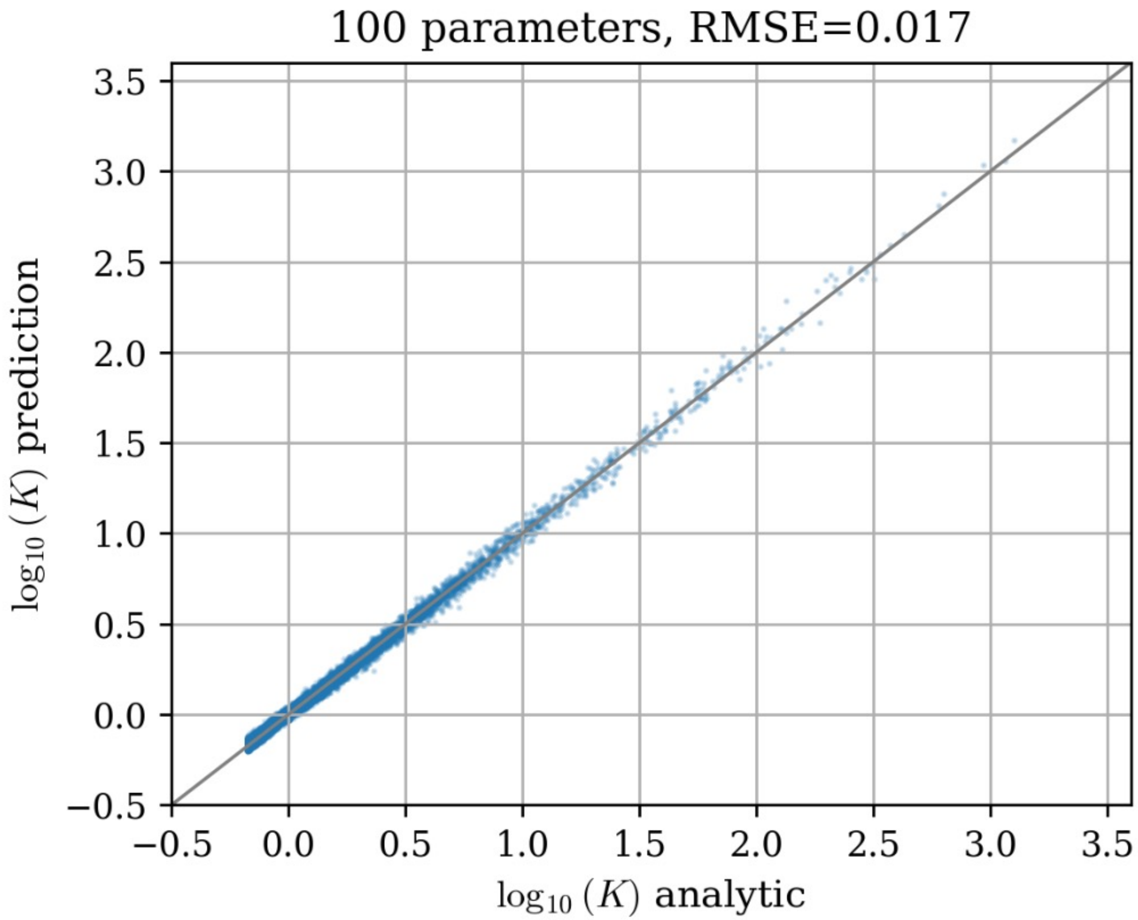
$$\Longrightarrow P(d|M) = \int P(d|\theta,M)P(\theta|M)d\theta$$

Jeffrey & Wandelt, arXiv:2305.11241

Example: evidence ratio with 100 parameters

Evidence Networks trained to compute evidence ratio based only on simulated examples and a custom loss function.

