Navigating the Cosmos: An Overview of Bayesian and Machine Learning Approaches in Modern Cosmology

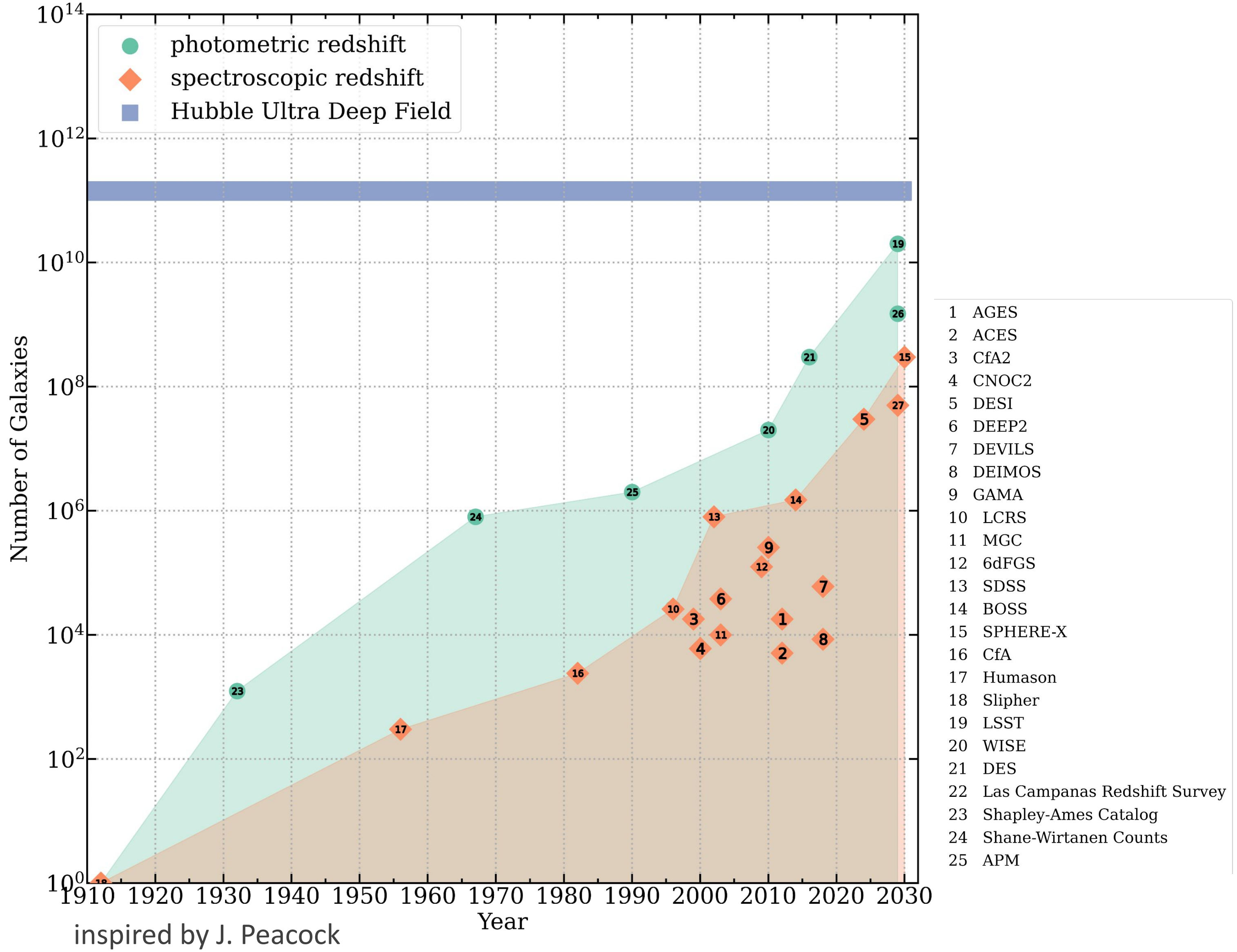
Debating the Potential of Machine Learning in Astrophysical Surveys #2
November 29, 2023
IAP, Paris / Flatiron institute, New York

