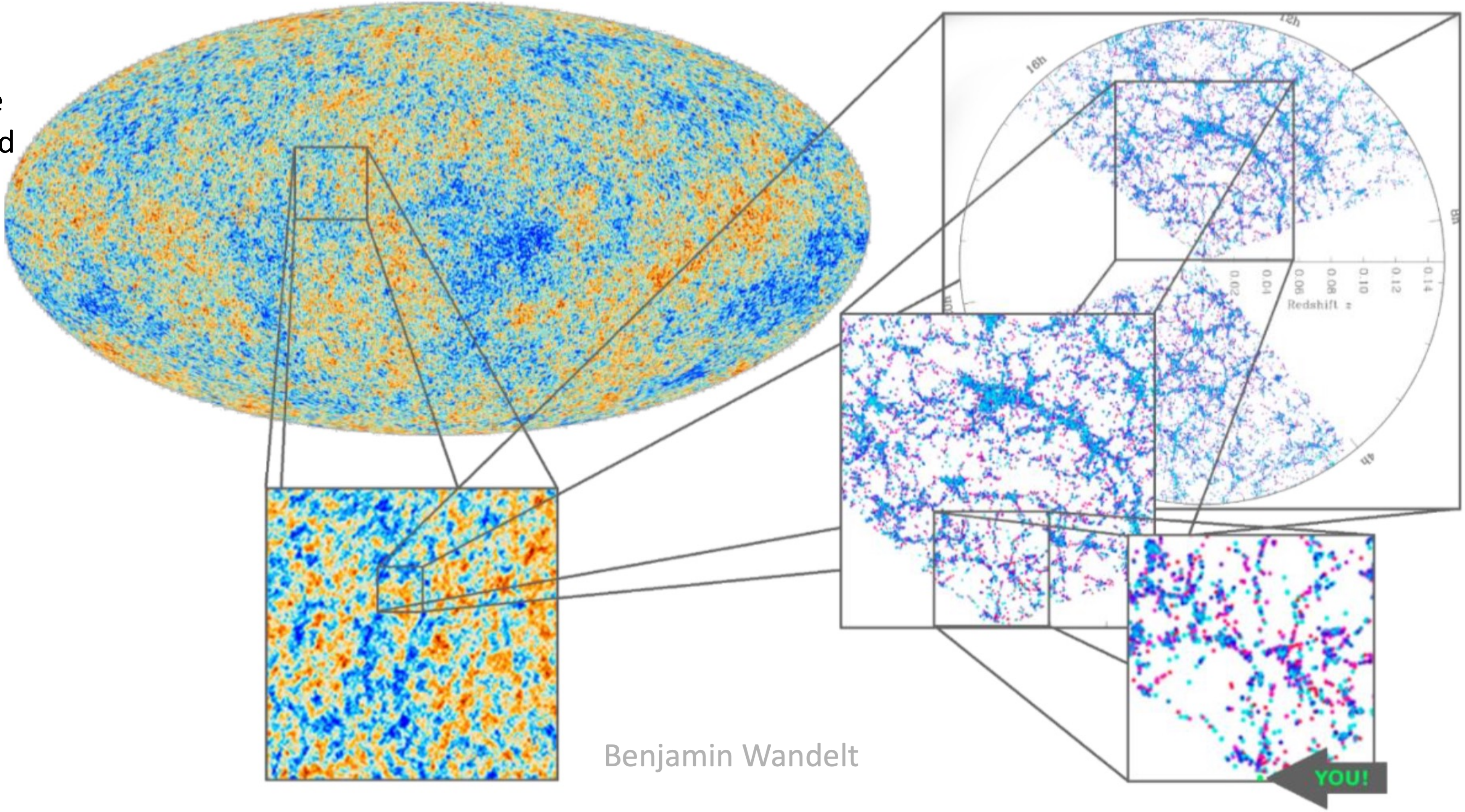
Implicit Inference

Justin Alsing (Oskar Klein Center), Benjamin D. Wandelt

Niall Jeffrey (UCL), Matt Ho (IAP), Xiaosheng Zhang (Tsinghua), Gabriel Jung (IAS, Orsay), Lucas Makinen (Imperial), Will Coulton (CCA), Stephen Feeney, ...

Cosmological data covers a hierarchy of scales on the past light cone. Smaller scales ➔ increasing complexity

Cosmic Microwave Background



Galaxy surveys

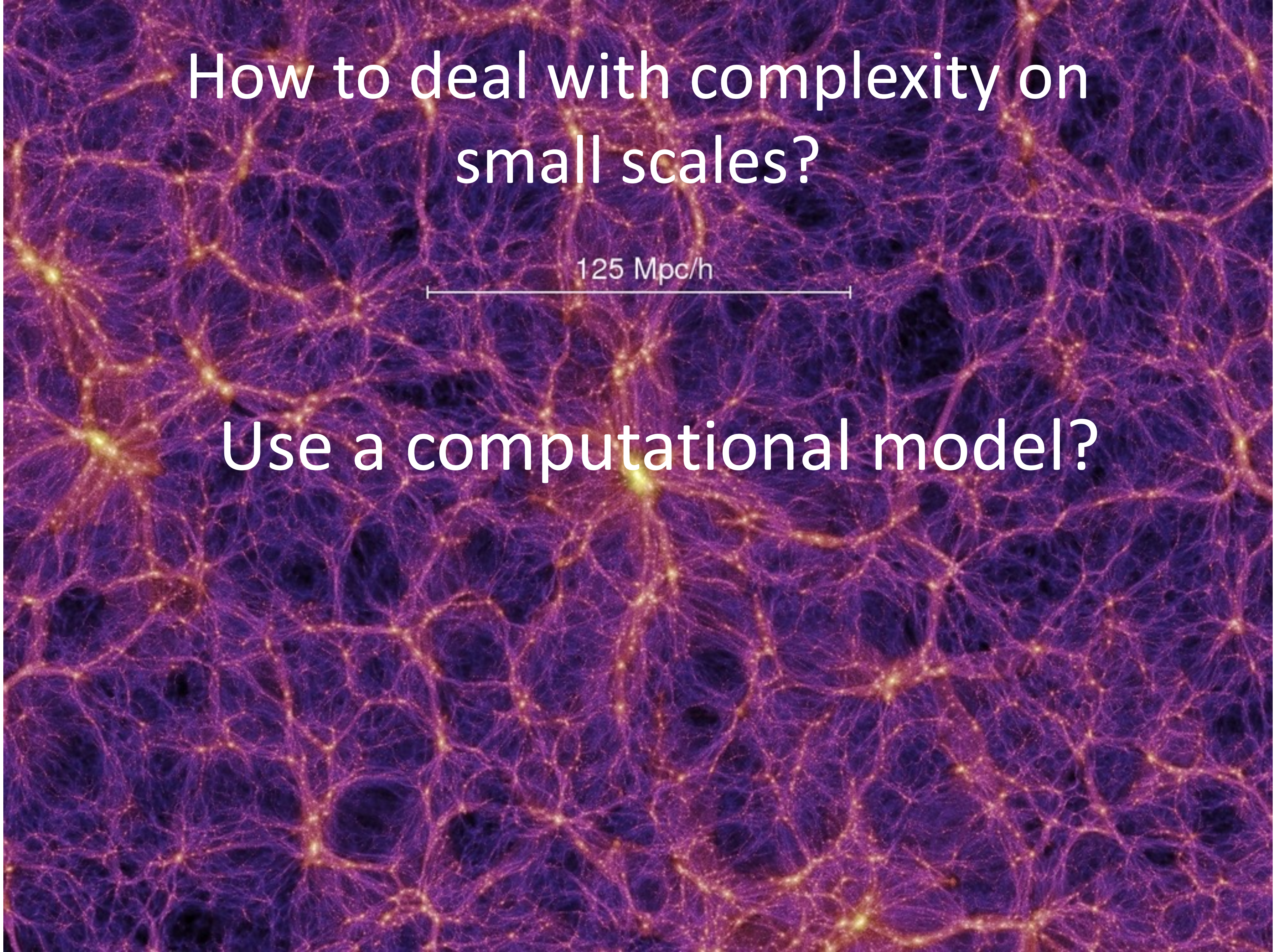
Benjamin Wandelt

How to deal with complexity on small scales?

How to deal with complexity on small scales?

Smooth away small scales?

How to deal with complexity on small scales?



Use a computational model?

Approach 1: "Explicit" inference

1. Write down an explicit model of the full physical and stochastic model of data given parameters θ. This is the **Likelihood**.
2. Get data d.
3. Specify **prior**
4. Write down **posterior**

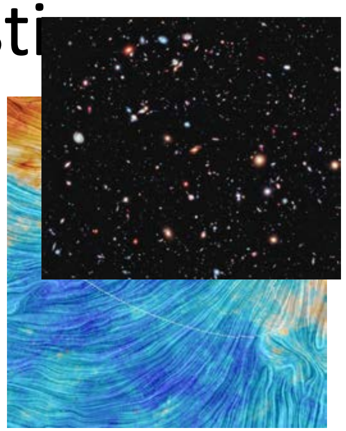
$$P(\boldsymbol{\theta}|\mathrm{d}) = \frac{P(\mathrm{d}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathrm{d})}$$

5. Explore/sample posterior for fixed data as a function of parameters.
6. Done!

Approach 1: “Explicit” inference

1. Write down an explicit model of the full physical and stochastic model of data given parameters θ. This is the **Likelihood**.
2. Get data d.
3. Specify **prior**
4. Write down **posterior**

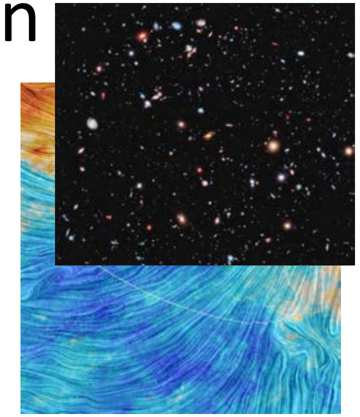
$$P(\boldsymbol{\theta}|\mathrm{d}) = \frac{\cancel{P(\mathrm{d}|\boldsymbol{\theta})}\,P(\boldsymbol{\theta})}{P(\mathrm{d})}$$

What if d =  ?

5. Explore/sample posterior for fixed data as a function of parameters.
6. Done!

Approach 1: “Explicit” inference

1. Work out the full physical and stochastic model of data given parameters θ. This is the **Likelihood**.
2. Get data d.
3. Specify **prior**
4. Write down **posterior**

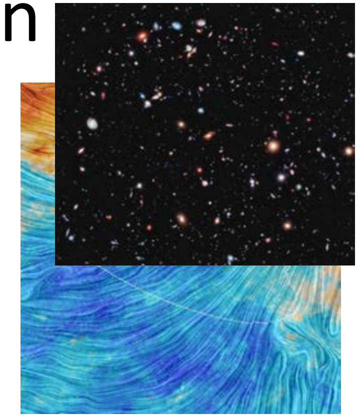
What if d =  ?

$$P(\theta|\mathrm{d}) = \frac{\cancel{P(\mathrm{d}|\theta)}\,\cancel{P(\theta)}}{P(\mathrm{d})}$$

5. Explore/sample posterior for fixed data as a function of parameters.

Approach 1: “Explicit” inference

1. Work out the full physical and stochastic model of data given parameters θ. This is the **Likelihood**.
2. Get data d.
3. Specify **prior**
4. Write down **posterior**

What if d =  ?

$$P(\boldsymbol{\theta}|\mathrm{d}) = \frac{\cancel{P(\mathrm{d}|\boldsymbol{\theta})}\,\cancel{P(\boldsymbol{\theta})}}{P(\mathrm{d})}$$

5. **Explore/sample** posterior for fixed data as a function of parameters.

What we want

To succeed we need more freedom than a traditional likelihood approach can provide:

- FREEDOM to make our physical model anything we want
- FREEDOM to project/summarize/cut/mask our data any way we want

Simulating data is often **much easier** than deriving an accurate likelihood.

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

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

Yes!

A major shift over the last 5 years.

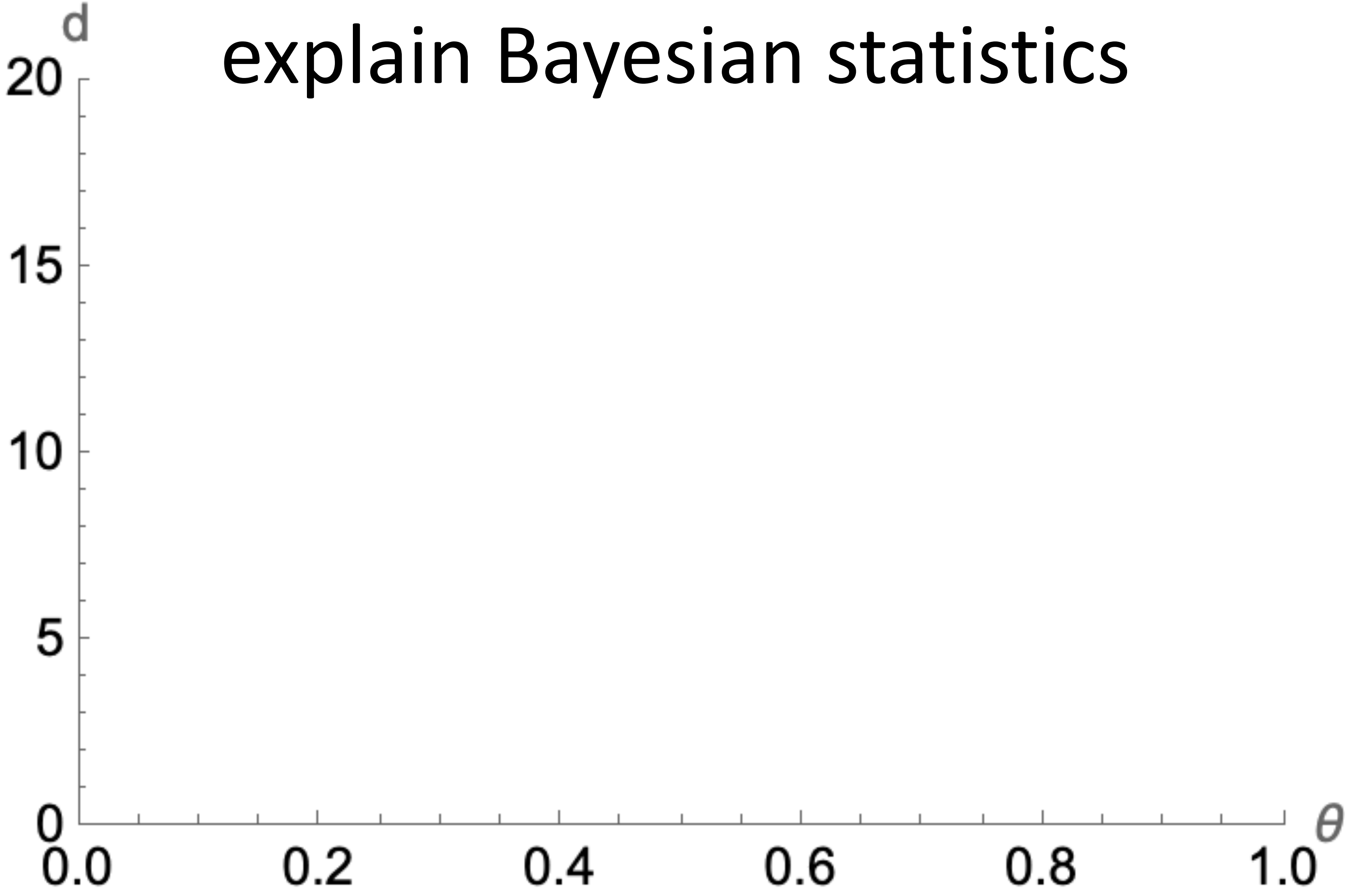
Likelihood is represented *implicitly* through simulations

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

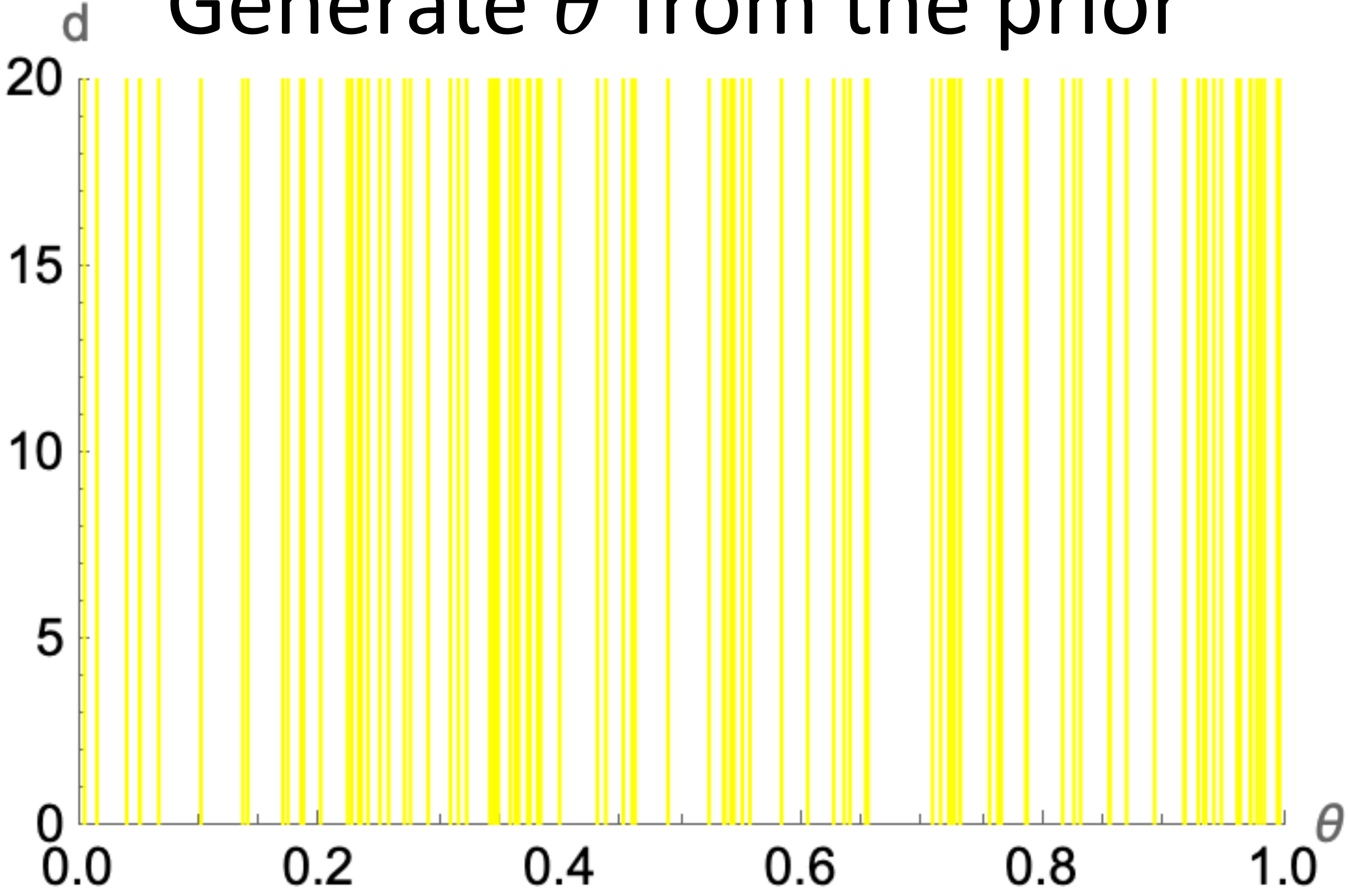
Let's do a simple example.