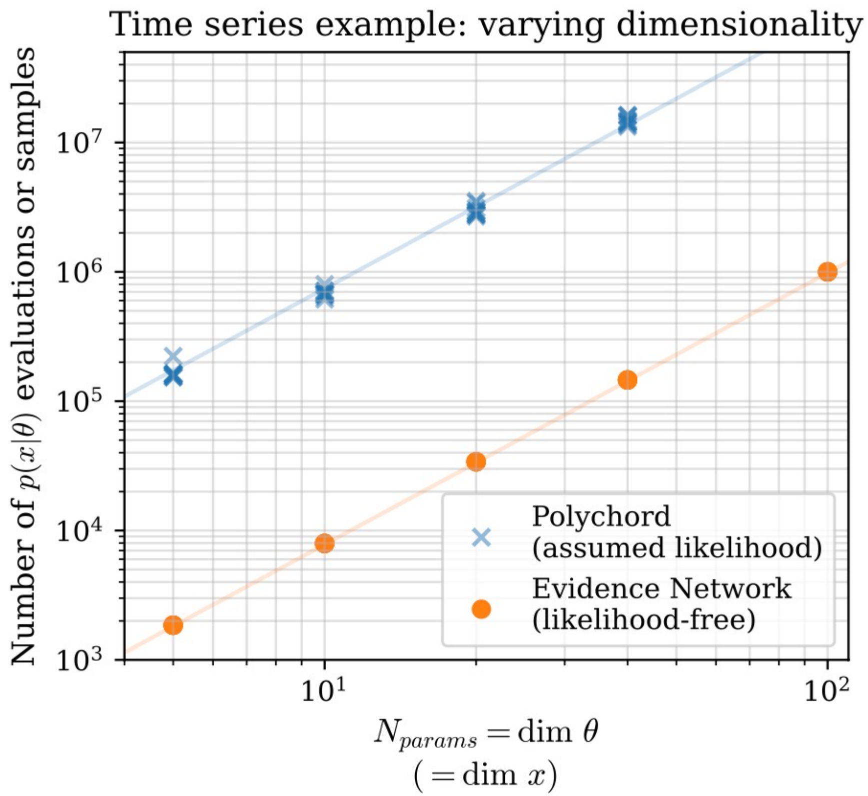
This evidence computation does not explicitly depend on number of parameters!

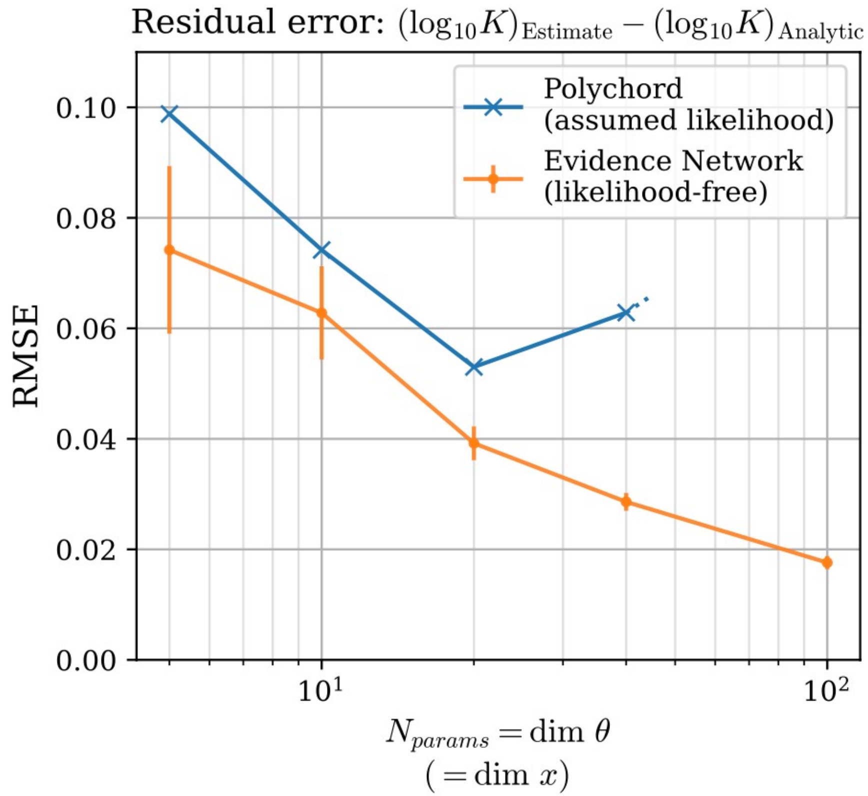


Jeffrey & Wandelt, arXiv:2305.11241

Evidence nets: more accurate and faster than best-of-class nested sampling method