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

The easiest diagram to explain Bayesian statistics

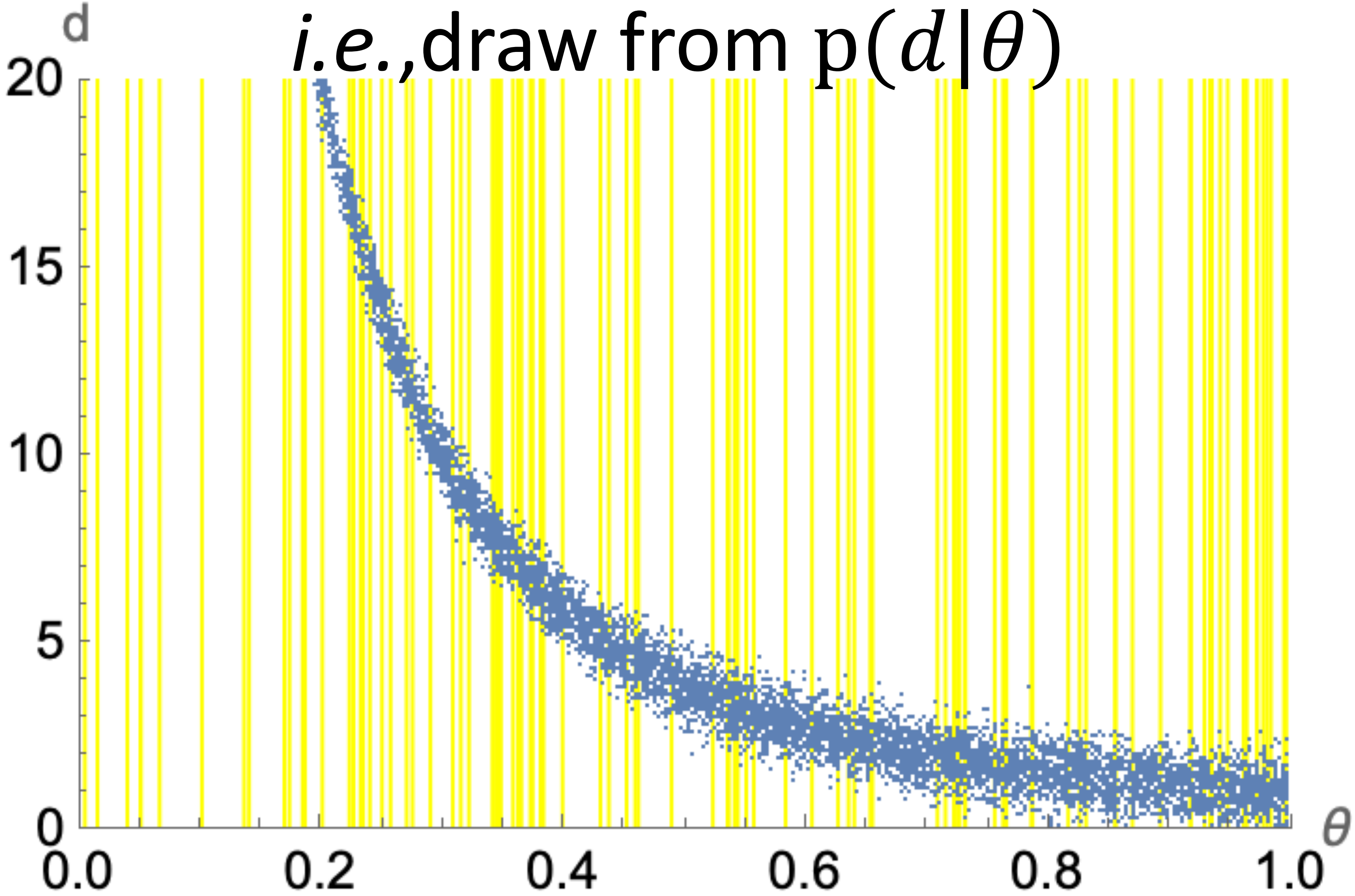


Generate θ from the prior

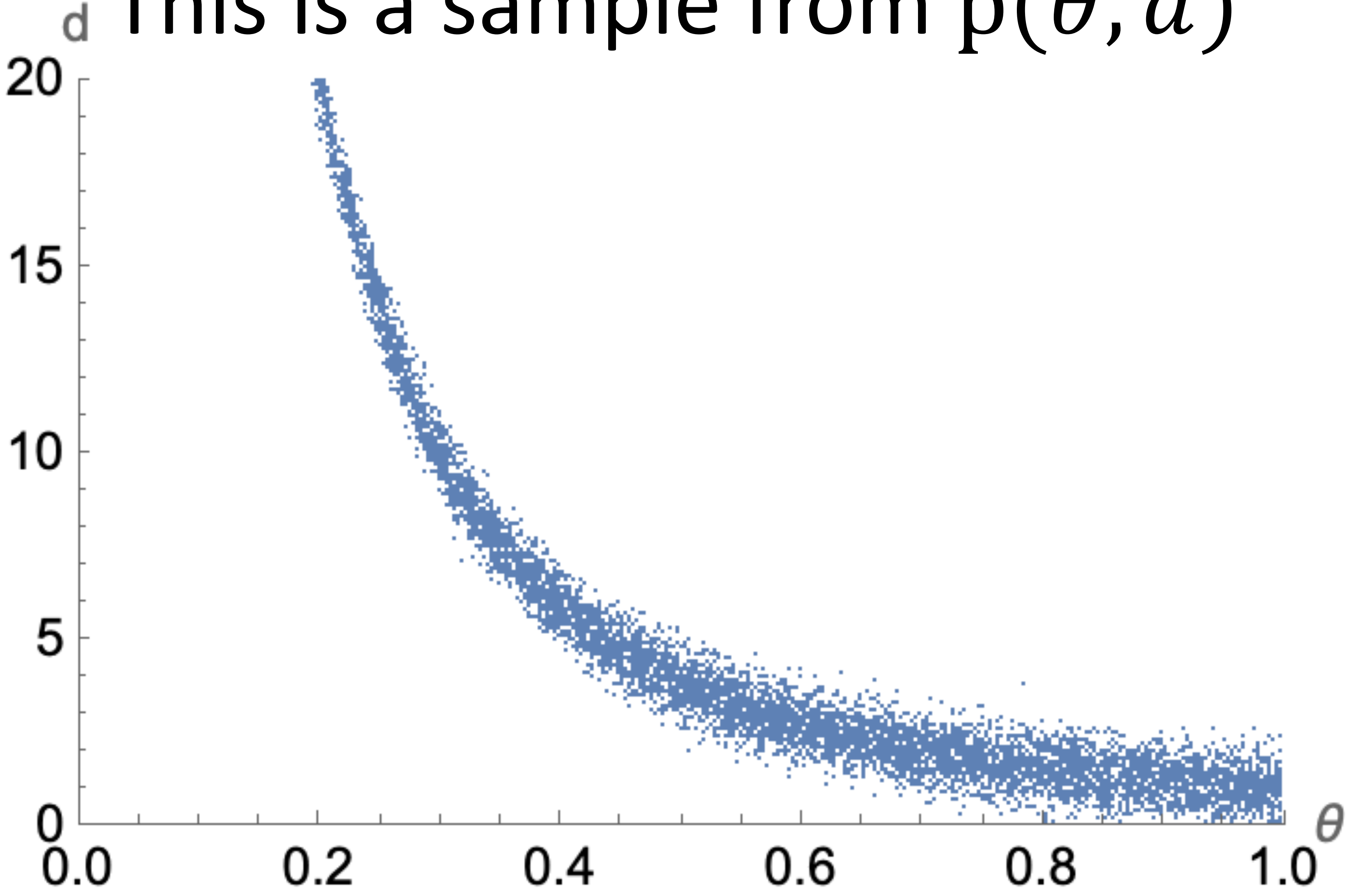


Simulate/generate d given θ

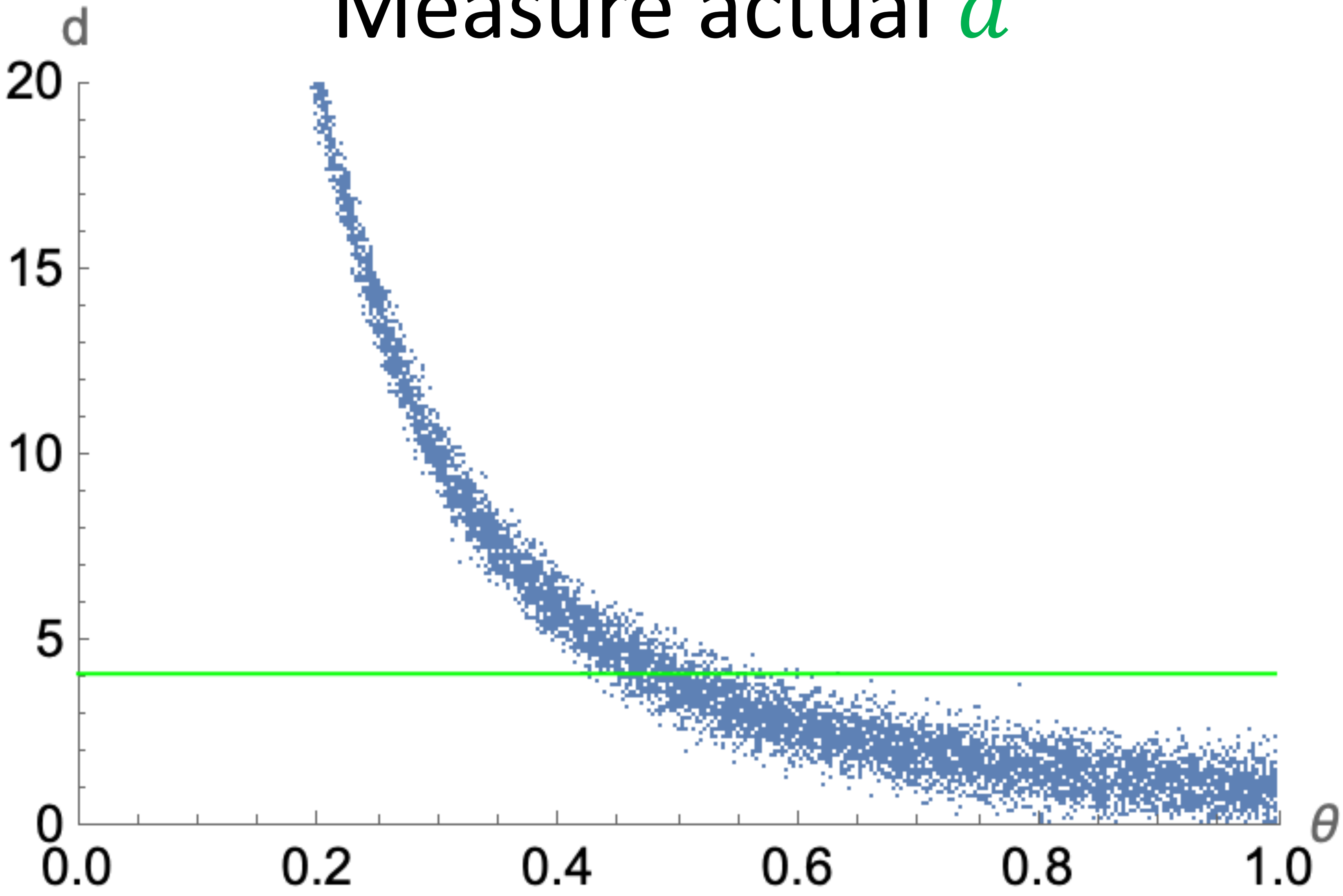
i.e.,draw from $\mathrm{p}(d|\theta)$



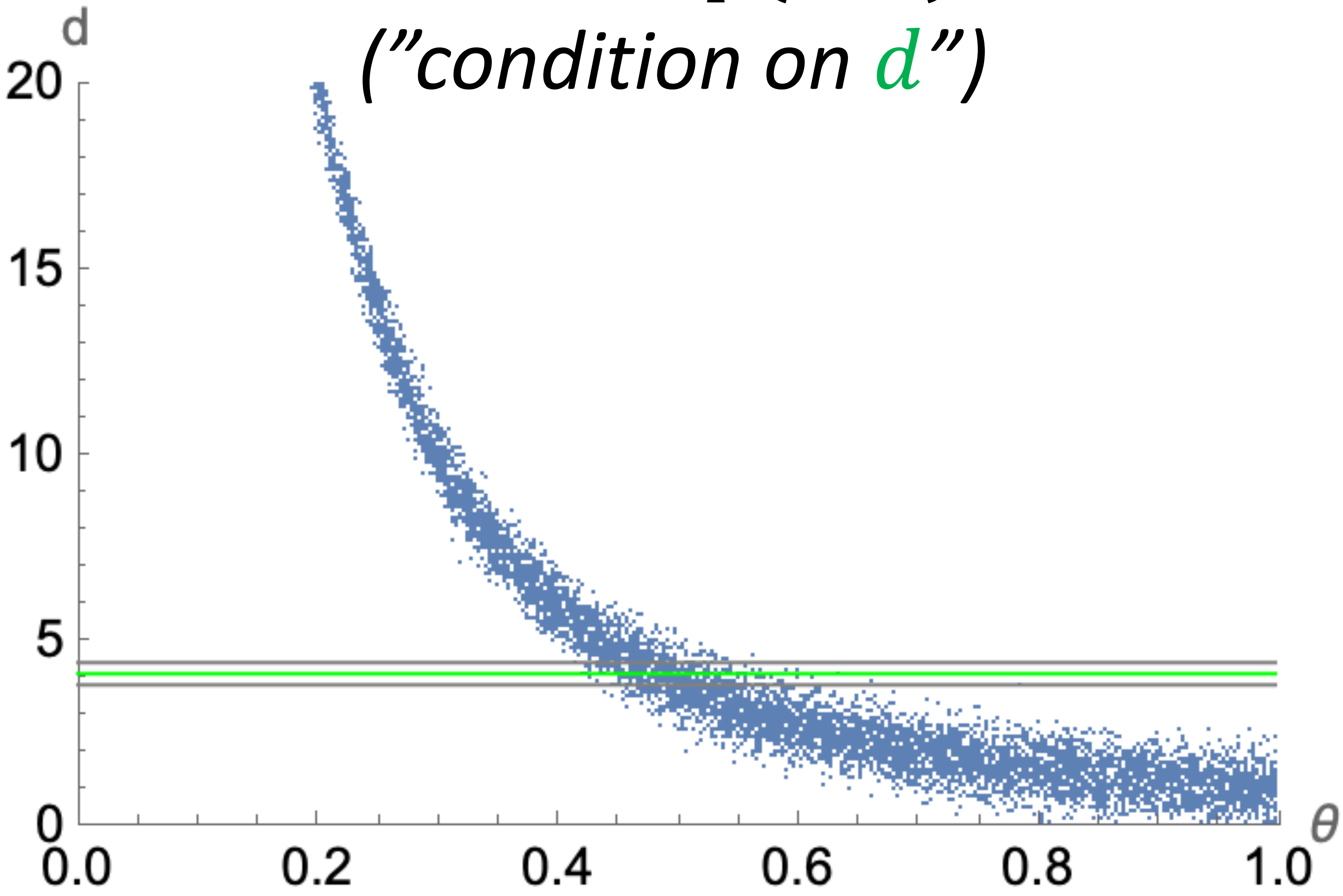
This is a sample from $\mathrm{p}(\theta, d)$



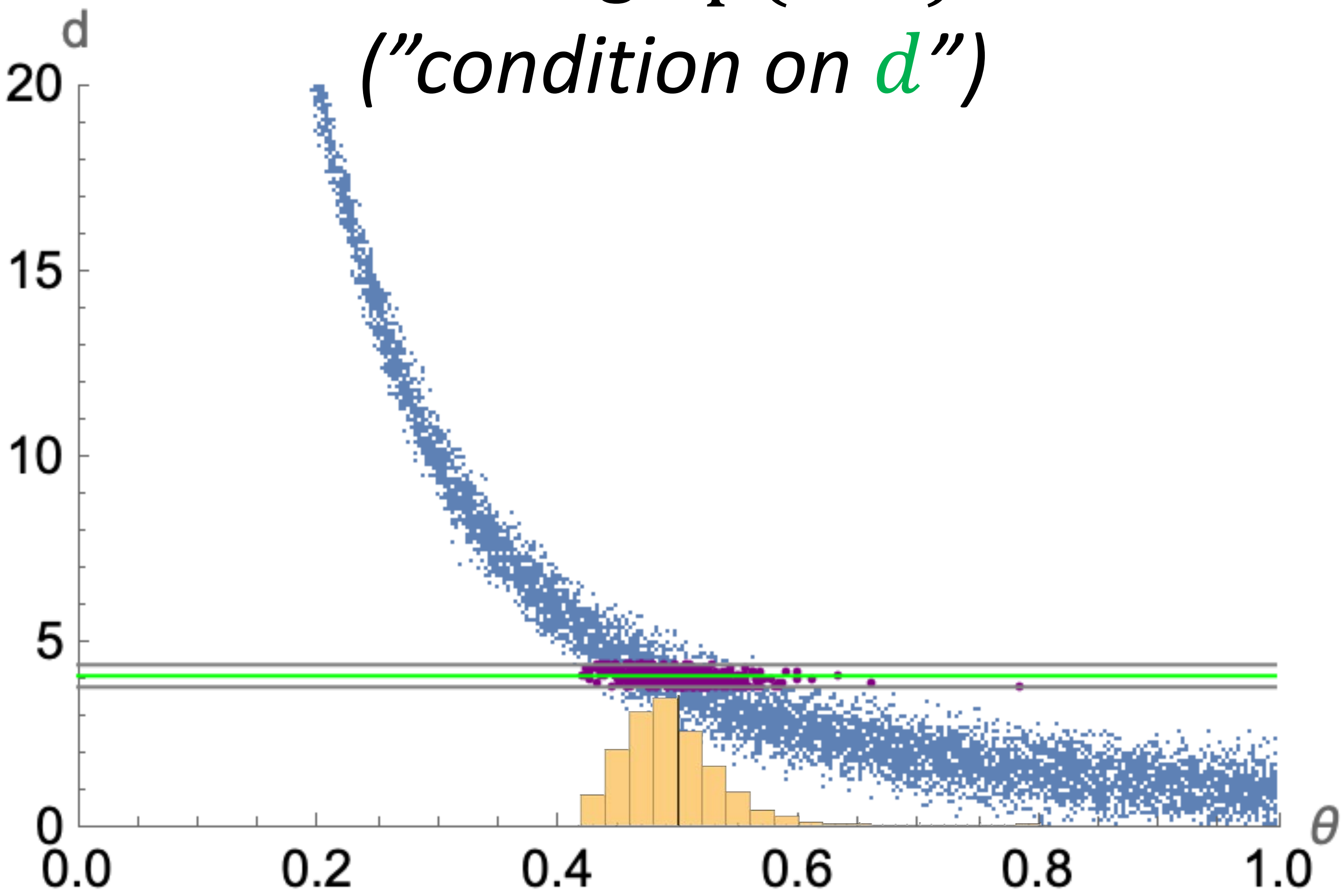
Measure actual d



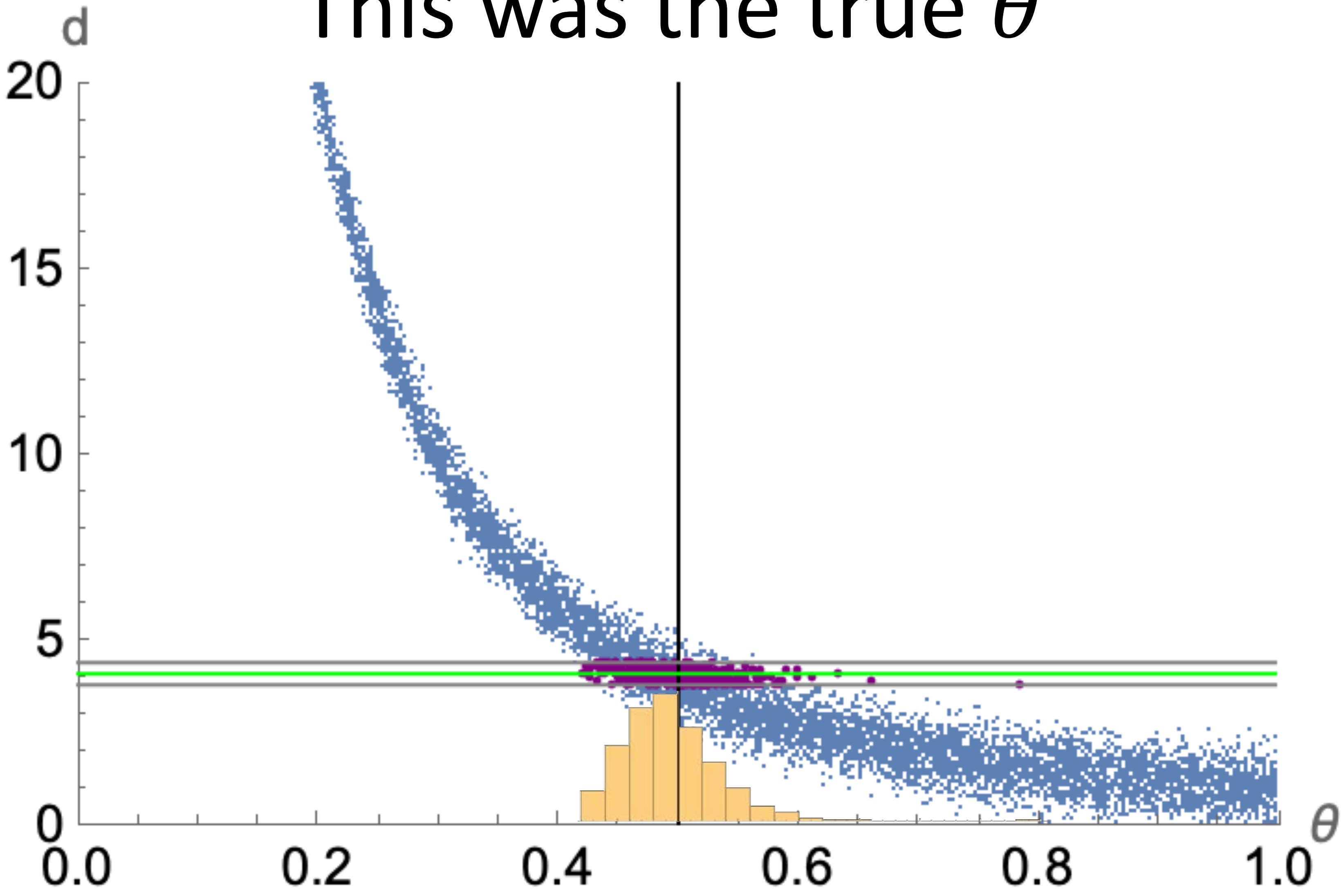
Slice through $\mathrm{p}(\theta, d)$ at d
(”condition on d”)



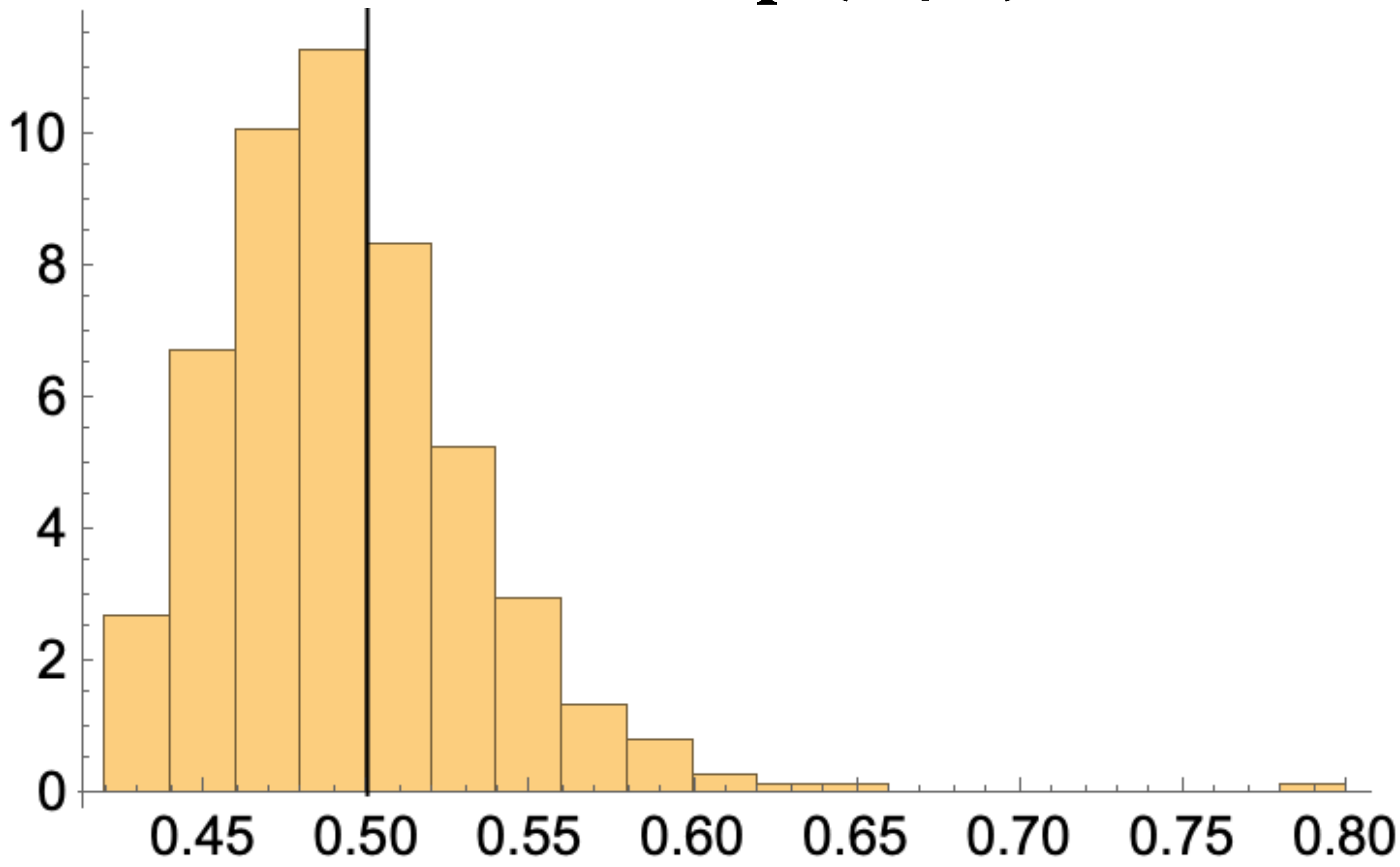
Slice through $\mathrm{p}(\theta, d)$ at d
("condition on d")



This was the true θ



Posterior $p(\theta|d)$



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)

The Machine Learning revolution in computational astrophysics and cosmology

- Many problems that we considered impossible **solved** in the last ~5 years:
  - Automated finding of informative data summary statistics
    - computing score functions and Fisher Information 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 posterior moments (*e.g.,* Moment Networks)
    - Posterior samples for huge non-linear inverse problems (*e.g.,* Initial Conditions)
  - Bayesian Evidence for intractable models
    - Simulation-based model comparison(*e.g.,* Evidence Networks)

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

Machine learning takes us the rest of the way

- ***Recast inference problems as optimization problems***.

- Write down a loss that defines the problem
  - Parameterize the solution using a neural network
  - Minimize
  - Validate

First example: variational Bayes

- Define a parameterized family of distributions
- Minimize Kullback-Leibler loss between neural family and true likelihood

When using a neural density estimator this is DELFI, a (now) classic example of simulation-based inference.

$$D_{\mathrm{KL}}(p^* \mid p) = \int p^*(\mathbf{t}|\boldsymbol{\theta}) \ln\left(\frac{p(\mathbf{t}|\boldsymbol{\theta};\mathbf{w})}{p^*(\mathbf{t}|\boldsymbol{\theta})}\right) d\mathbf{t}$$

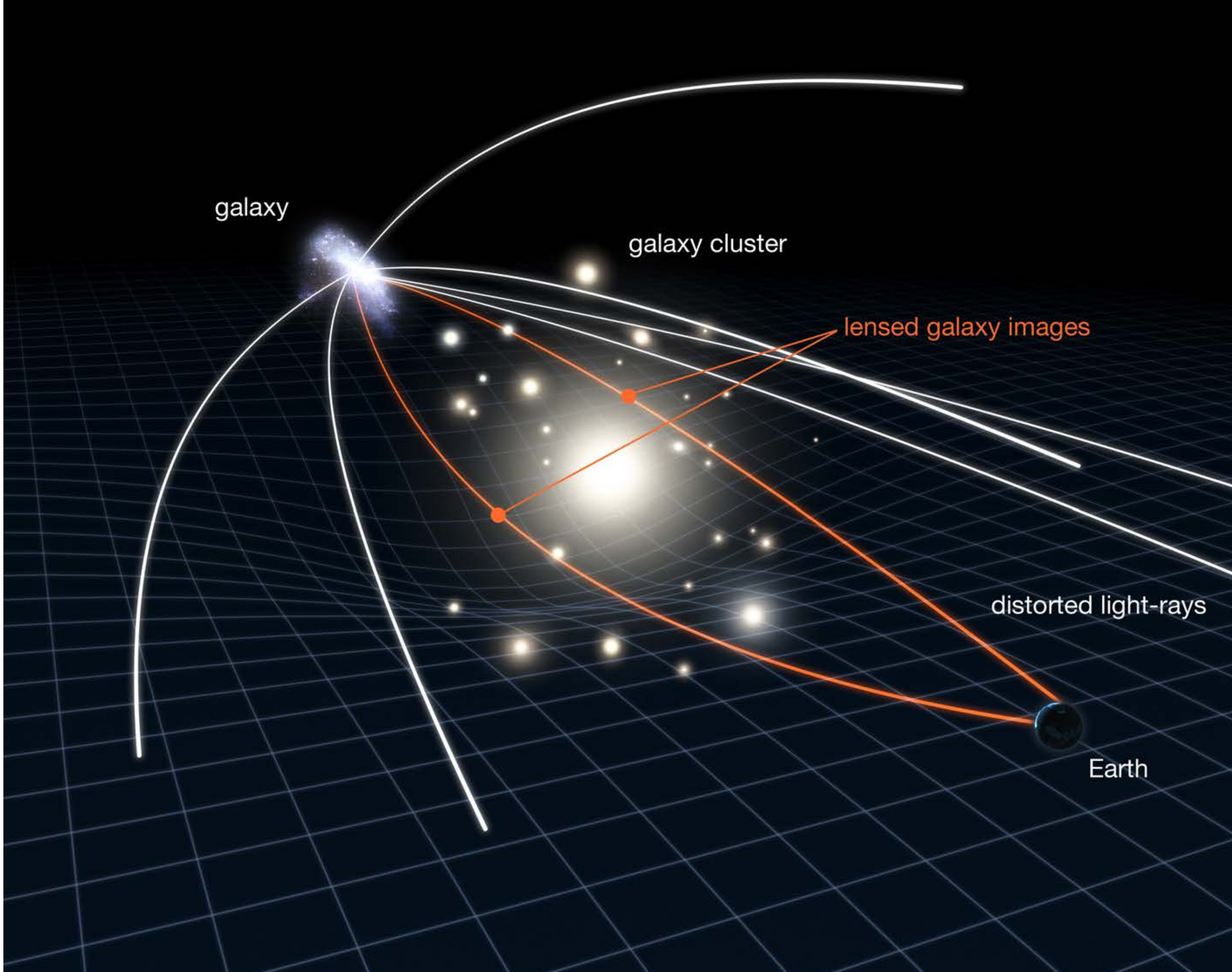
$$-\ln U(\mathbf{w}|\{\boldsymbol{\theta},\mathbf{t}\}) = -\sum_{i=1}^{N_{\mathrm{samples}}} \ln p(\mathbf{t}_i|\boldsymbol{\theta}_i;\mathbf{w})$$

Papamakarios, Murray + coauthors, arXiv:1605.06376, 1705.07057, 1805.07226
Alsing, Feeney & Wandelt, arXiv: 1801.01497, 1903.01473

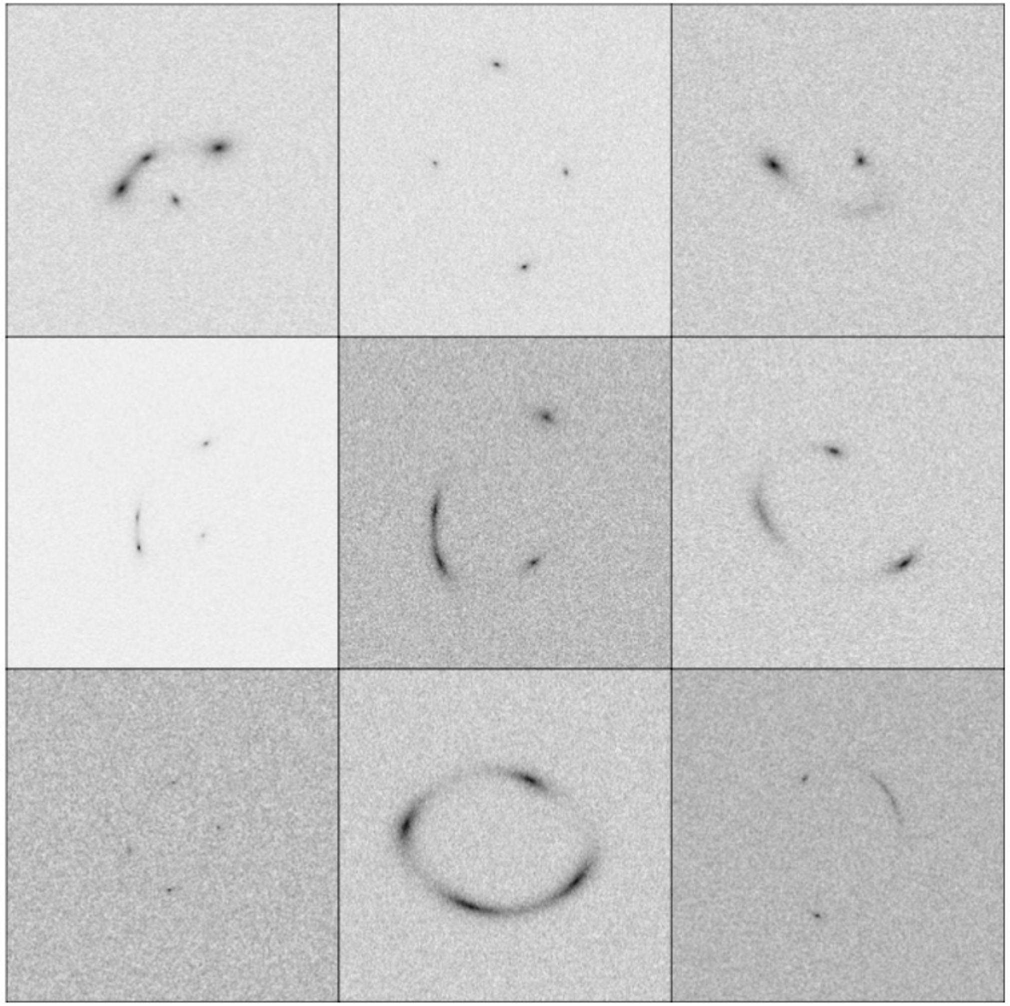
Example:

Strong Gravitational Lensing

Legin et al. arXiv:2212.00044

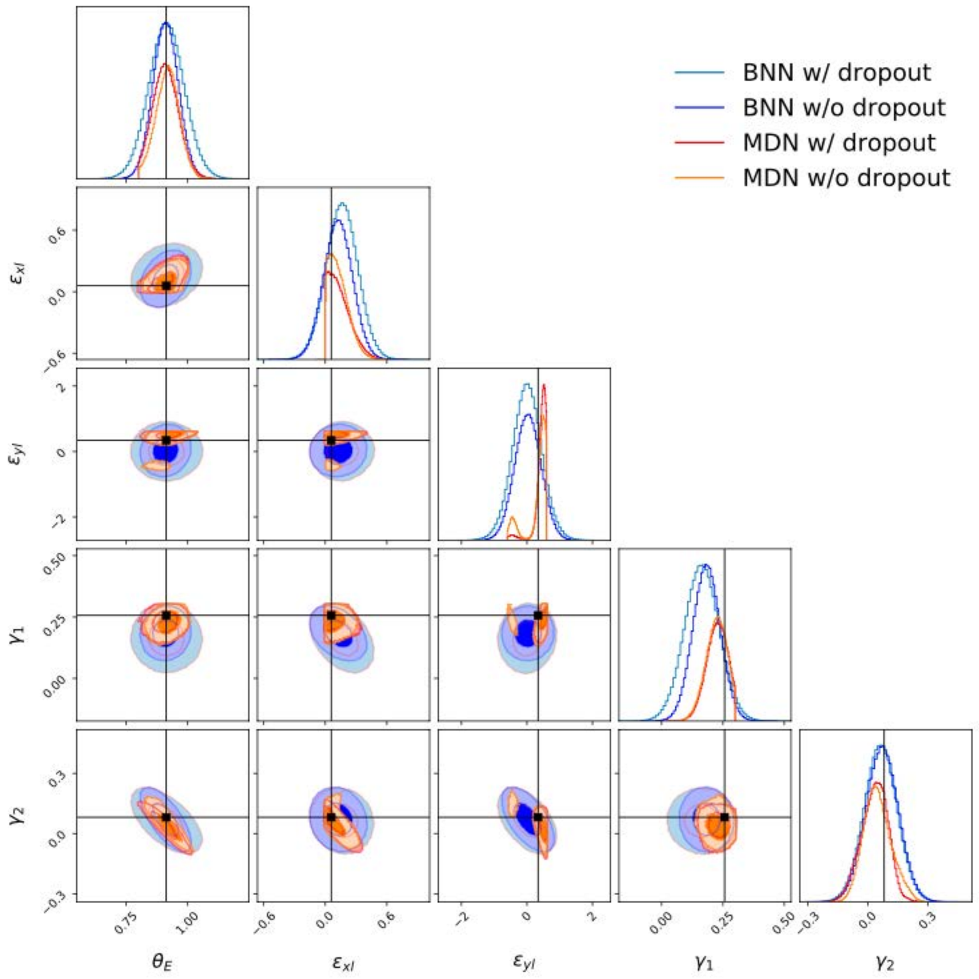


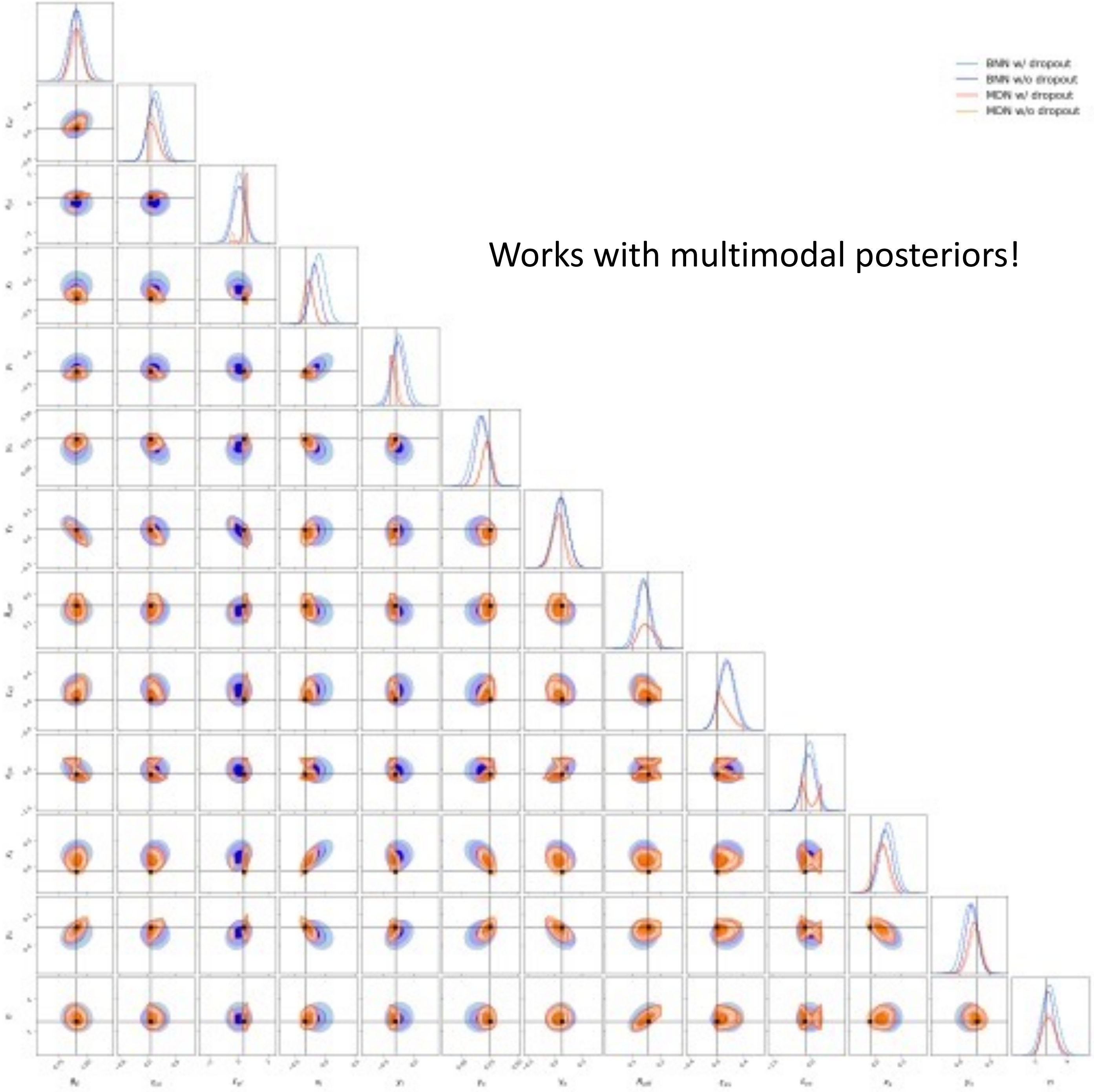
Simulated images



Legin et al arXiv 2212.00044

Inference

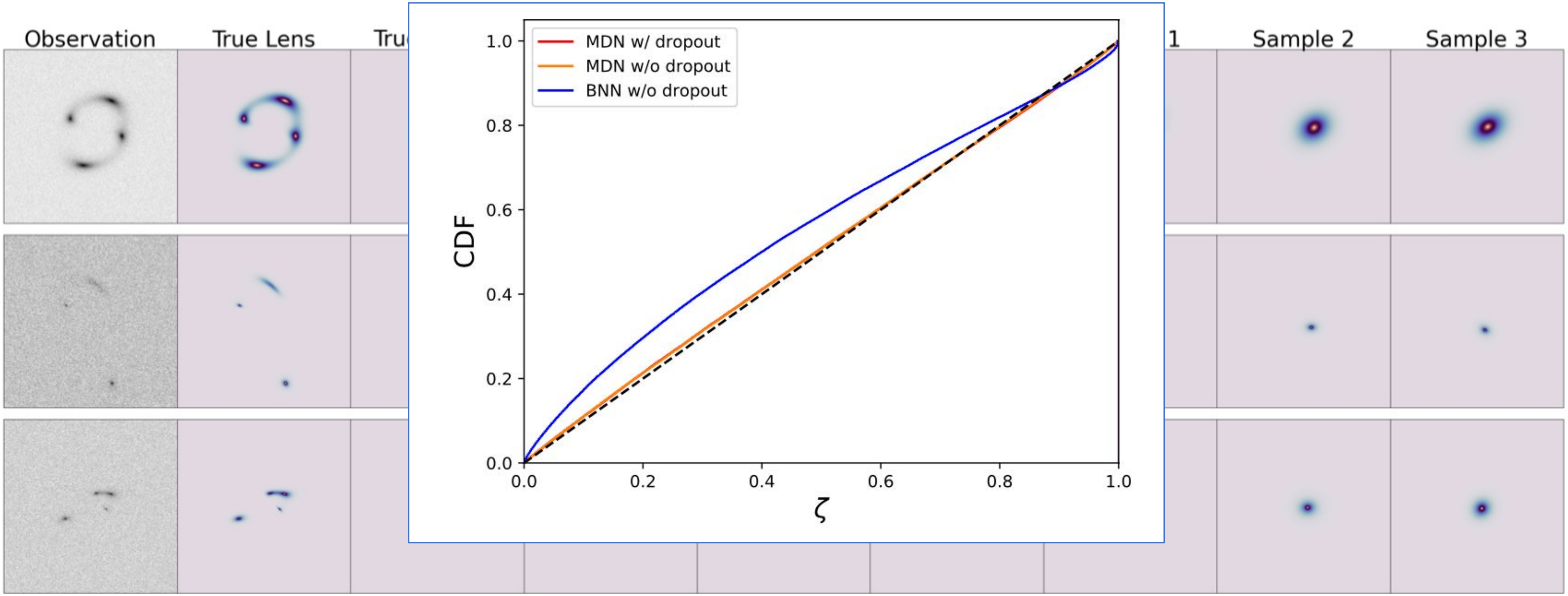




Works with multimodal posteriors!

Legin et al arXiv 2212.00044

Validation



Legin et al arXiv 2212.00044

What if the number of parameters is large or simulations are scarce?

- In general, neural density estimation becomes exponentially hard as number of dimensions increases.
- How do we handle high-dimensional problems?

- Simplify.

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

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

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

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

Moment Network Example

Cosmology and astrophysics from full hydrodynamical simulations including black holes, star formation,…

CAMELS

Cosmology and Astrophysics with Machine Learning

Large suites of full, cosmological hydrosimulations as a function of cosmological parameters and astrophysics models with multiple codes (AREPO/Illustris, GIZMO/SIMBA, Astrid,…).

F. Villaescusa-Navarro, S. Genel, D. Angles-Alcazar et al. arXiv:2109.10915
F. Villaescusa-Navarro, D. Angles-Alcazar, S. Genel et al. arXiv:2010.00619

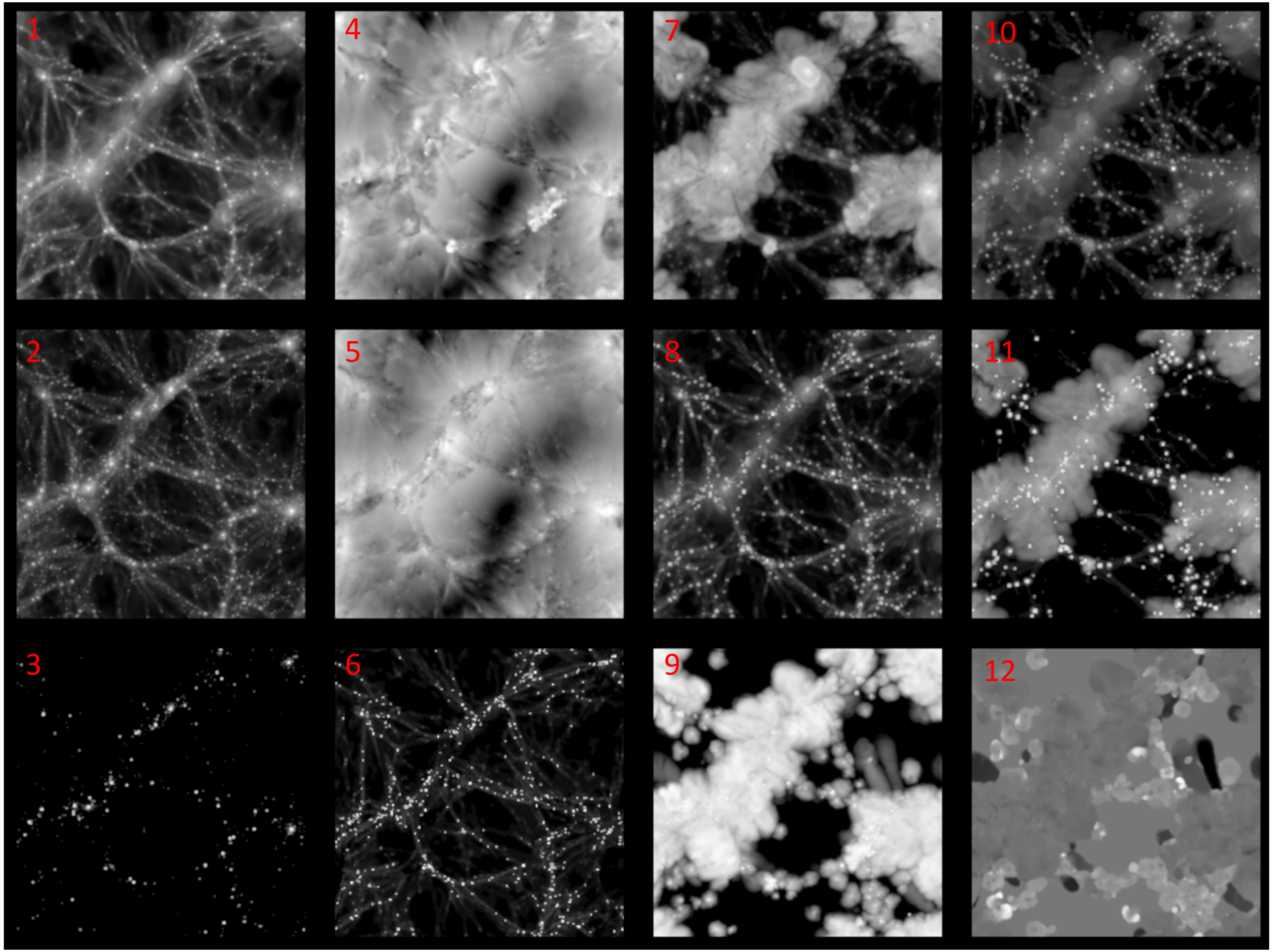
Cosmology on small scales with baryons

15 different 2-dimensional fields:

1. Gas mass
2. Dark matter mass
3. Stellar mass
4. Gas velocity
5. Dark matter velocity
6. Neutral hydrogen mass
7. Gas temperature
8. Electron density
9. Gas metallicity
10. Gas pressure
11. Magnetic fields
12. Mg/Fe
13. Total mass
14. N-body
15. All fields except dark matter

15,000 images per field from 1,000 CAMELS-IllustrisTNG simulations.
Each image:

- 250x250 pixels
- 25x25 $(\mathrm{Mpc}/h)^2$
- 100 kpc/h resolution



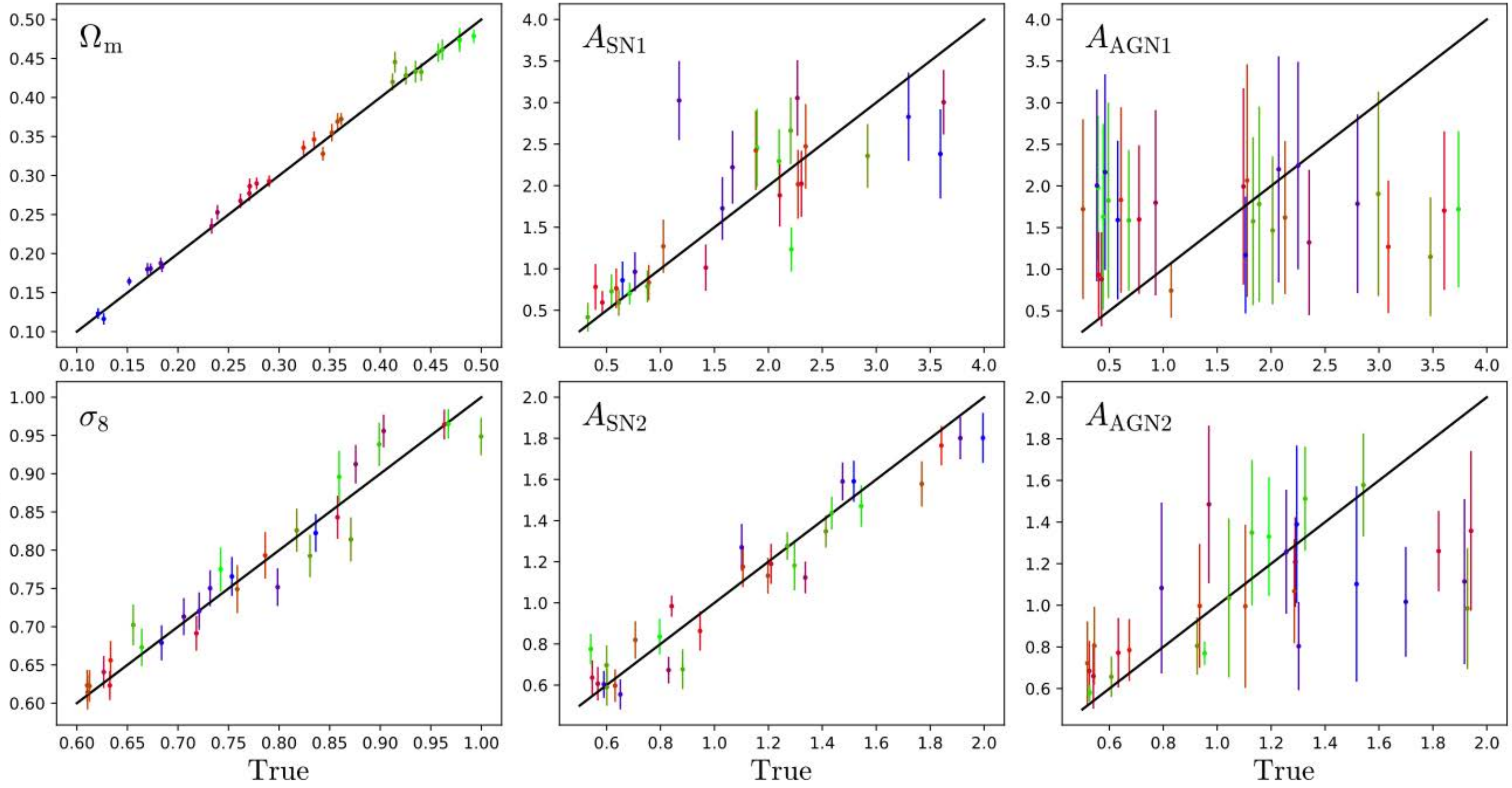
SBI: COSMOLOGY FROM SMALL-SCALE HYDRO

Villaescusa-Navarro, arXiv:2109.09747

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

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

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



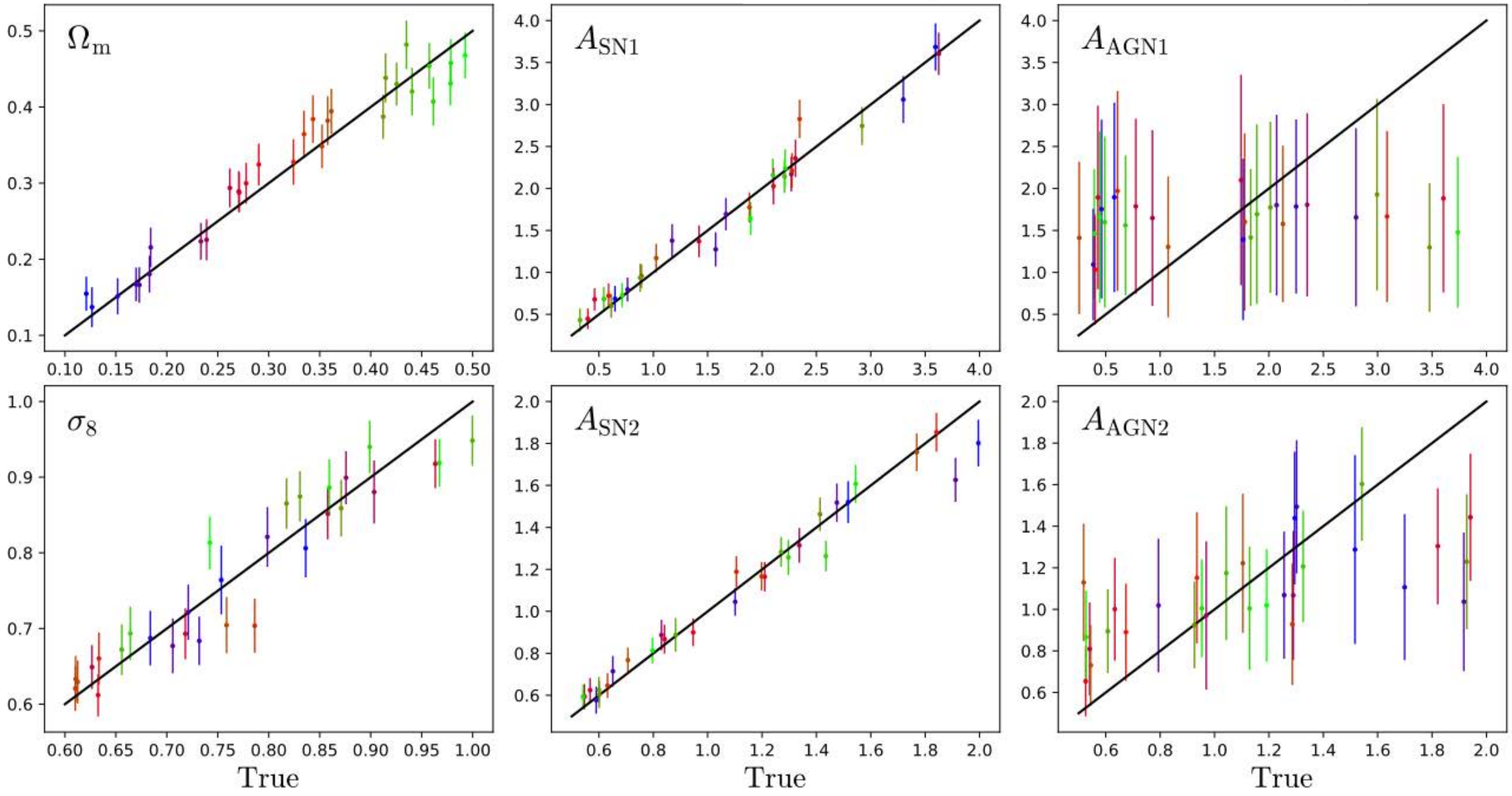
SBI: Cosmology from Small-Scale Hydro

Villaescusa-Navarro, arXiv:2109.09747

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

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

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

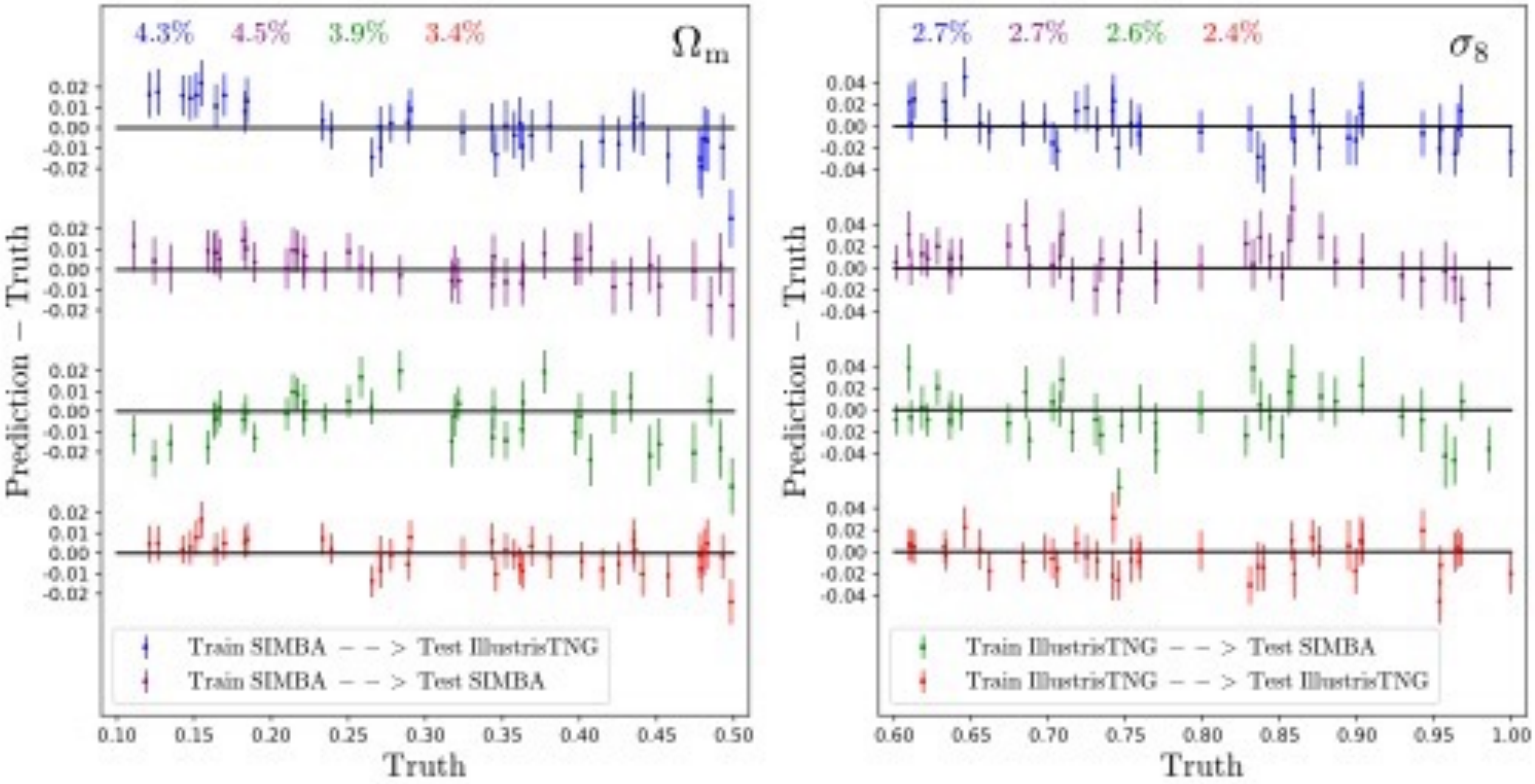


What the cosmological AI tells us about the CAMELS Multifield Data set

1. There is cosmological information on very small scales (100 kpc)
2. The hydro outputs contain *more* information than the dark matter density
3. For *total matter,* inferences are *robust* to baryonic physics (good news for weak lensing!)

Villaescusa-Navarro et al., arXiv:2109.09747, arXiv:2109.10360

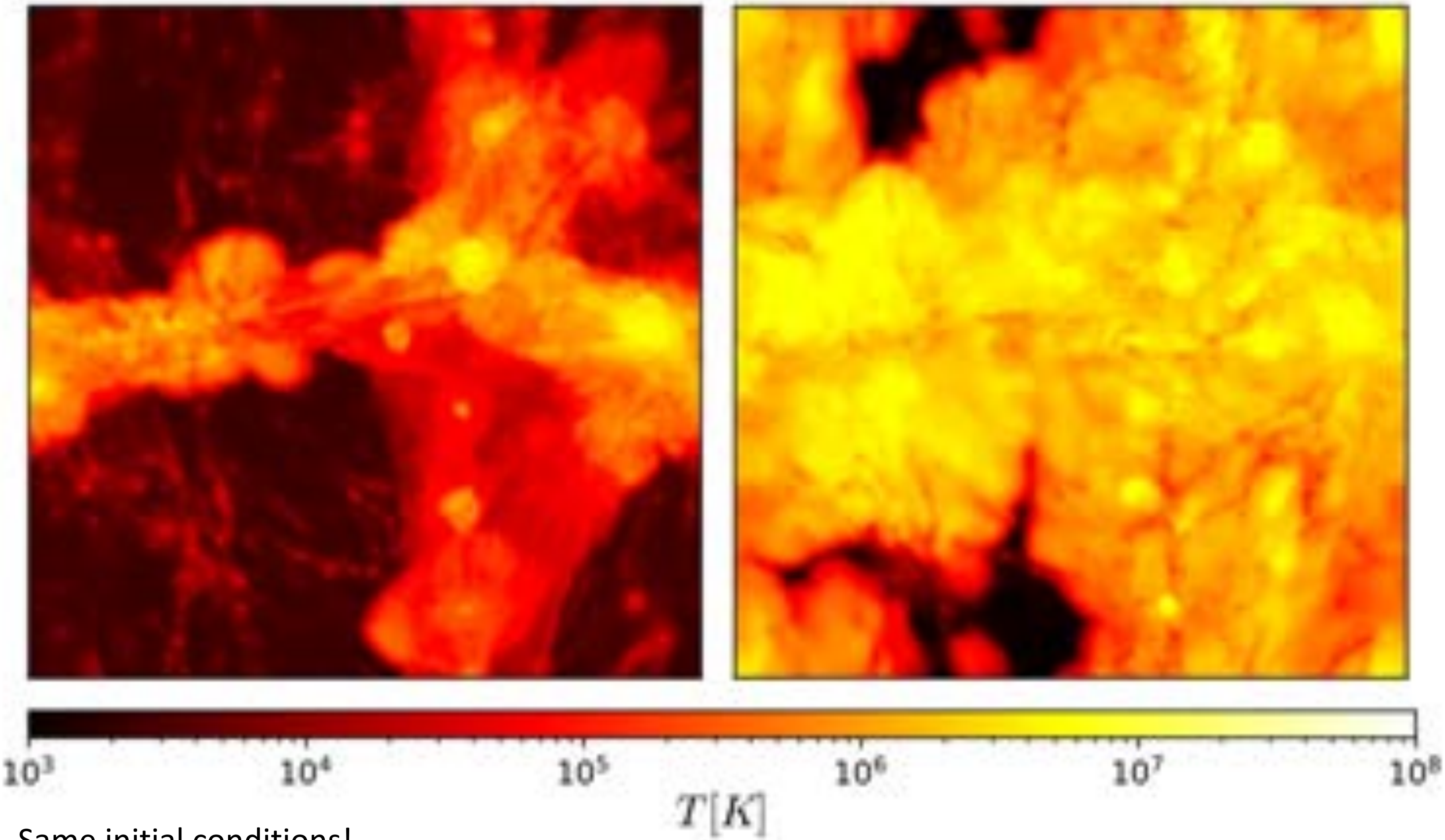
Cosmology robust to baryonic physics



Villaescusa-Navarro et al., arXiv:2109.10360

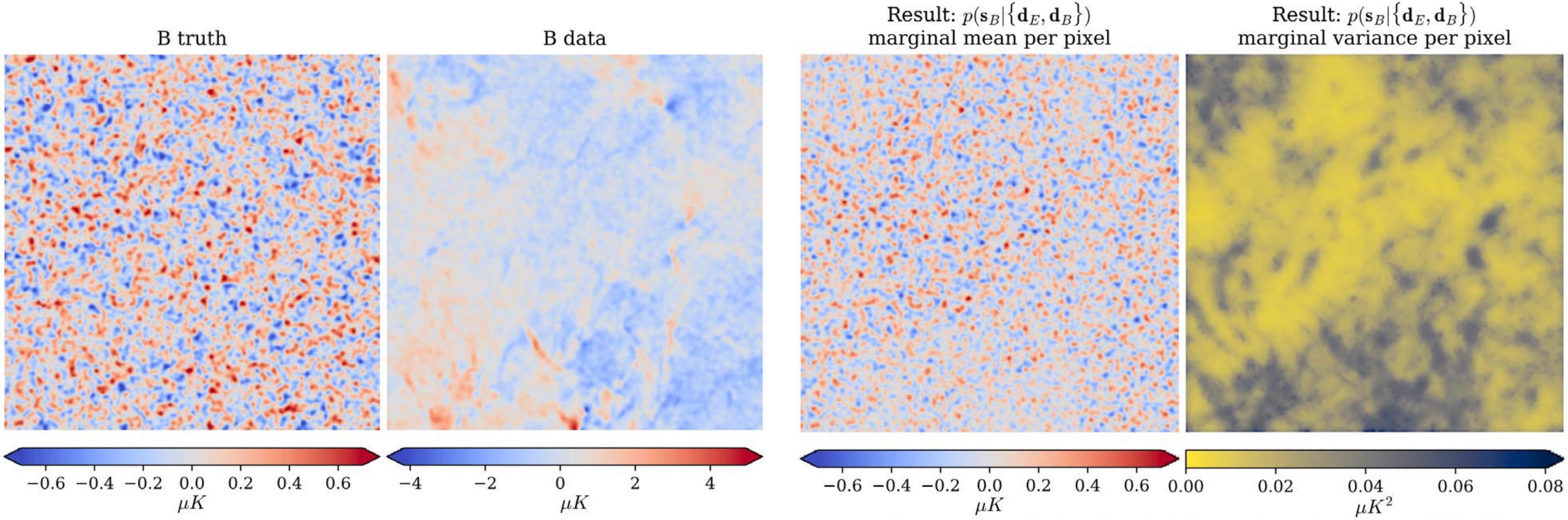
Villaescusa-Navarro et al., arXiv:2109.09747

Illustris TNG SIMBA



Same initial conditions!

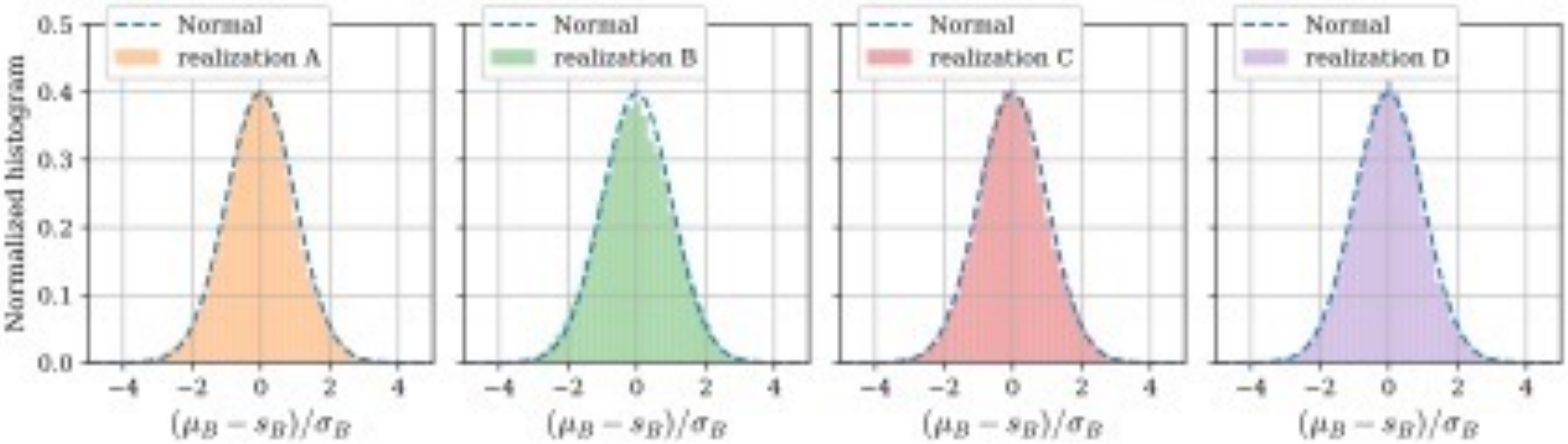
High dimensional application of Moment Networks (with **a single training image)**



Uses a generative model based on Wavelet Phase Harmonics

Jeffrey, Boulanger, Wandelt, Regaldo-Saint Blancard, Allys, Levrier 2021, arXiv:2111.01138
(Allys et al. 2020; Regaldo-Saint Blancard et al. 2021; Jeffrey & Wandelt, arXiv:2011.05991)

Moment networks: Posterior means and variances pass quantile test



Jeffrey, Boulanger, Wandelt, Regaldo-Saint Blancard, Allys, Levrier 2021, arXiv:2111.01138

Summary Statistics and Compression

Nuisance hardened data compression for fast likelihood-free inference

Justin Alsing[1,2,3]⋆ and Benjamin Wandelt[2,4]

[1] *Oskar Klein Centre for Cosmoparticle Physics, Stockholm University, Stockholm SE-106 91, Sweden*
[2] *Center for Computational Astrophysics, Flatiron Institute, 162 5th Ave, New York City, NY 10010, USA*
[3] *Imperial Centre for Inference and Cosmology, Department of Physics, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK*
[4] *Sorbonne Université, Institut Lagrange de Paris (ILP), 98 bis boulevard Arago, F-75014 Paris, France*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

In this paper we show how nuisance parameter marginalized posteriors can be inferred directly from simulations in a likelihood-free setting, without having to jointly infer the higher-dimensional interesting and nuisance parameter posterior first and marginalize a posteriori. The result is that for an inference task with a given number of interesting parameters, the number of simulations required to perform likelihood-free inference can be kept (roughly) the same irrespective of the number of additional nuisances to be marginalized over. To achieve this we introduce two extensions to the standard likelihood-free inference set-up. Firstly we show how nuisance parameters can be re-cast as latent variables and hence automatically marginalized over in the likelihood-free framework. Secondly, we derive an asymptotically optimal compression from N data down to n summaries – one per interesting parameter – such that the Fisher information is (asymptotically) preserved, but the summaries are insensitive (to leading order) to the nuisance parameters. This means that the nuisance marginalized inference task involves learning n interesting parameters from n "nuisance hardened" data summaries, regardless of the presence or number of additional nuisance parameters to be marginalized over. We validate our approach on two examples from cosmology: supernovae and weak lensing data analyses with nuisance parameterized systematics. For the supernova problem, high-fidelity posterior inference of Ω_m and w_0 (marginalized over systematics) can be obtained from just a few hundred data simulations. For the weak lensing problem, six cosmological parameters can be inferred from just $O(10^3)$ simulations, irrespective of whether ten additional nuisance parameters are included in the problem or not. If needed, an approximate posterior for the nuisance parameters can be re-constructed a posteriori as a pseudo-Blackwell-Rao estimator (without running any additional simulations).

Key words: data analysis: methods

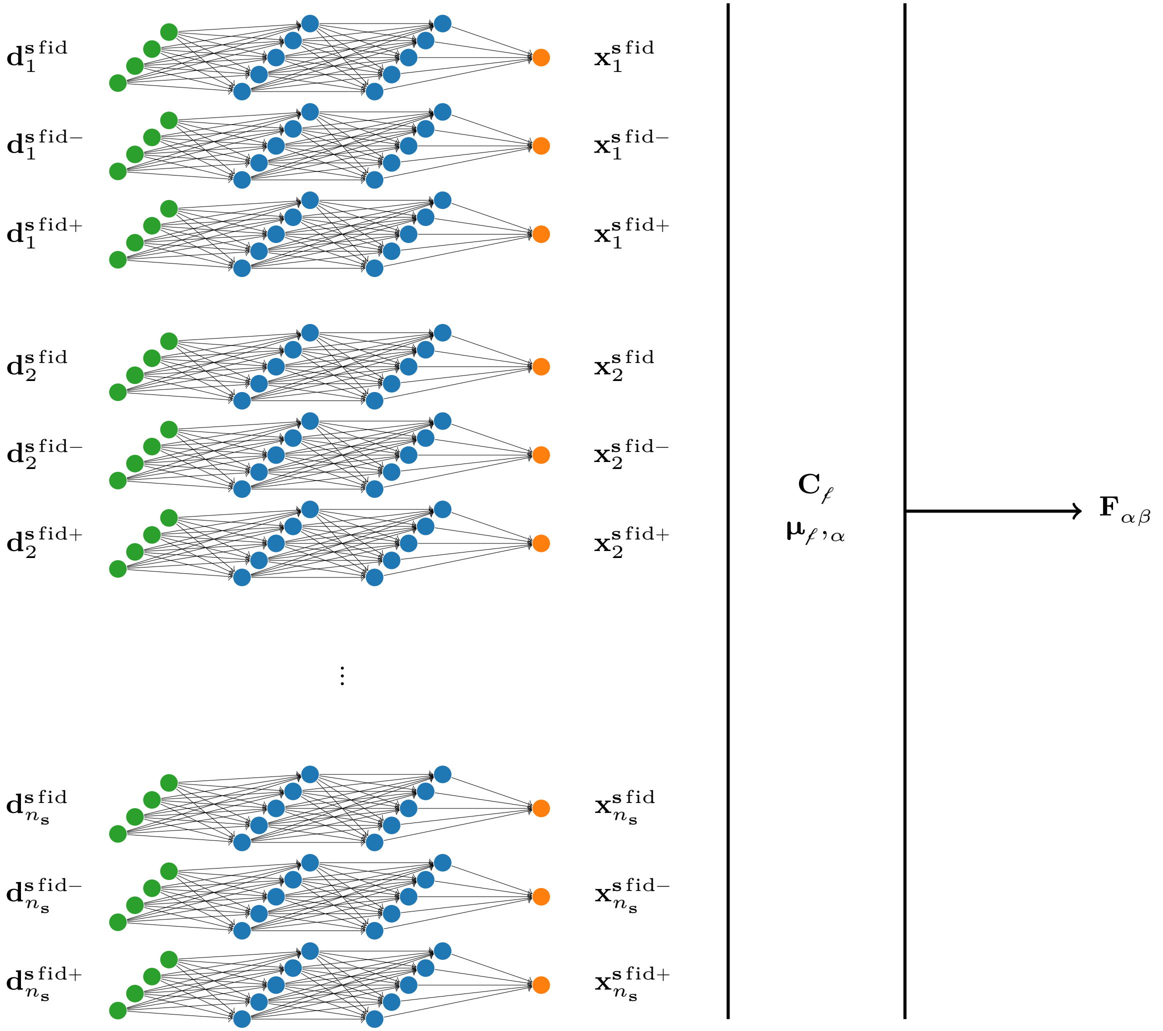
Automatic Physical Inference with Information Maximizing Neural Networks

- Goal: obviate the need to “guess” heuristic, informative summaries of the data
- Setup: design a neural network that maps the data into a small set of informative *summaries*.
- The training loss is (– the Fisher information) under an assumed simple likelihood for the summaries.

- Training uses physical simulations of the model to maximize the information in the summaries about the parameters of the model.
- The achieved loss on a test set is *meaningful* – it’s the information content of the data.

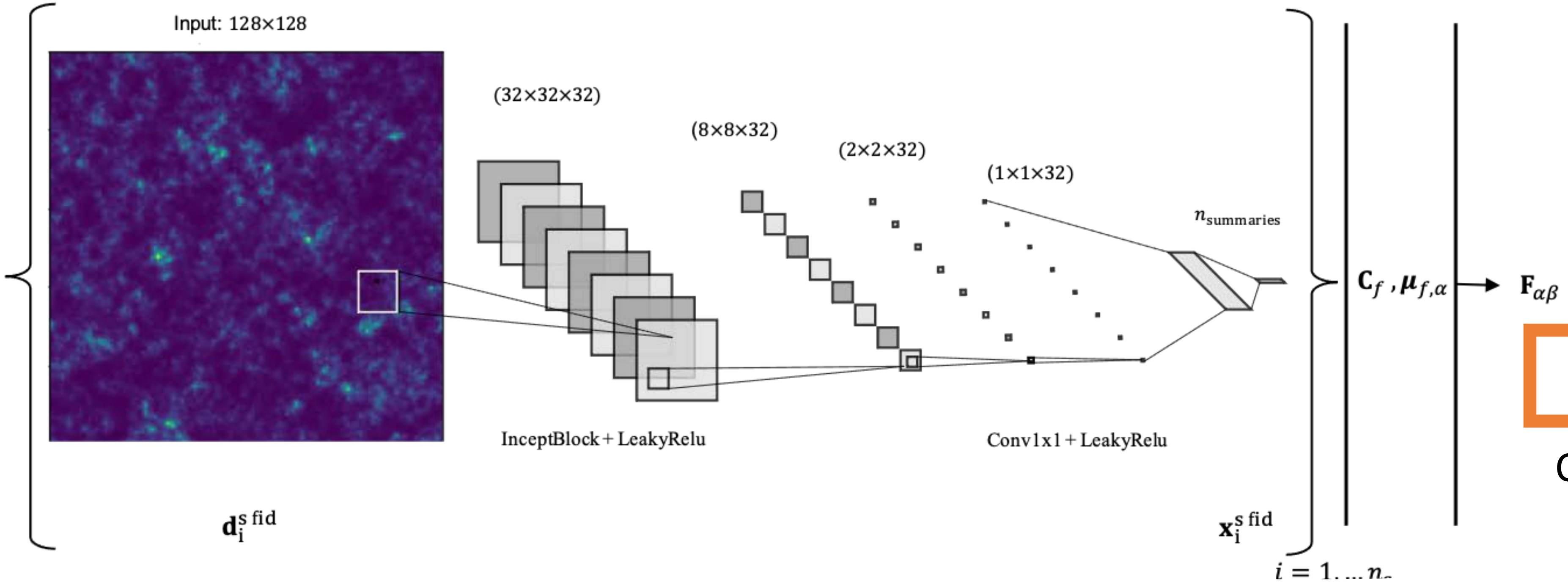
Charnock, Lavaux, Wandelt (arXiv:1802:03537)

Information maximizing neural network



Charnock, Lavaux, Wandelt (arXiv:1802:03537)

Information maximizing neural networks: asymptotically optimal analysis, Information Matrix, score computation *if you don't know the likelihood*



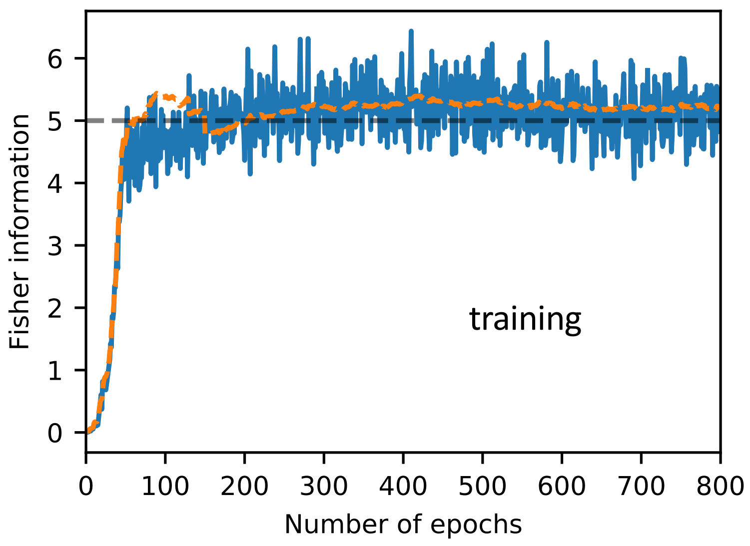
Optimal compression network minimizes information loss

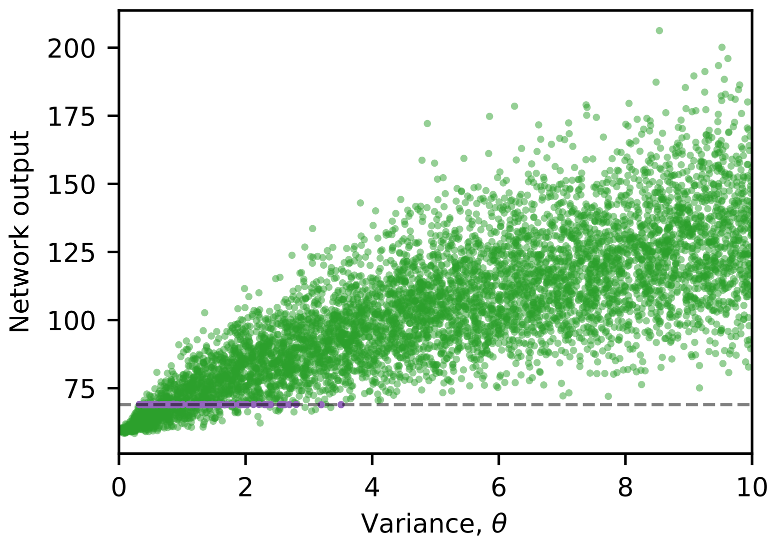
$$\boxed{\mathcal{L} = -\ln\det\langle F\rangle_{p(d|\theta_{\mathrm{fid}})}}$$