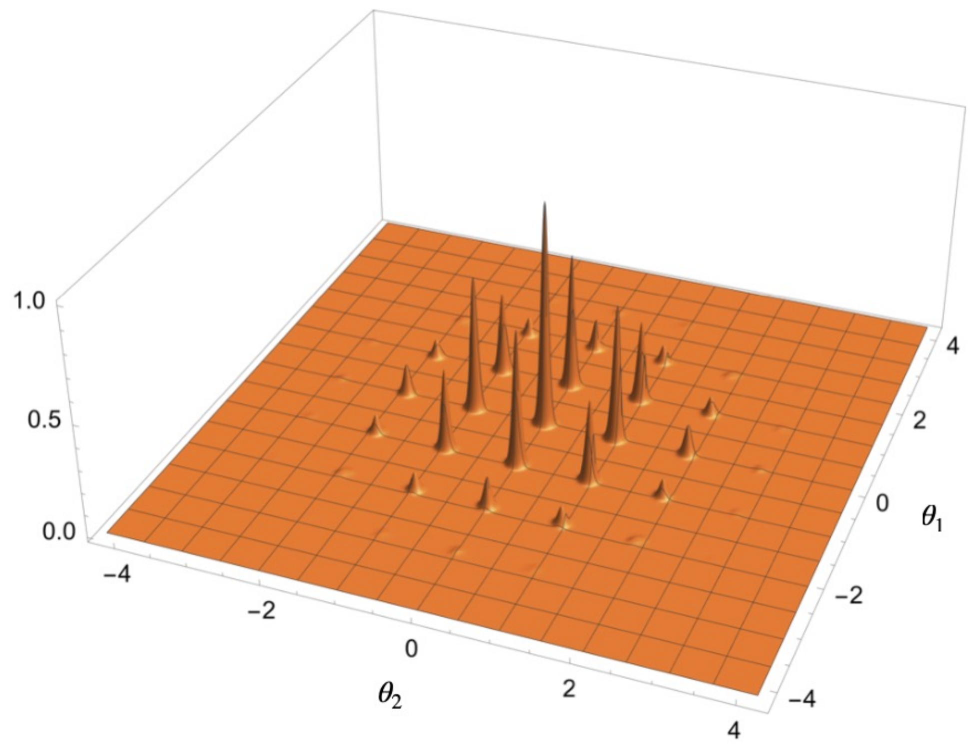
Jeffrey & Wandelt, arXiv:2305.11241

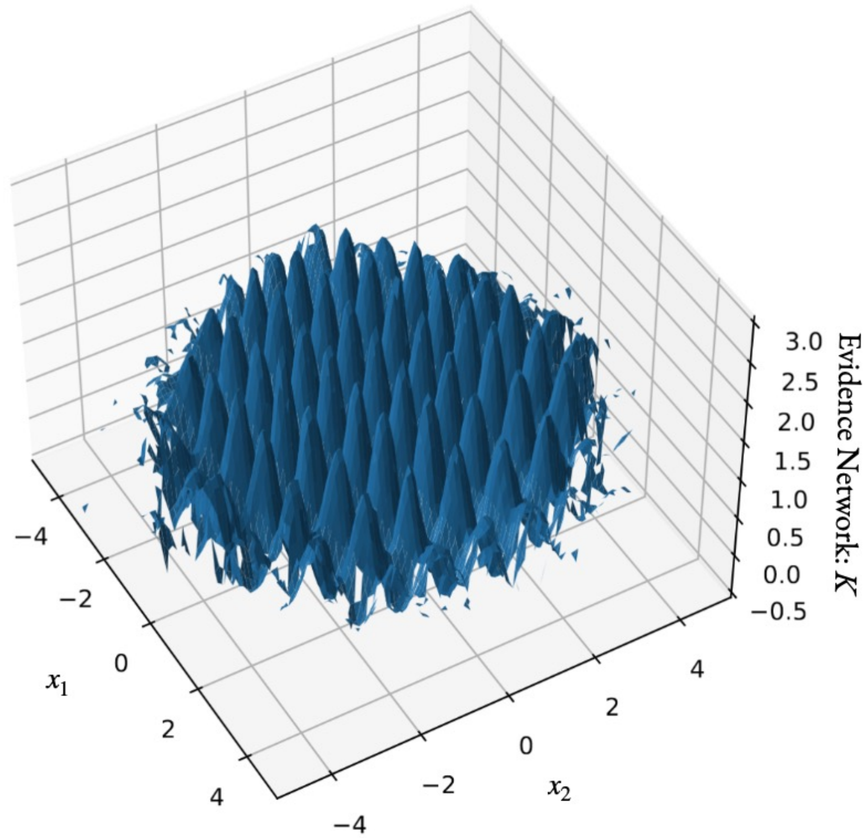




Computational cost of evidence network includes time to generate sims and train. Application to a given data set is nearly instantaneous.