Charnock, Lavaux, Wandelt (arXiv:1802:03537)

Example 1: inference of variance

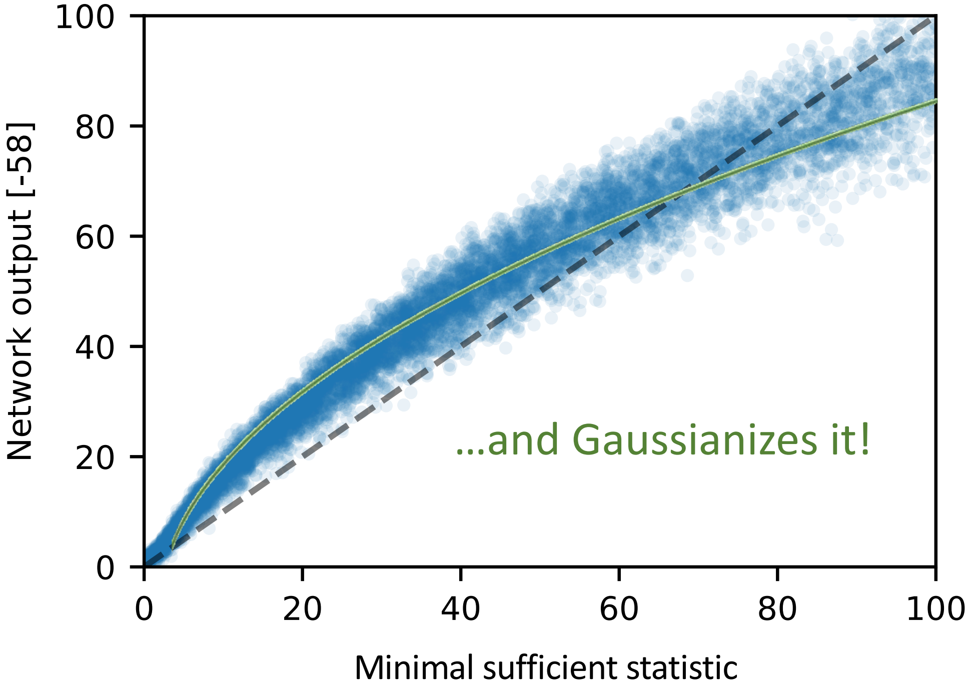
- Perfect information gives $|F| = 5$ in this problem
- Any linear summary gives $|\mathrm{F}| = 0.5$





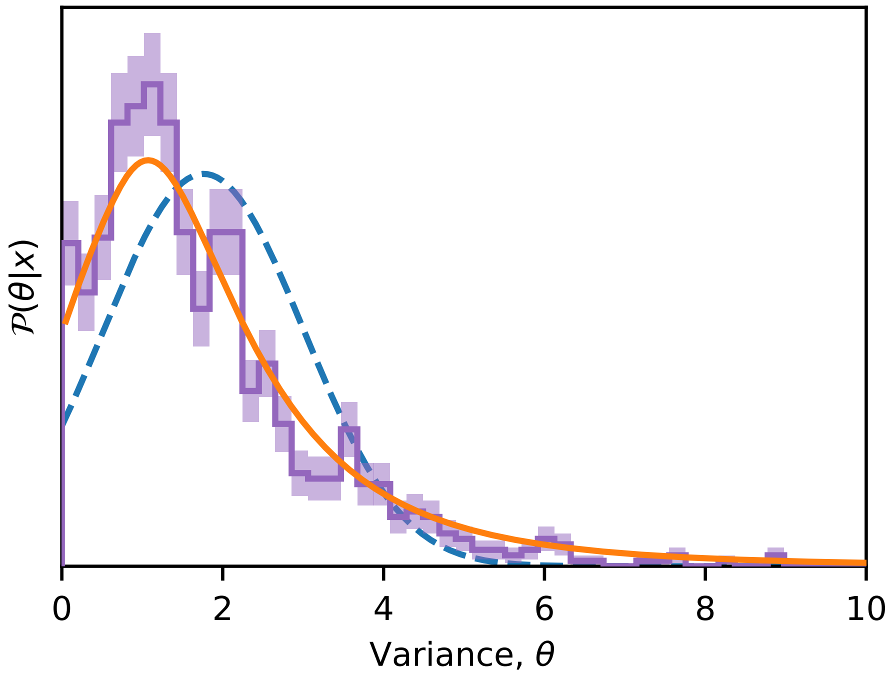
Charnock, Lavaux, Wandelt (arXiv:1802:03537)

The IMNN finds a minimal sufficient statistic for this inference problem



Charnock, Lavaux, Wandelt (arXiv:1802:03537)

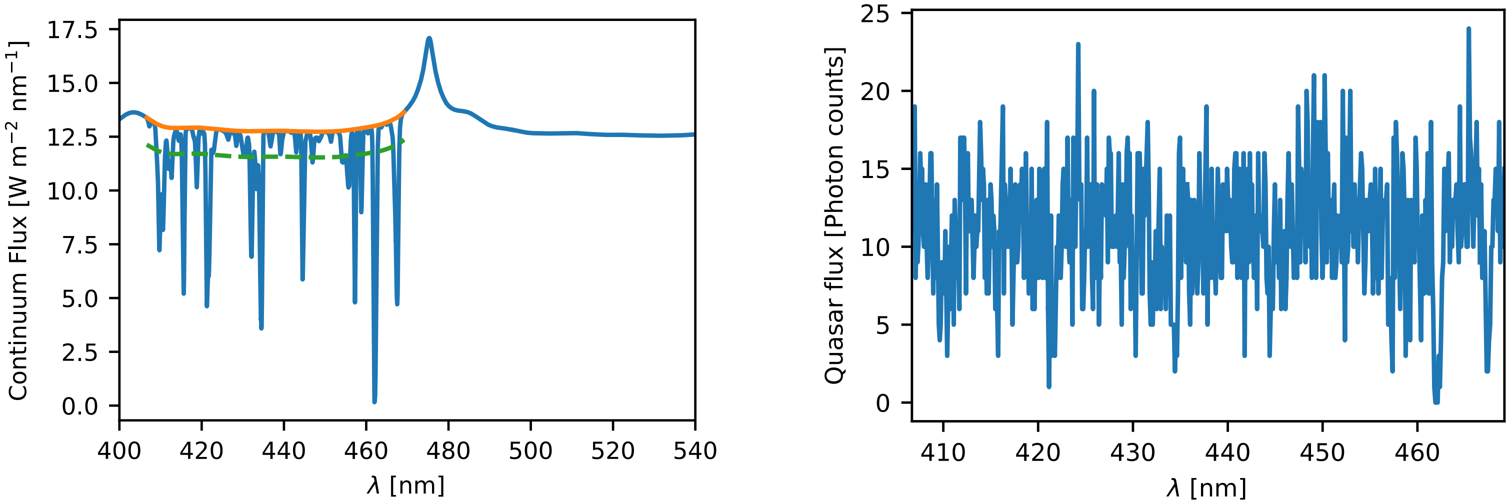
Example 2: Automatic physical inference with unknown noise



Charnock, Lavaux, Wandelt (arXiv:1802:03537)

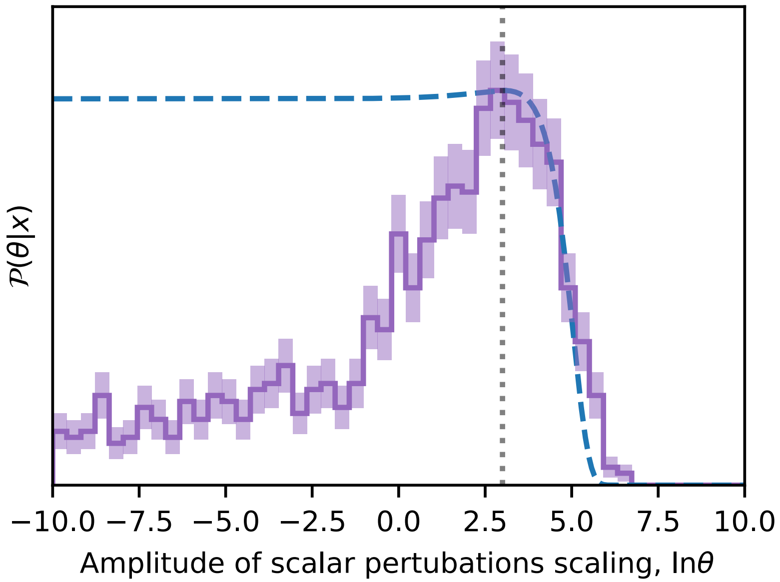
Example 3: Lyman-α forest inference

- The idea is to infer the variance of the underlying density field from a non-linearly transformed, photon-noise dominated Lyman-α forest spectrum



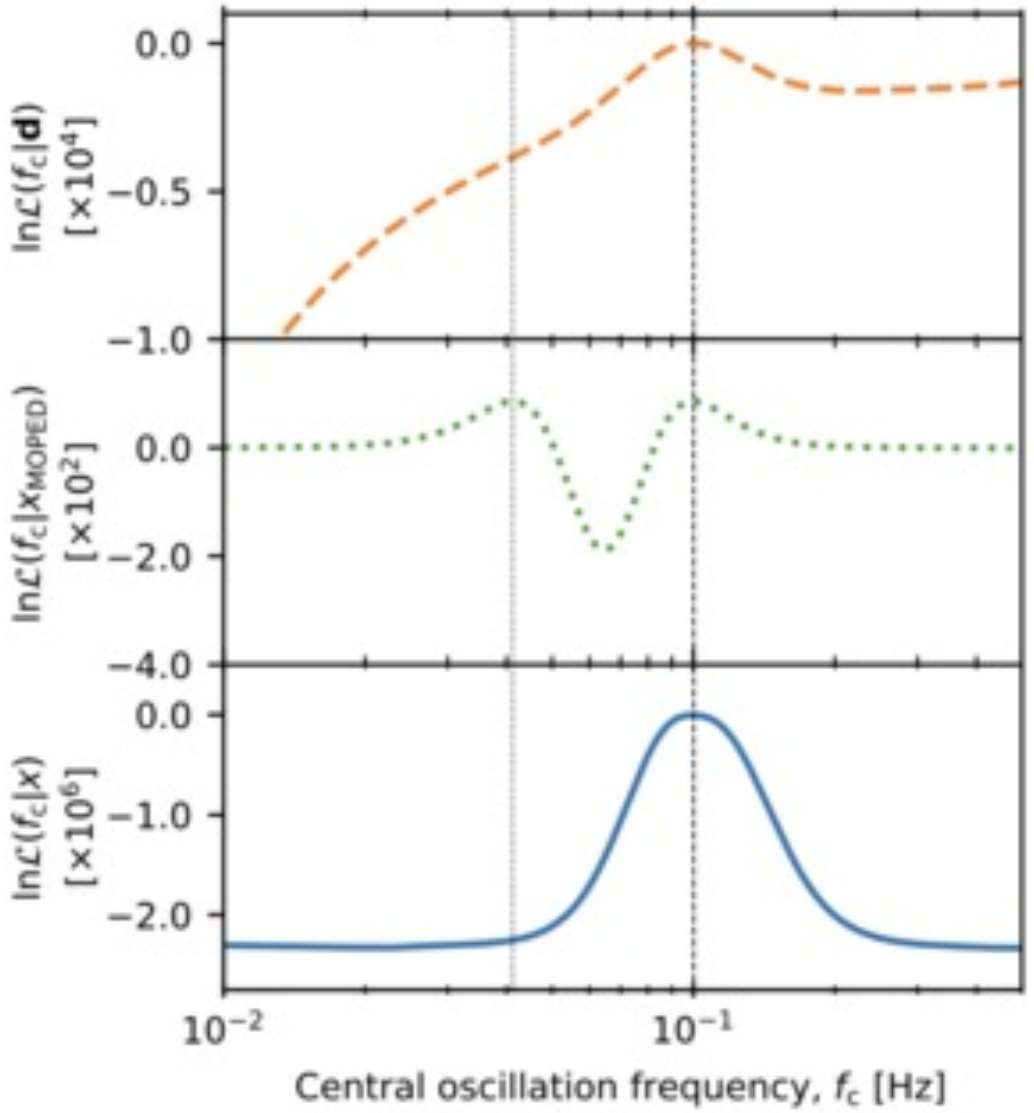
Charnock, Lavaux, Wandelt (arXiv:1802:03537)

Example 3: Lyman-α forest inference



Charnock, Lavaux, Wandelt (arXiv:1802:03537)

Example 4: excellent generalization from fiducial model



Infer frequency of LISA gravitational wave chirp

Full log-likelihood (LLH)

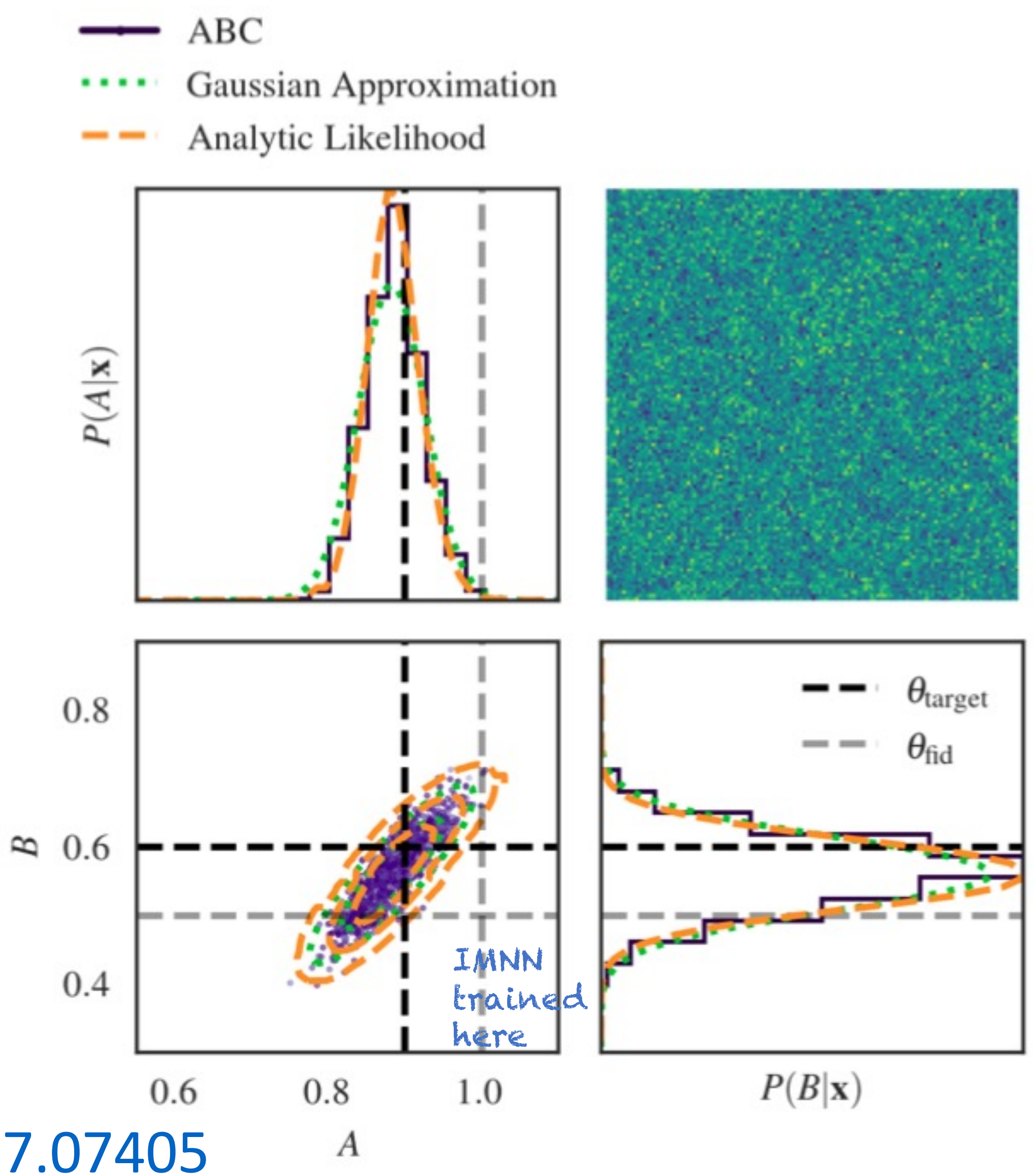
Gaussian LLH based on linear compression

Gaussian LLH based on IMNN compression

The Information Maximizing Network summary gives the correct unique likelihood peak.

Charnock, Lavaux, Wandelt (arXiv:1802:03537)

IMNN recovers full info directly from the field



IMNN posterior

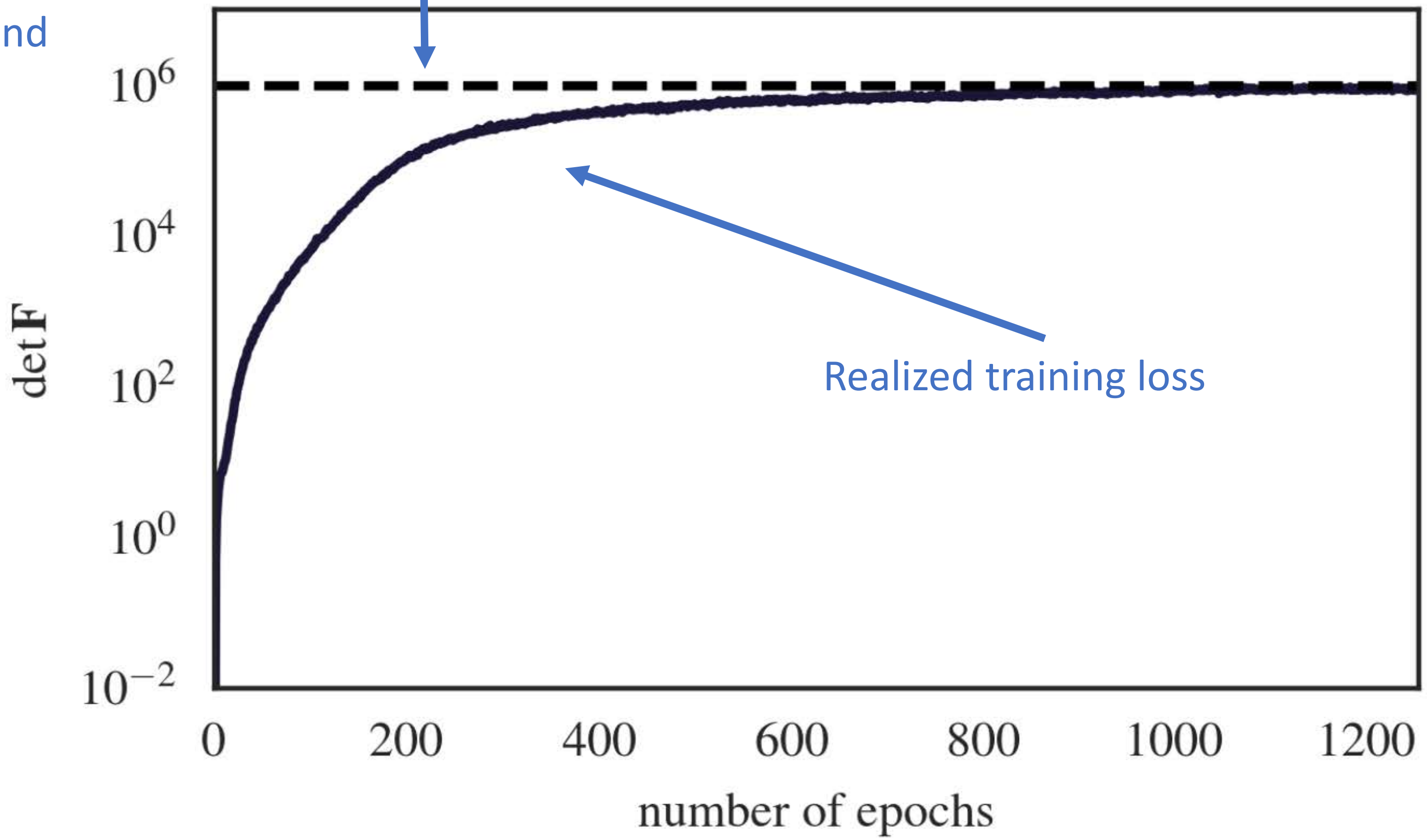
IMNN PML estimator

Exact likelihood

Makinen et al., arXiv:[2107.07405](https://arxiv.org/abs/2107.07405)

The IMNN recovers the full information

Theoretical (Cramer-Rao) information bound



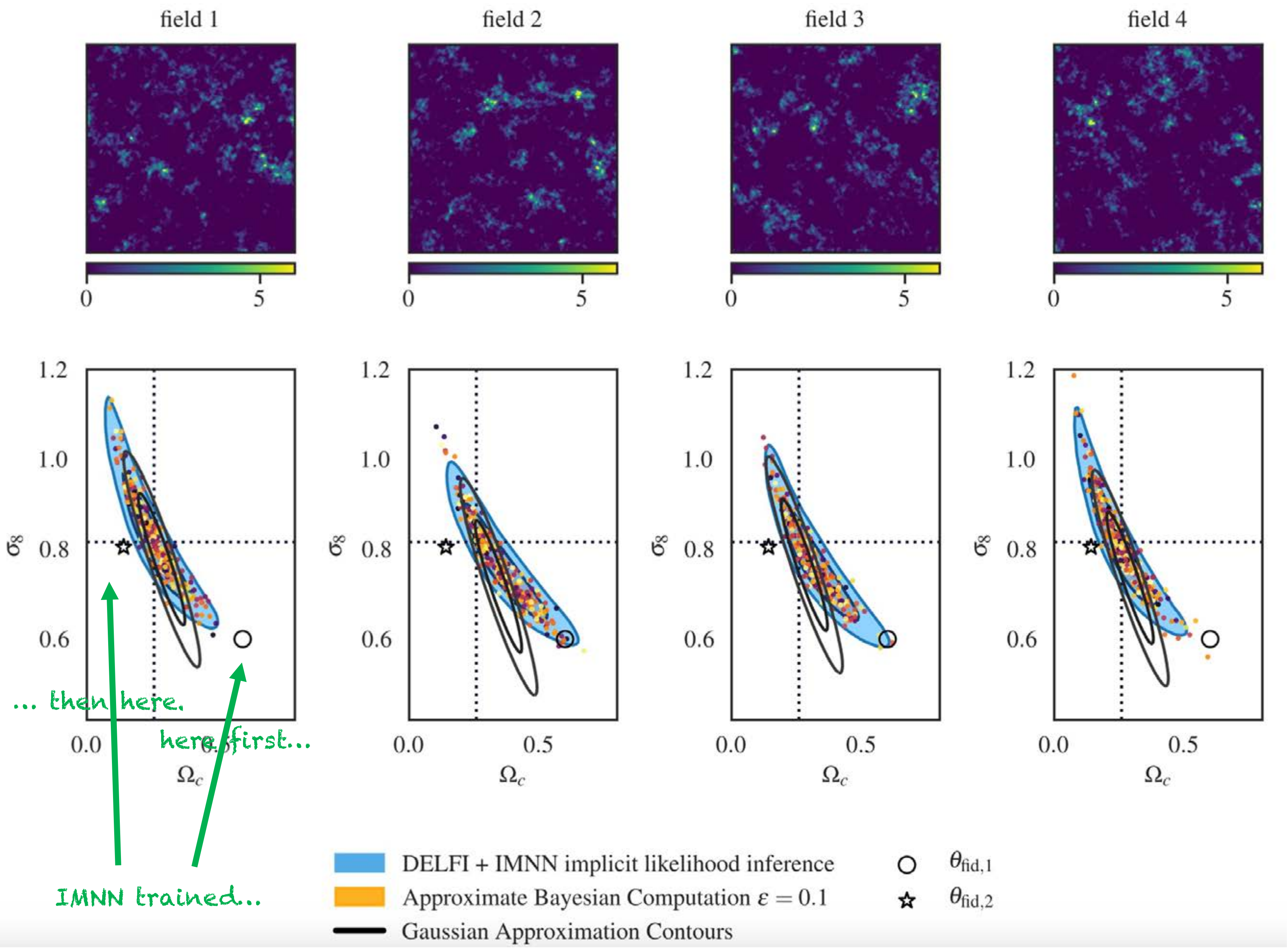
Realized training loss

11 minutes on 1 GPU

Makinen et al., arXiv:[2107.07405](https://arxiv.org/abs/2107.07405)

Non-Gaussian field inference with IMNN and DELFI

Available as interactive notebook tutorial at https://bit.ly/imnn-cosmo

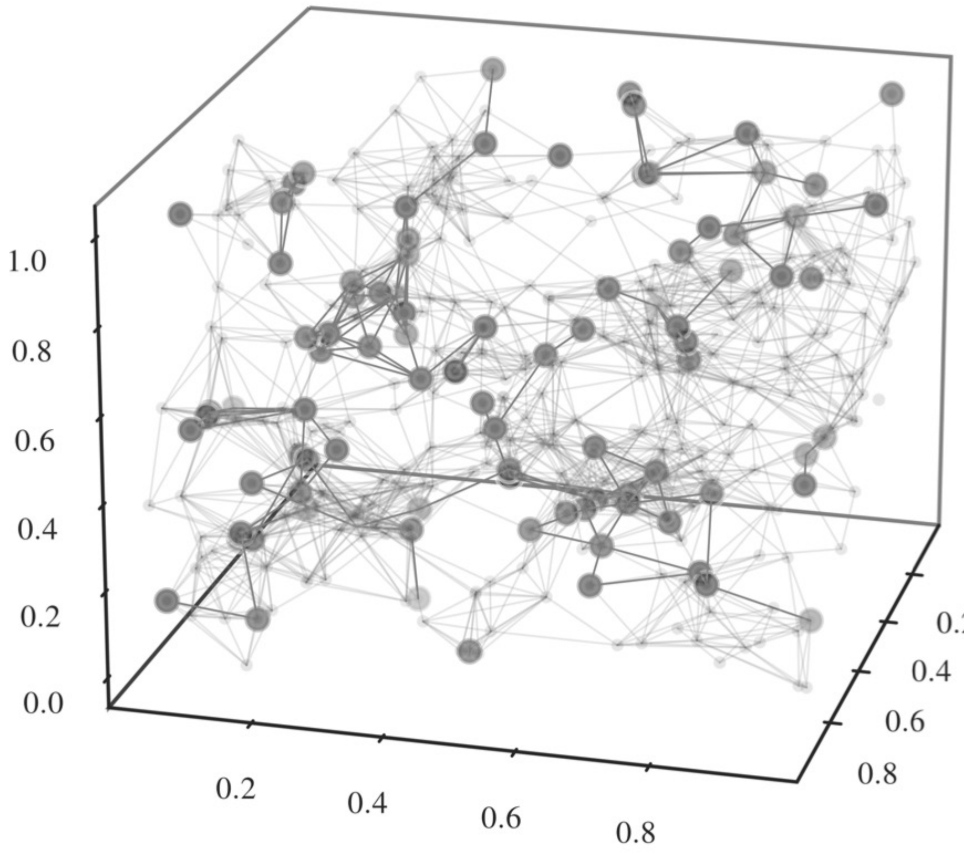


Makinen et al., arXiv:2107.07405

Can define both information matrix and score function on distributions of graphs

Example of using clusters of galaxies to infer cosmological parameters

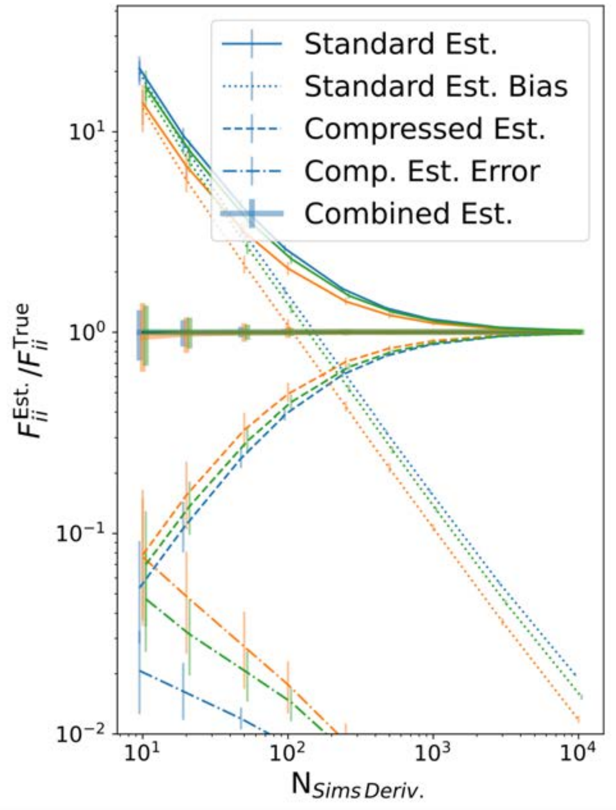
Uses neurally derived Fisher score within pyDELFI.



Makinen et al. arXiv:2207.05202

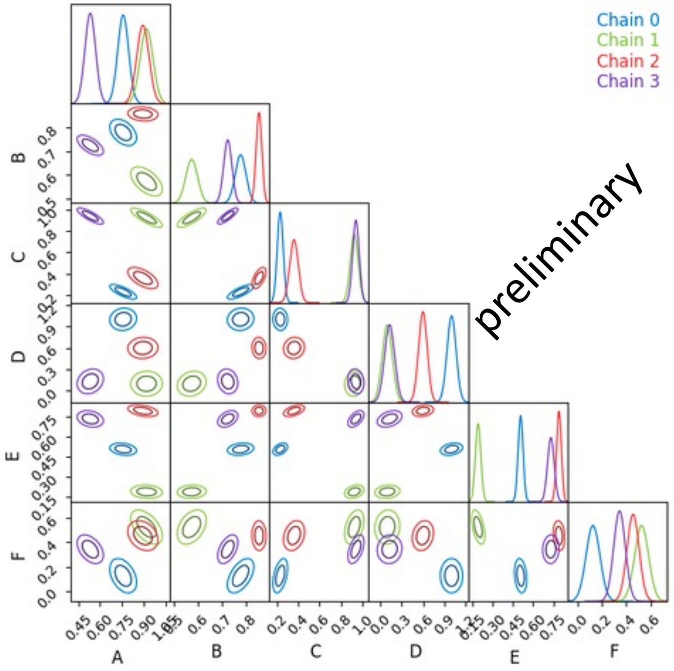
Other ways to compute the Information Matrix *implicitly*

- Coulton & Wandelt (arXiv:2305.08994): *A new, efficient estimator of Fisher Information from simulations*



- Typically, compute at a fiducial parameter point.
  Can we compute Fisher information efficiently ***everywhere*** *in parameter space?*

➔Fishnets (Makinen et al, in prep)



Reasoning about models with Bayesian Machine Learning

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

Actually, $p(d|M_i)$

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***

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

Niall's talk tomorrow!

Sampling high-dimensional posteriors in Implicit Inference

and

Information-Ordered Bottlenecks

Benjamin D. Wandelt

Ronan Legin (Montreal), Niall Jeffrey (UCL), Matt Ho (IAP), Xiaosheng Zhang (Tsinghua), Gabriel Jung (IAS, Orsay), Lucas Makinen (Imperial), Will Coulton (CCA), Stephen Feeney, ...

“Score”-based Diffusion

- Consider a random walk of images
- Initialise with initial conditions
- Add Gaussian noise at every step
- Central limit theorem: this has an *attractor* a Gaussian noise distribution

- Then sample by solving 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

Exploring high-dimensional posterior pdfs with “score”-based diffusion

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

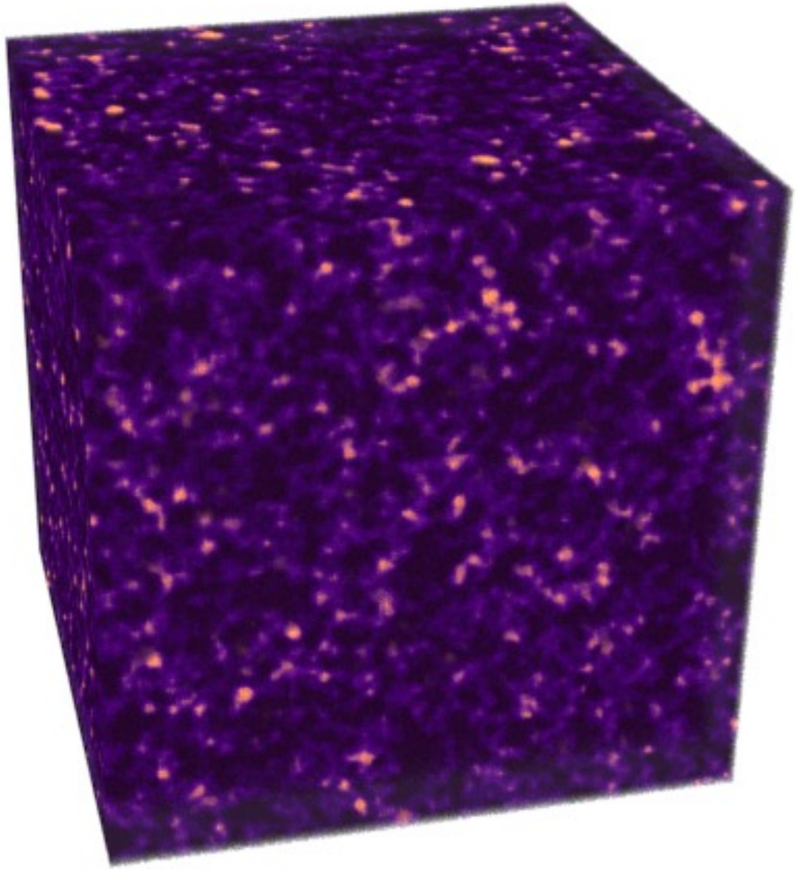
Train the “score” on QUIJOTE n-body simulations

- 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**

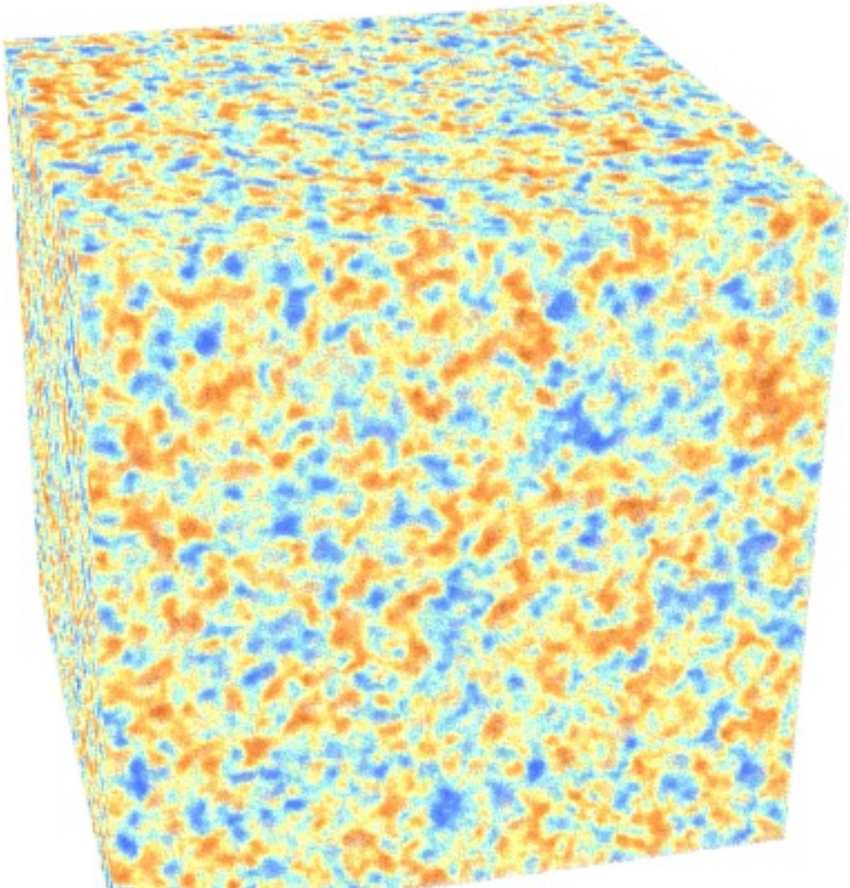
First full-field inference of initial conditions from fully non-linear density field

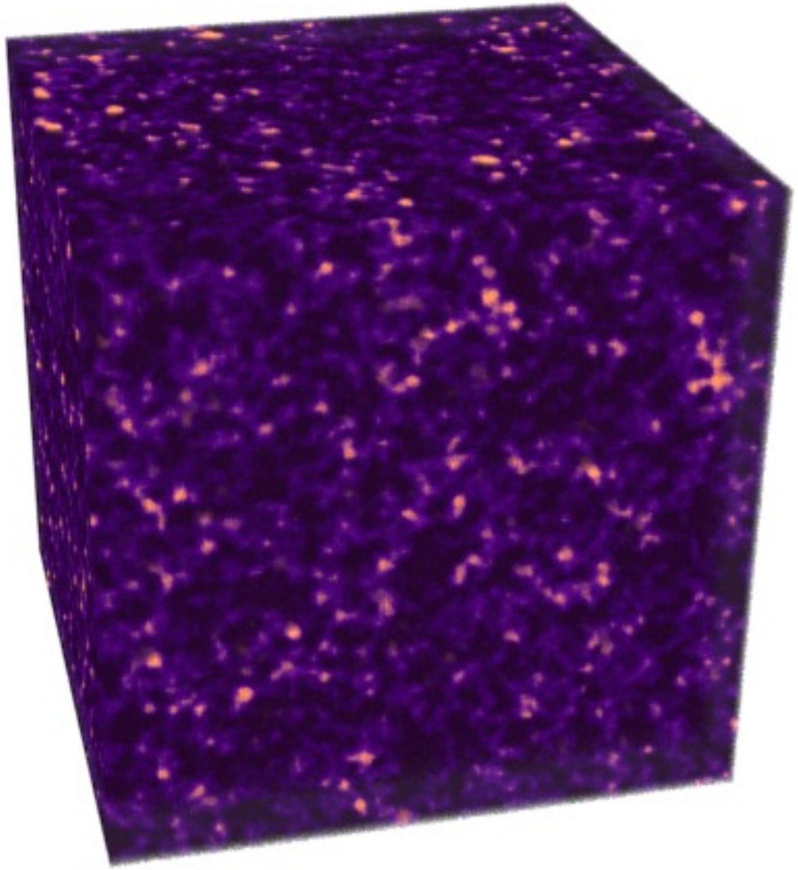
- 1 Gpc GADGET1024^3 simulation at z=0
- Binned on 128^3 grid



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

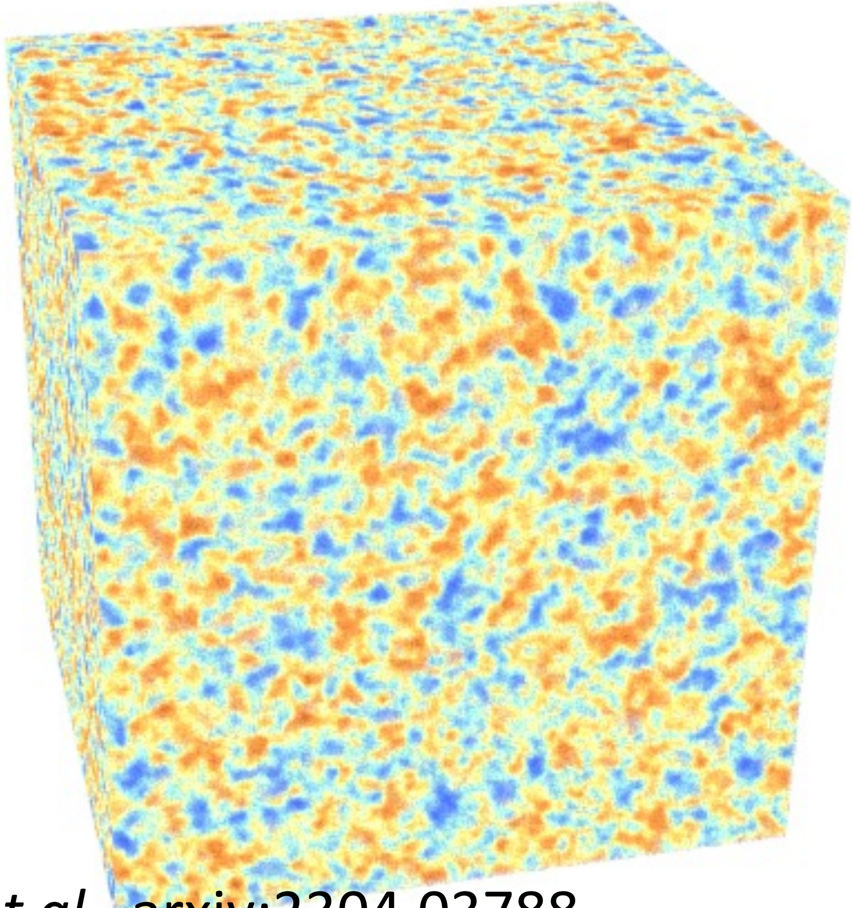
First full-field inference of initial conditions from fully non-linear density field

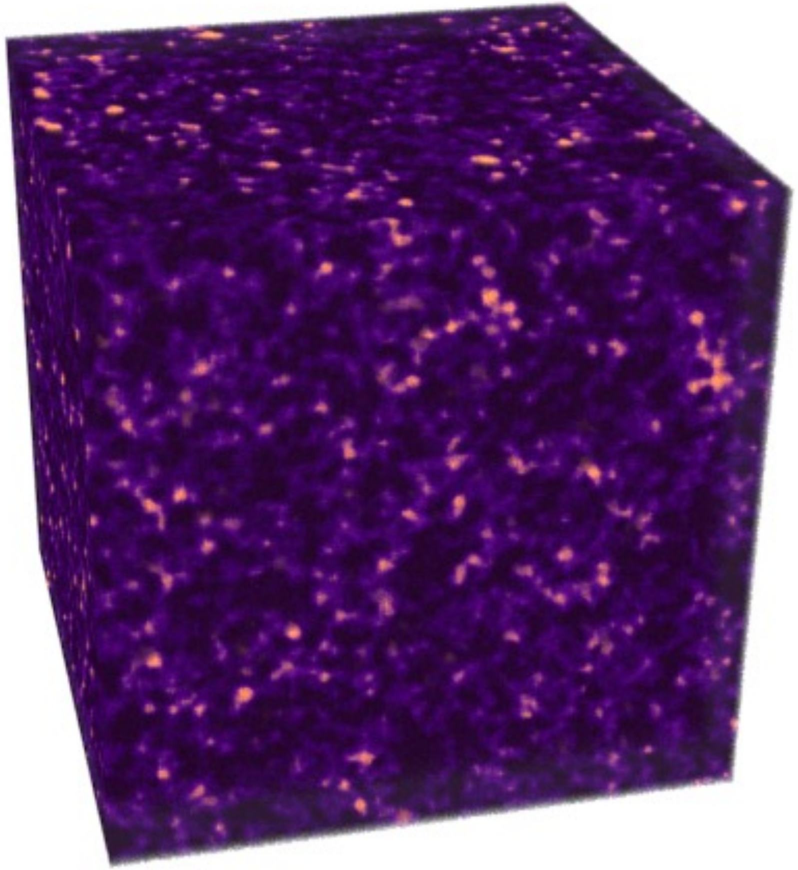




R. Legin *et al.*, arxiv:2304.03788

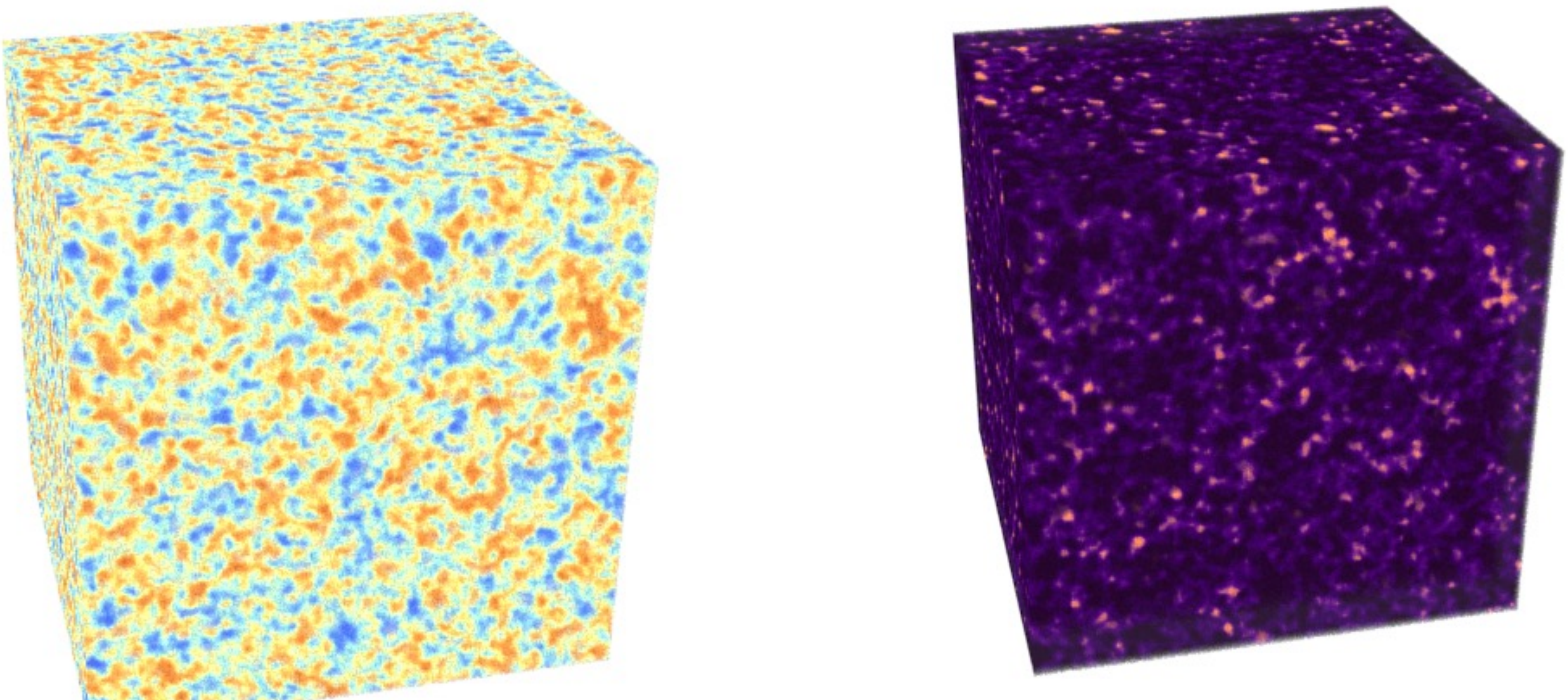
First full-field inference of initial conditions from fully non-linear density field





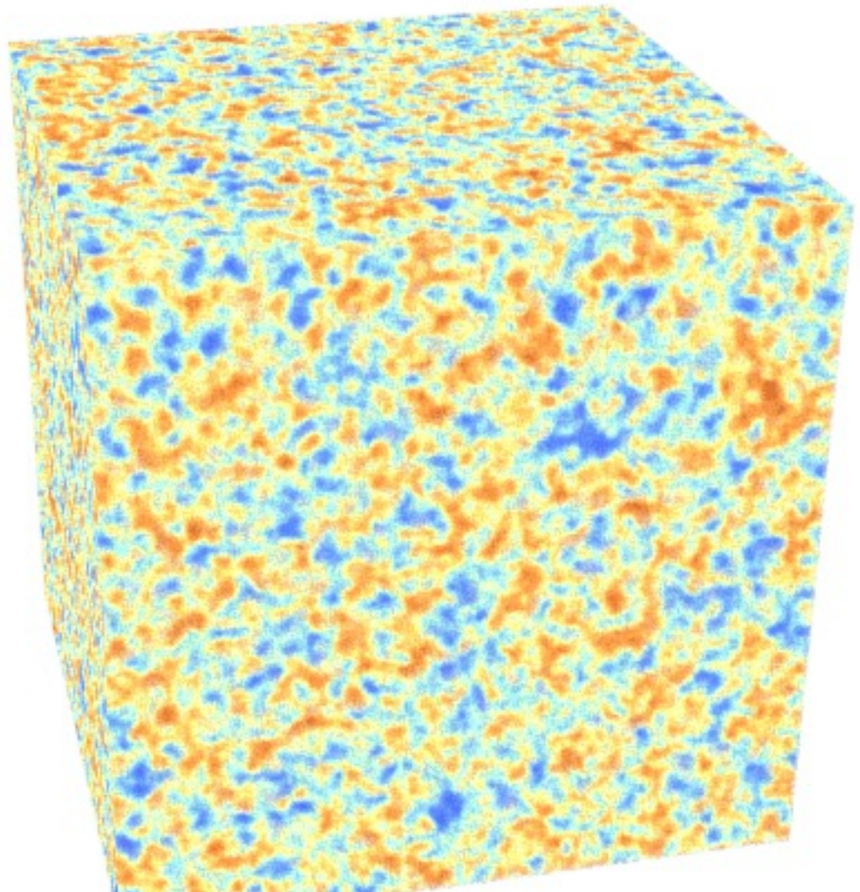
R. Legin *et al.*, arxiv:2304.03788

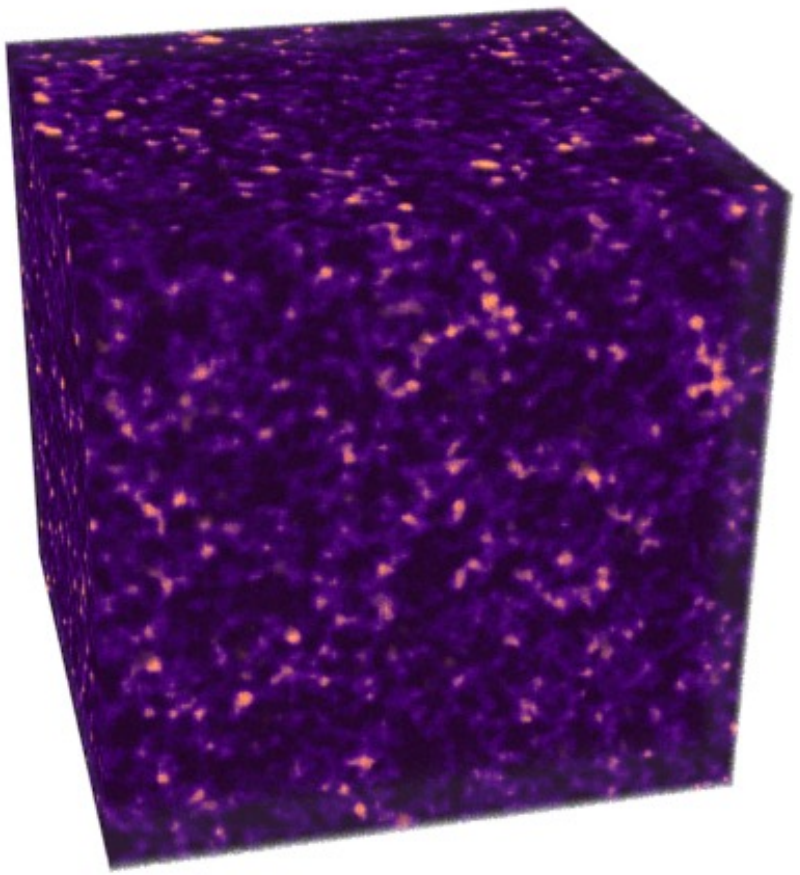
First full-field inference of initial conditions from fully non-linear density field