Works on traditionally intractable examples



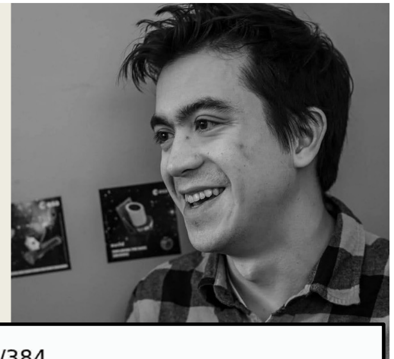


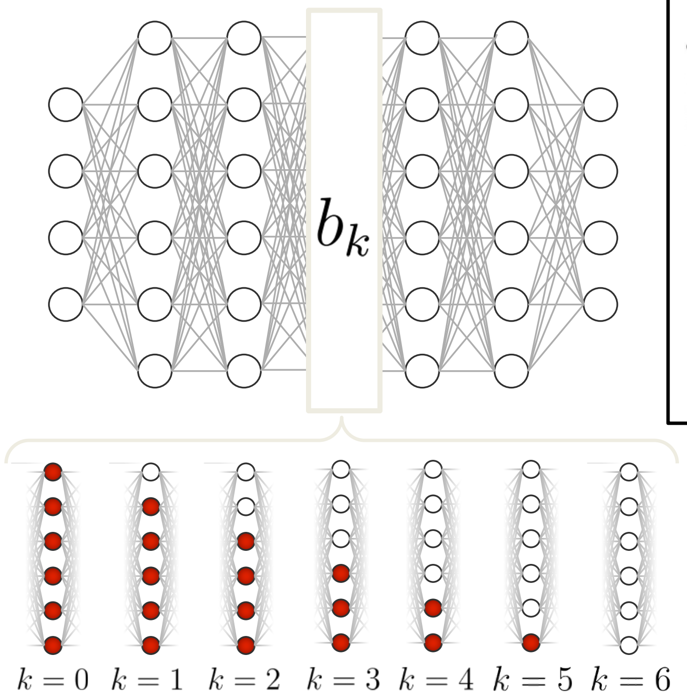
Jeffrey & Wandelt, arXiv:2305.11241

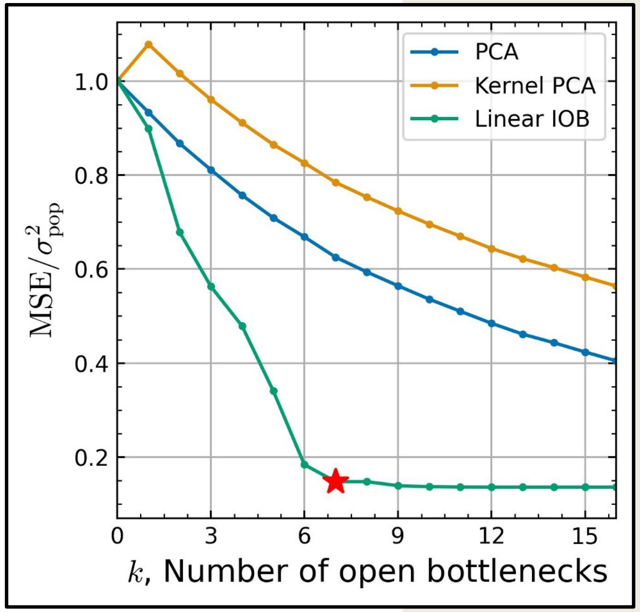
Information Ordered Bottlenecks

Scientific Discovery from Ordered Information Decomposition

Matthew Ho (UvA 1 @ 4:20pm)







$$\mathcal{L} := \sum_{k=0}^{k_{\max}} \mathcal{L}_k$$

Key insights

- Cosmology is no longer data-limited but model-limited.
- ML allows us to recast physics questions as optimization problems.
- AI cannot do it alone: AI + physics/astronomy
- Combine new machine learning methods, fast simulation techniques, and statistical methods
- Need interpretability for insight and discovery