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

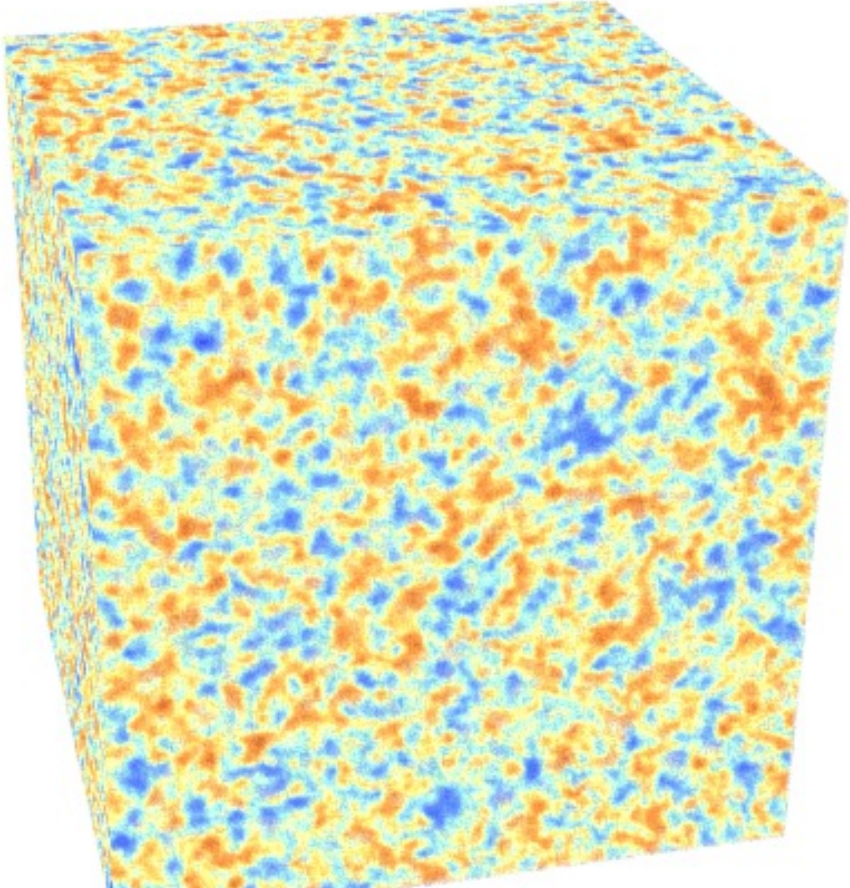
First full-field inference of initial conditions from fully non-linear density field

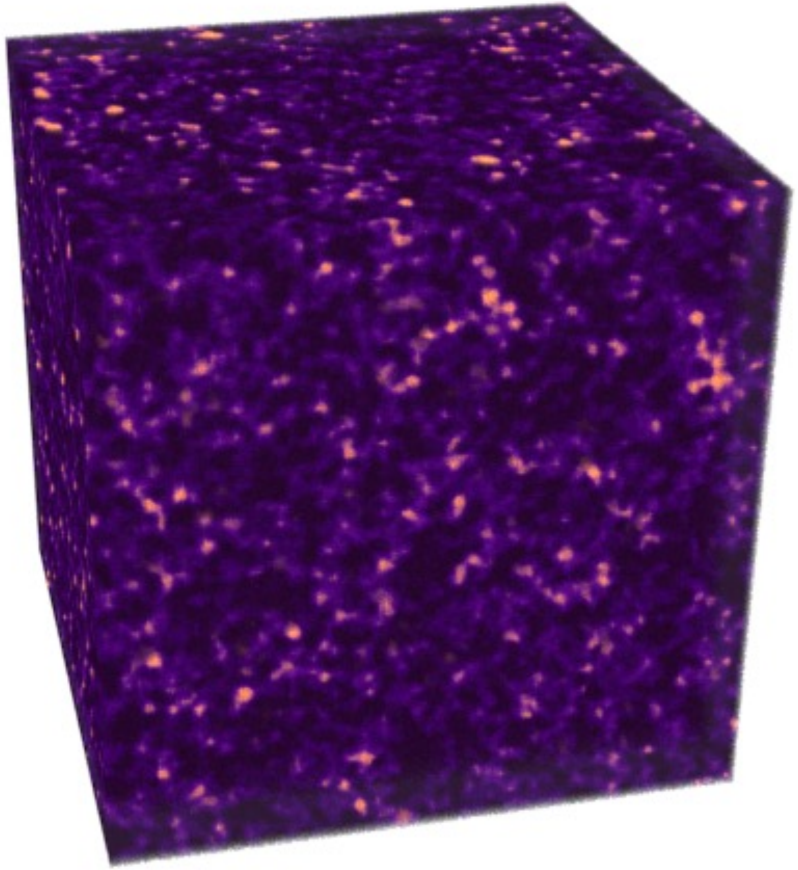




R. Legin *et al.*, arxiv:2304.03788

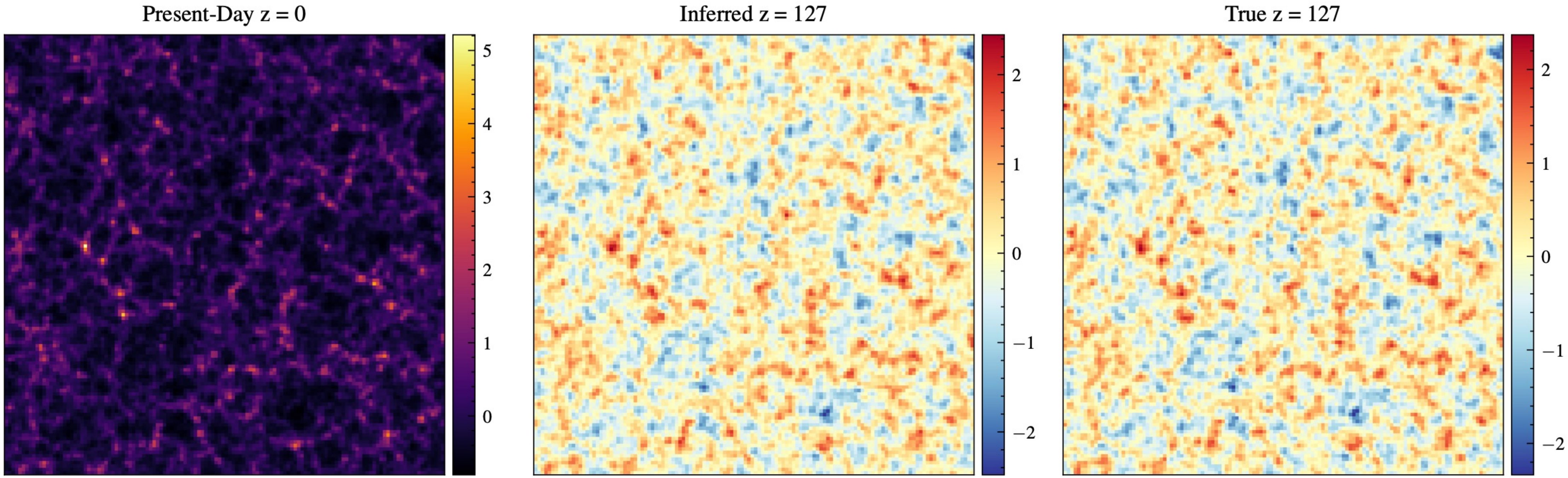
First full-field inference of initial conditions from fully non-linear density field





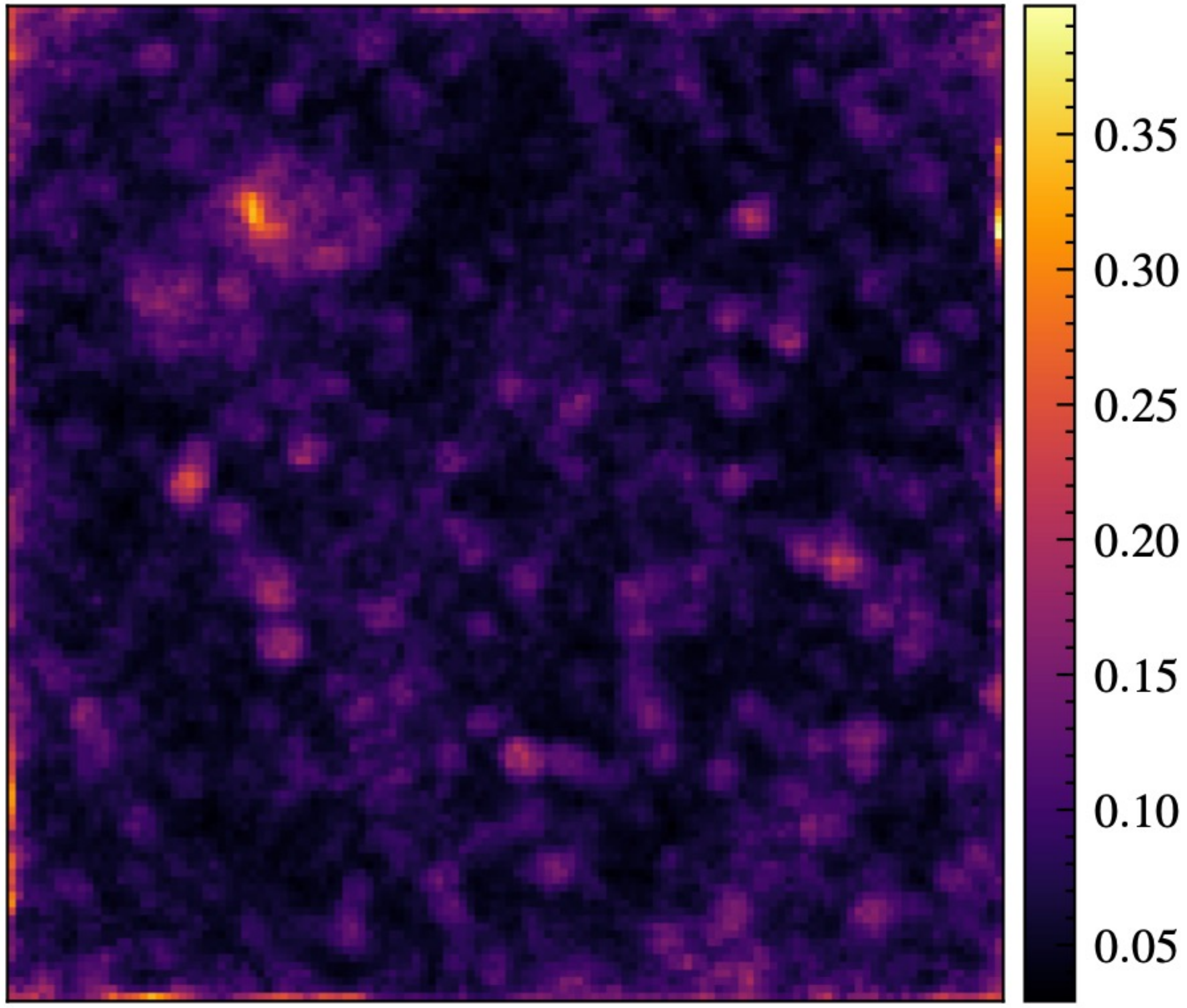
R. Legin *et al.*, arxiv:2304.03788

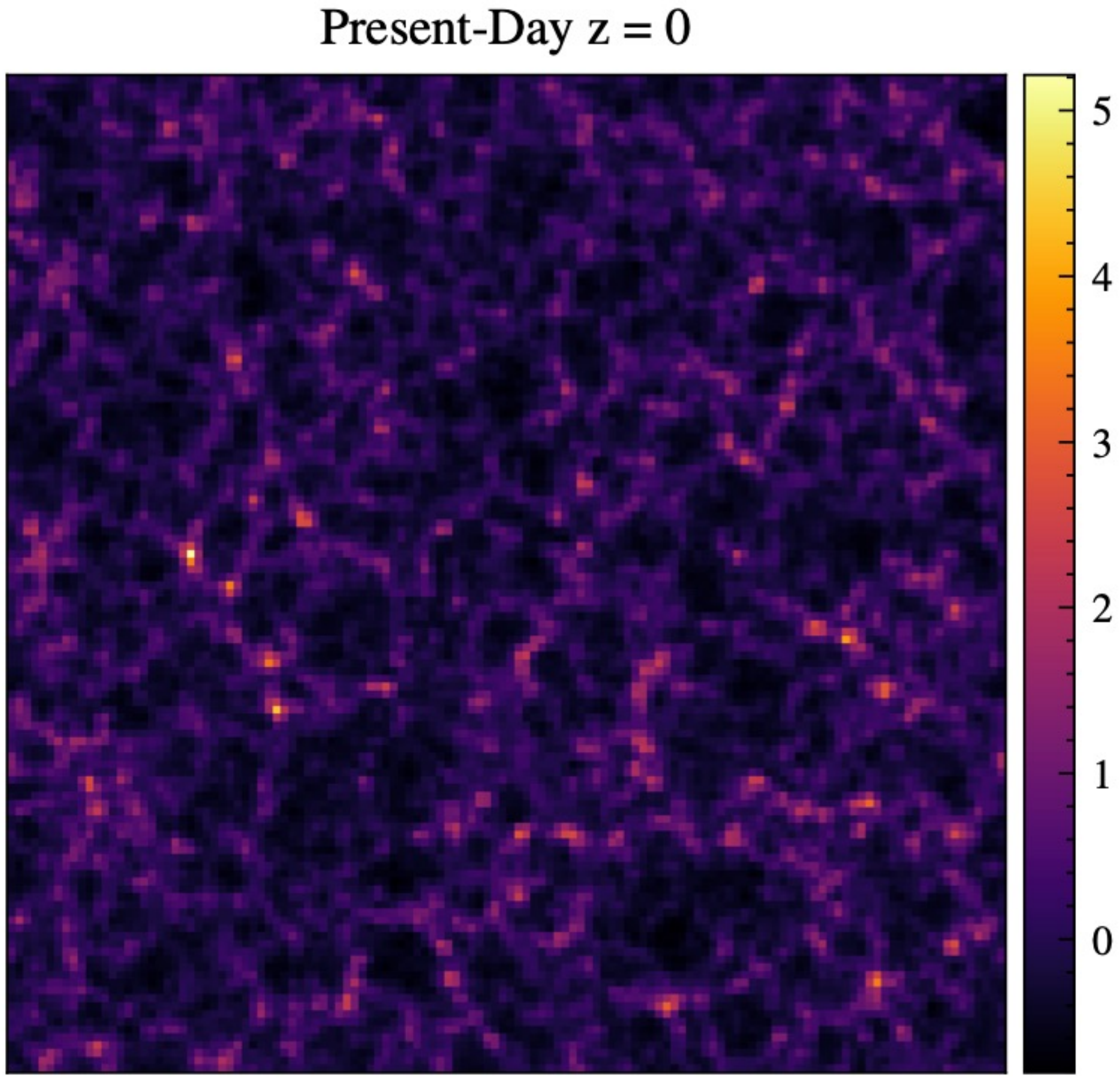
Faithful reconstruction…



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

… including uncertainties (posterior variance)



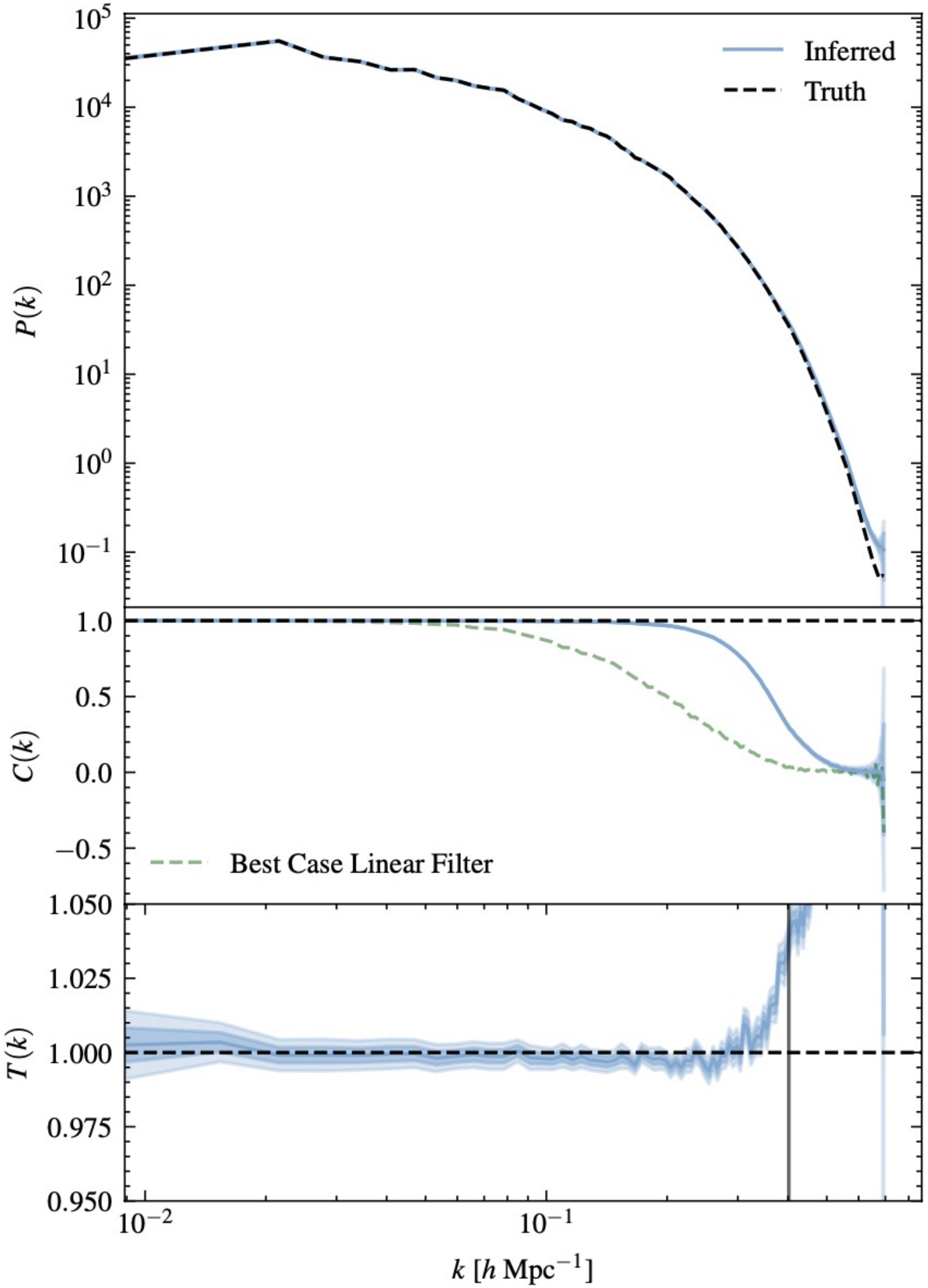


R. Legin *et al.*, arxiv:2304.03788

Accurate reconstructions

Points to note:

- full non-linear gravity
- No need for differentiability



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

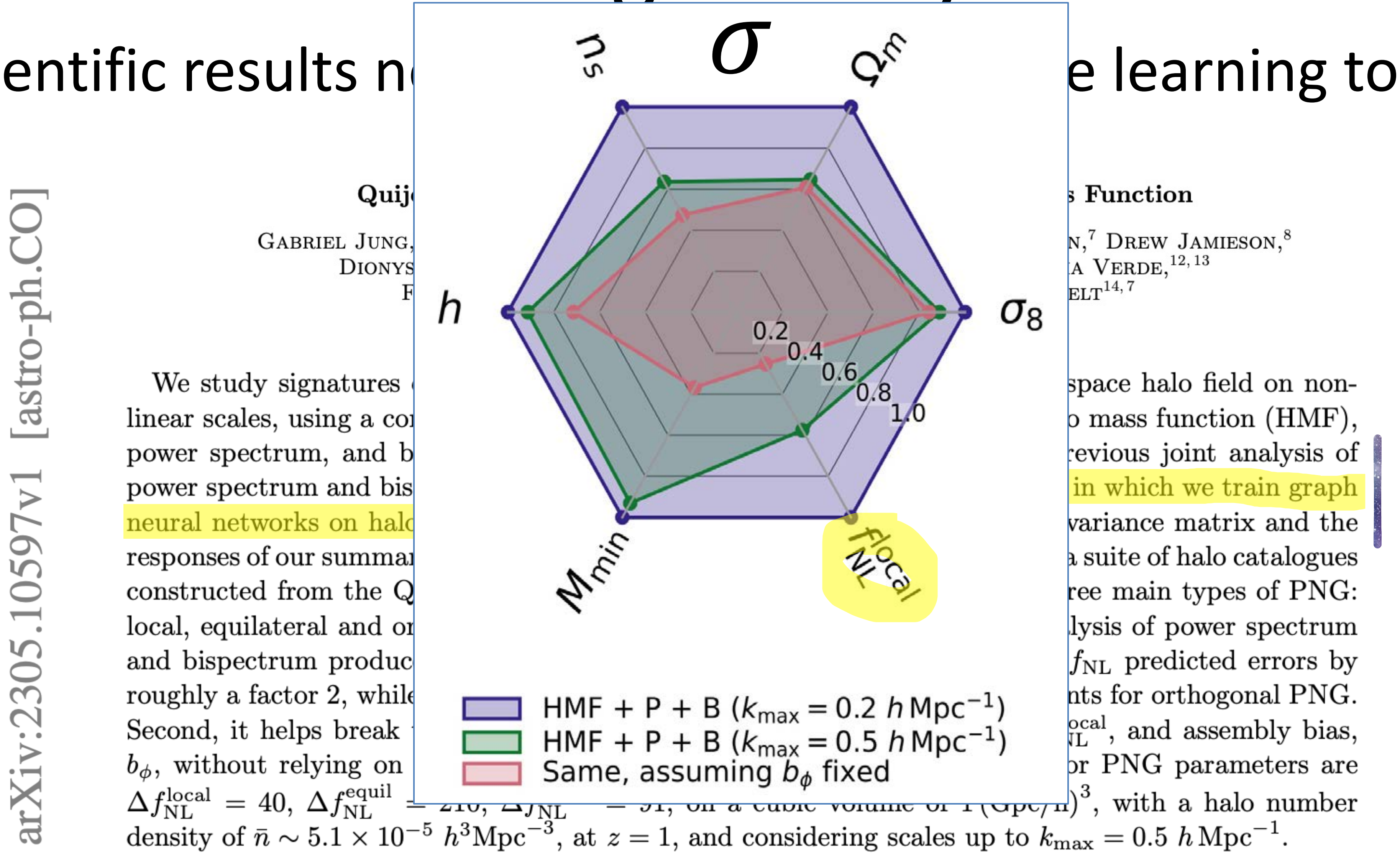
The principal impact of ML on cosmology so far

An extremely powerful tool to answer the question

“Does **A** contain information about **B**?”

Theoretical discovery example: breaking degeneracy

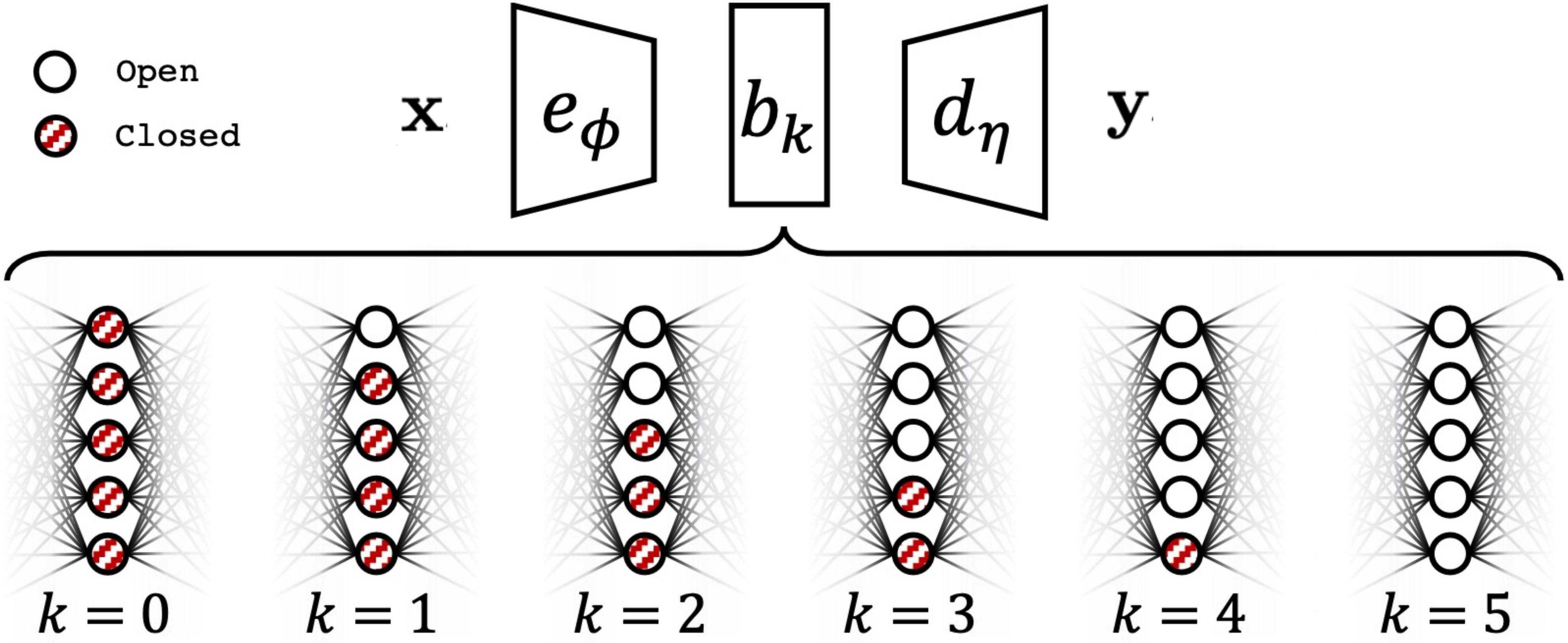
- Scientific results n… …e learning tools



From information to insight

Can we use machine learning to discover the most important degrees of freedom in data?

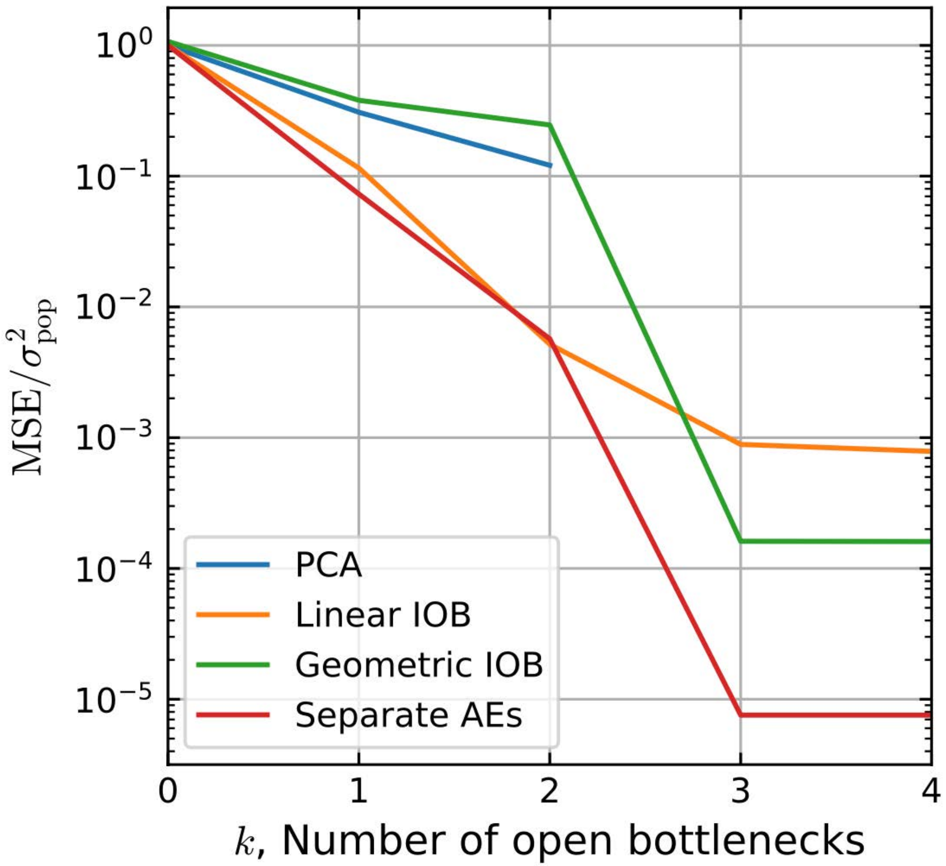
The Information-Ordered Bottleneck

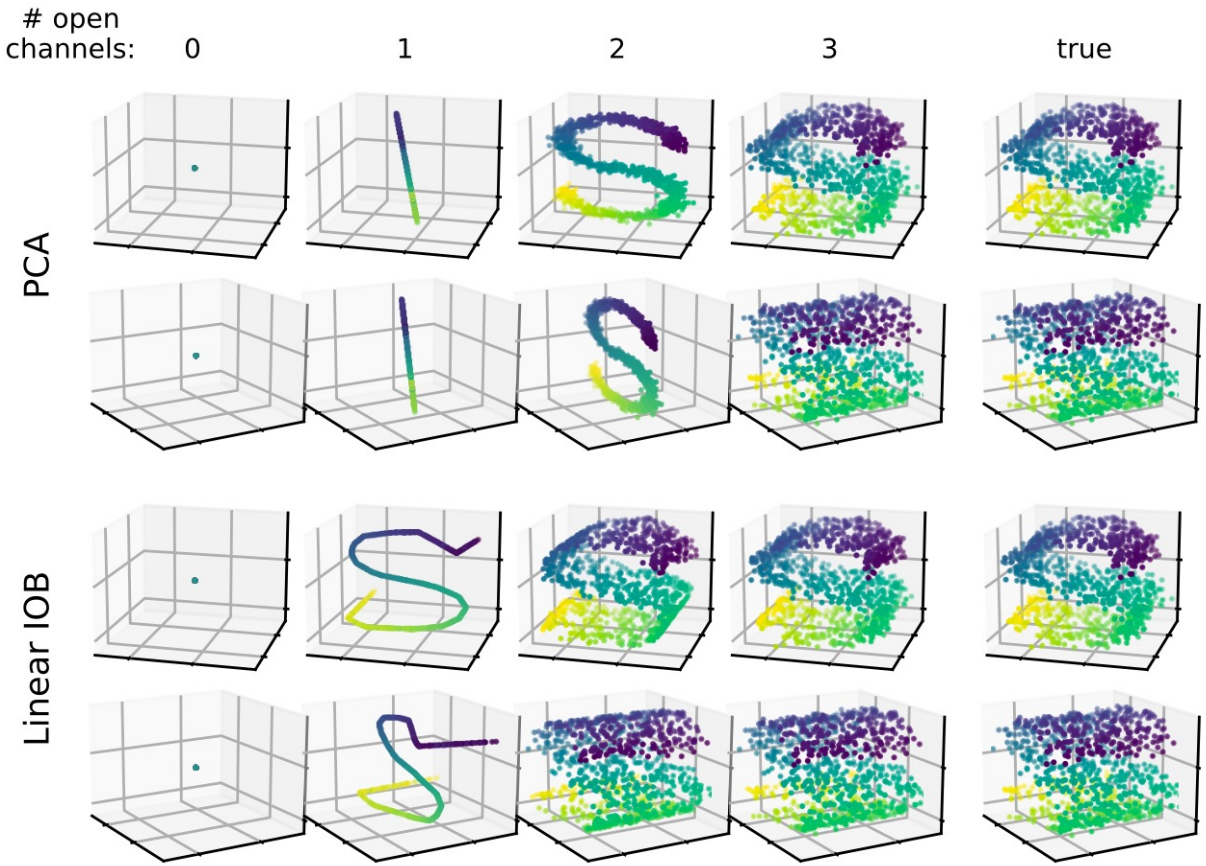


$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{N} \sum_{k=0}^{k_{\max}} \rho_k \ell \left[f_\theta^{(k)} (\mathbf{x}_i) , \mathbf{y}_i \right]$$

Ho, Zhao & Wandelt arXiv:2305.11213

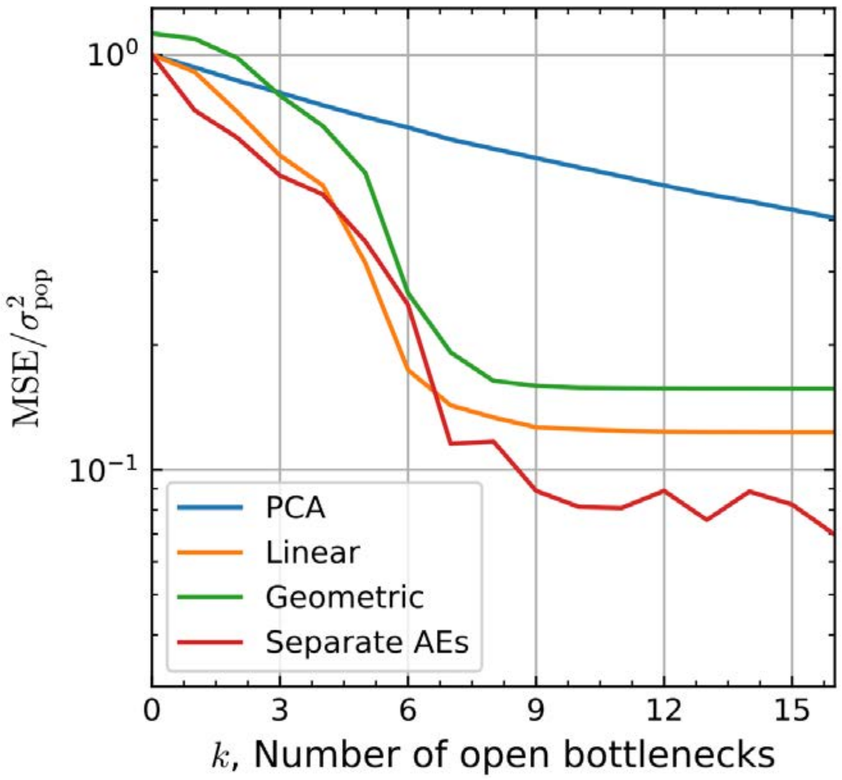
The IOB orders latents by information

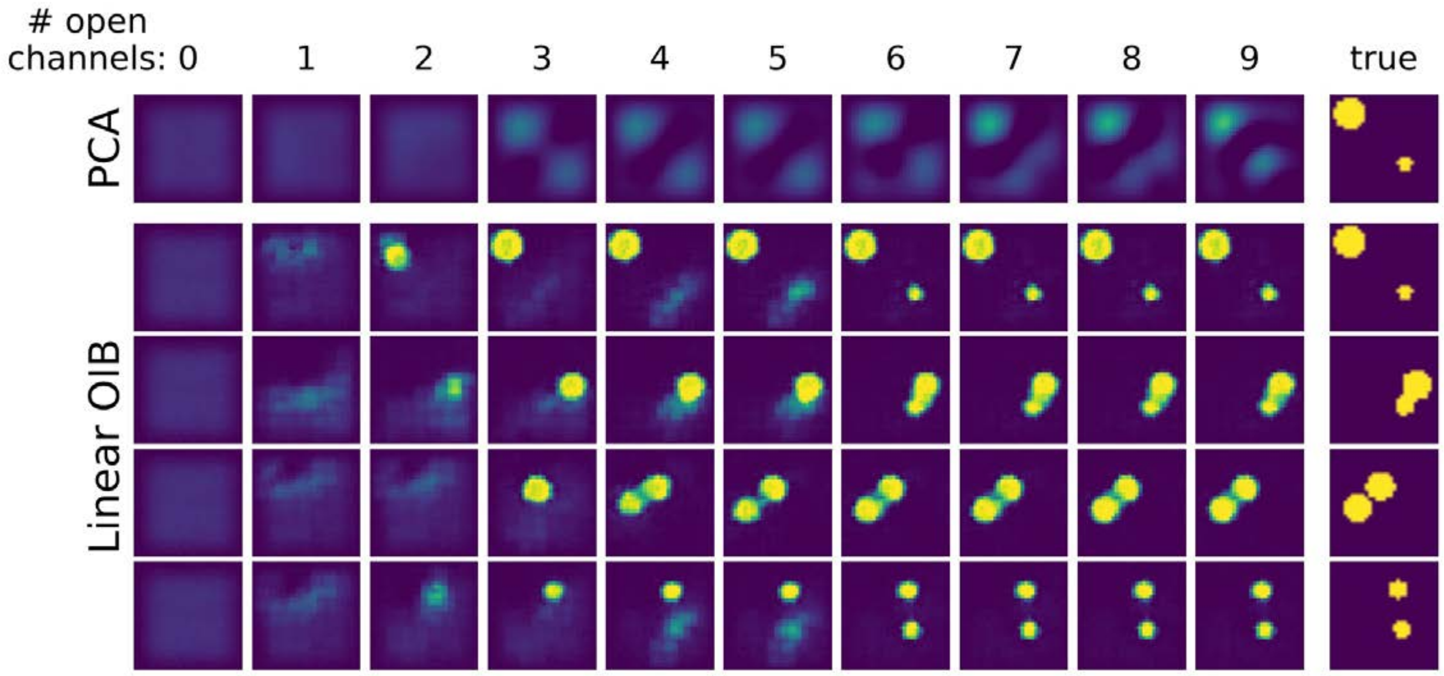




Ho, Zhao & Wandelt arXiv:2305.11213

The IOB discovers the number of degrees of freedom in data





Data: 2 disks of arbitrary size

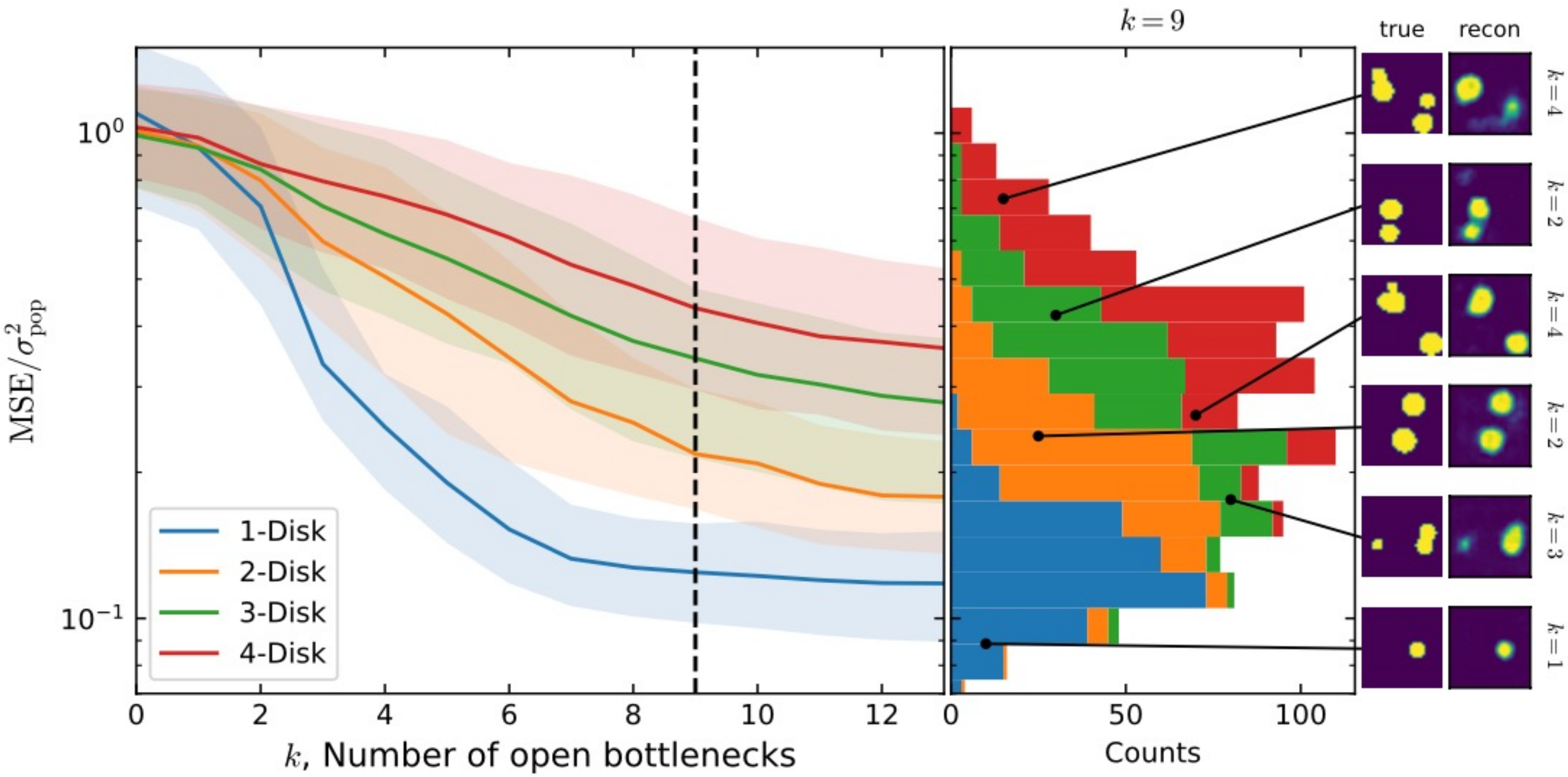
Ho, Zhao & Wandelt arXiv:2305.11213

Information Ordered Bottleneck: SOTA discovery of global intrinsic dimensionality

ID Estimator	S-curve	1-Ball	2-Ball	3-Ball	4-Ball	MS-COCO CLIP
PCA [23]	3	33	37	39	38	106
MADA [25]	2.5	inf	13.2	16.9	19.5	22.7
TwoNN [24]	2.9	5.3	13.6	16.3	21.4	21.4
Linear IOB*	**2**	**3**	**7**	**10**	14	322
Geometric IOB*	3	**3**	**7**	**10**	**12**	196
Data Dimensionality	3	1024	1024	1024	1024	768
True Dimensionality	2	3	6	9	12	≤ 768

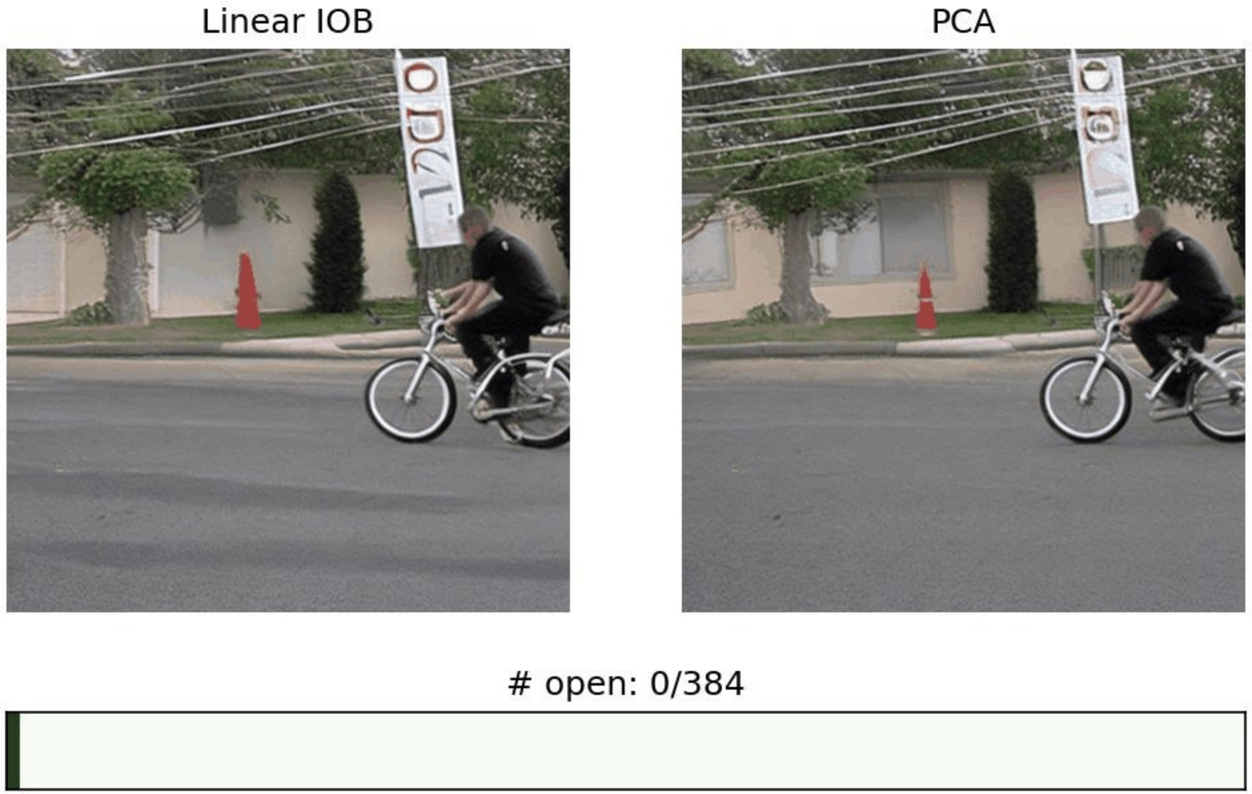
Ho, Zhao & Wandelt arXiv:2305.11213

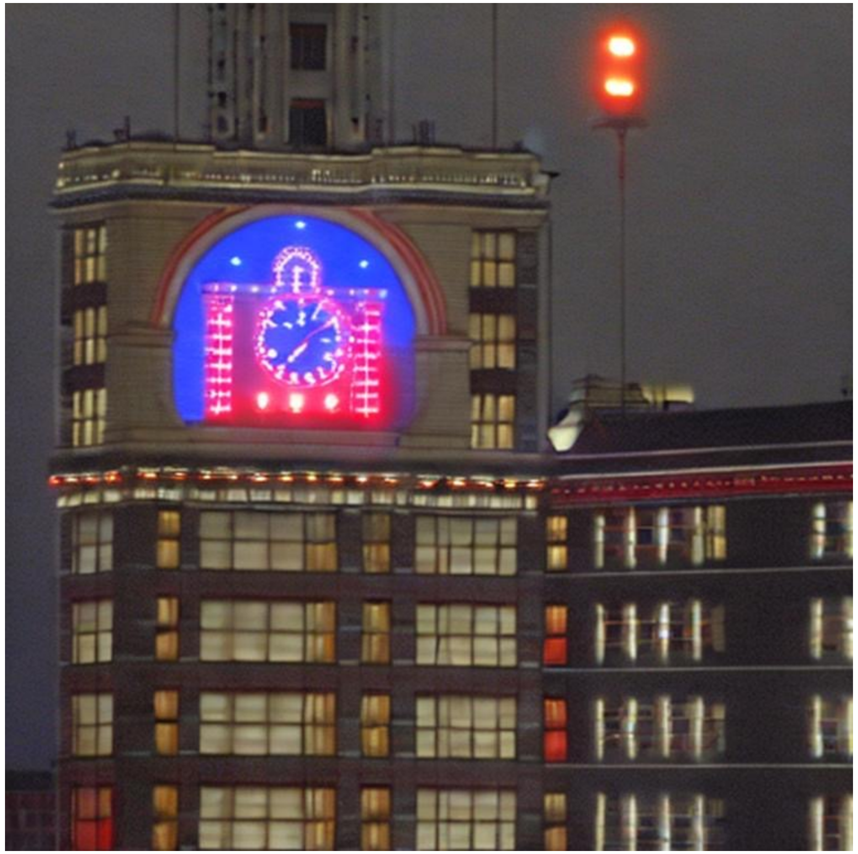
Training on heterogenous data sets can classify individual instances by complexity



Ho, Zhao & Wandelt arXiv:2305.11213

Adaptive IOB Compression in Semantic Latent Space



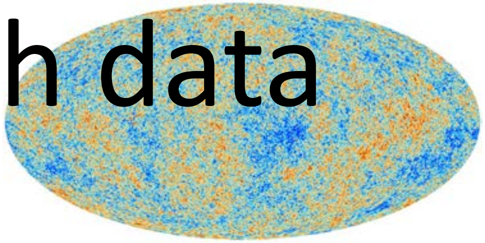


Information-Ordered Bottleneck applied to CLIP embeddings.

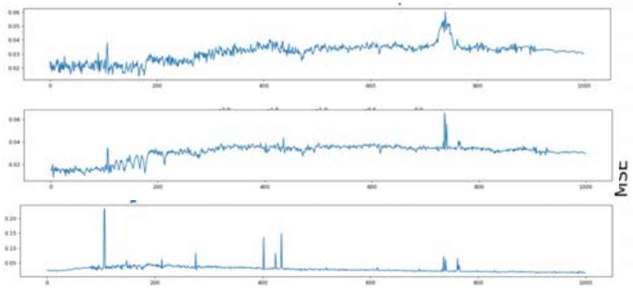
Dataset: MS-COCO

Check out Ho, Zhao & Wandelt arXiv:2305.11213 for more examples!

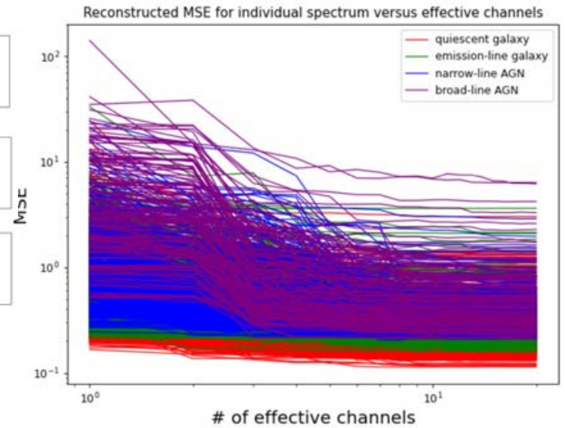
Information-Ordered Bottlenecks for insight in a world exploding with data



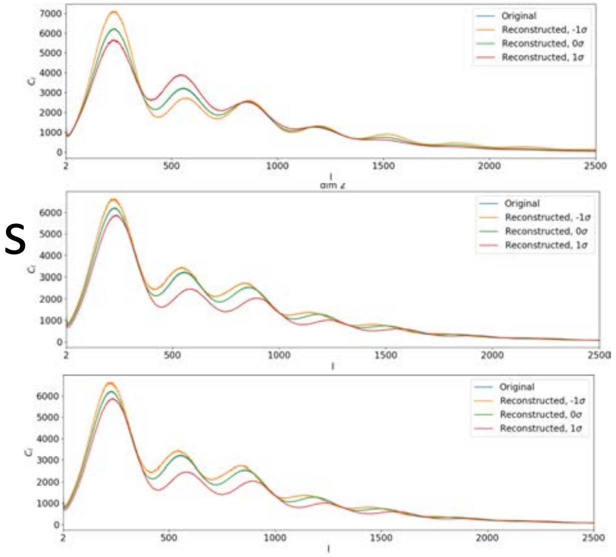
What is the (relative) complexity of Galaxies and Active Supermassive Black Holes spectra?



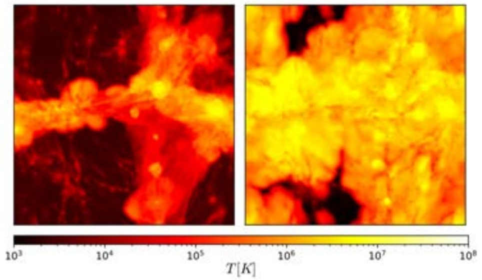
...



What parameter combinations does the CMB really constrain?



CAMELS



What are natural coordinates for the parameter space of cosmological hydro-simulations?