Jens Jasche

01 Motivation

Next-generation cosmological data is becoming available

A CENTURY OF GALAXY SURVEYS

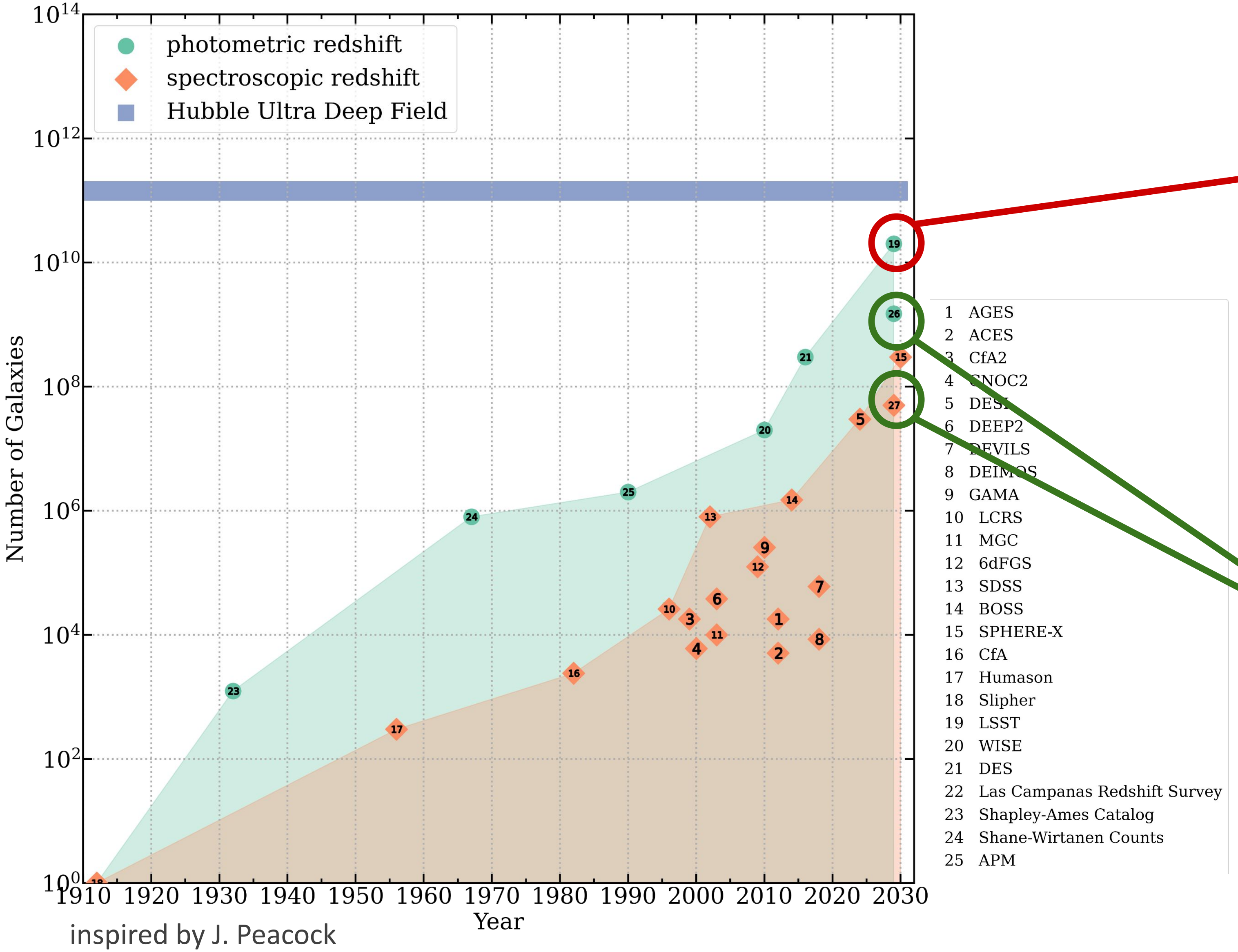


inspired by J. Peacock

credit for plotting scripts: Florent Leclercq

Next-generation cosmological data is becoming available

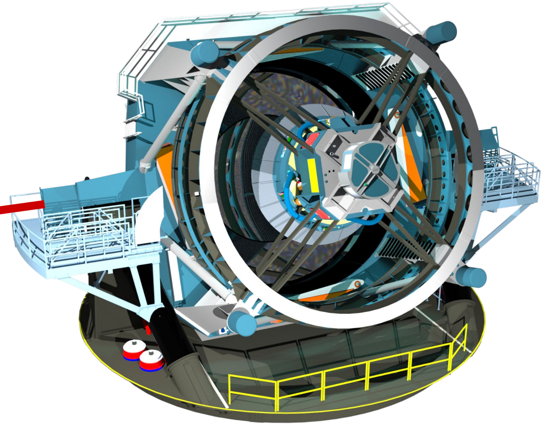
A CENTURY OF GALAXY SURVEYS



inspired by J. Peacock

credit for plotting scripts: Florent Leclercq

Large Synoptic Survey Telescope



ESA's Euclid Spacecraft

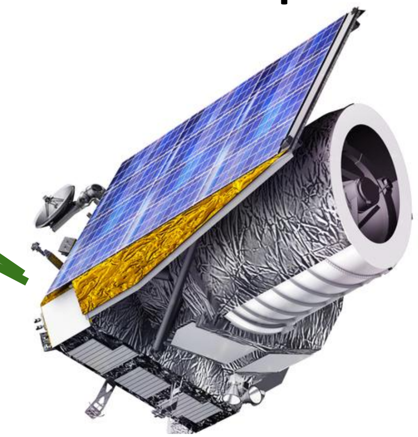


Image credit: ESA

The Promise of Bayesian inference for astrophysics (Loredo 1992)

Challenges of Cosmological Inference:

- Cosmology an observational science.
- Phenomena are inaccessible to direct manipulation.
- Cosmological inferences are generally uncertain.

The Bayes Inference Device:

$$\pi\left(\mathbf{x}\middle|\mathbf{d}\right) = \pi\left(\mathbf{x}\right)\frac{\pi\left(\mathbf{d}\middle|\mathbf{x}\right)}{\pi\left(\mathbf{d}\right)}$$

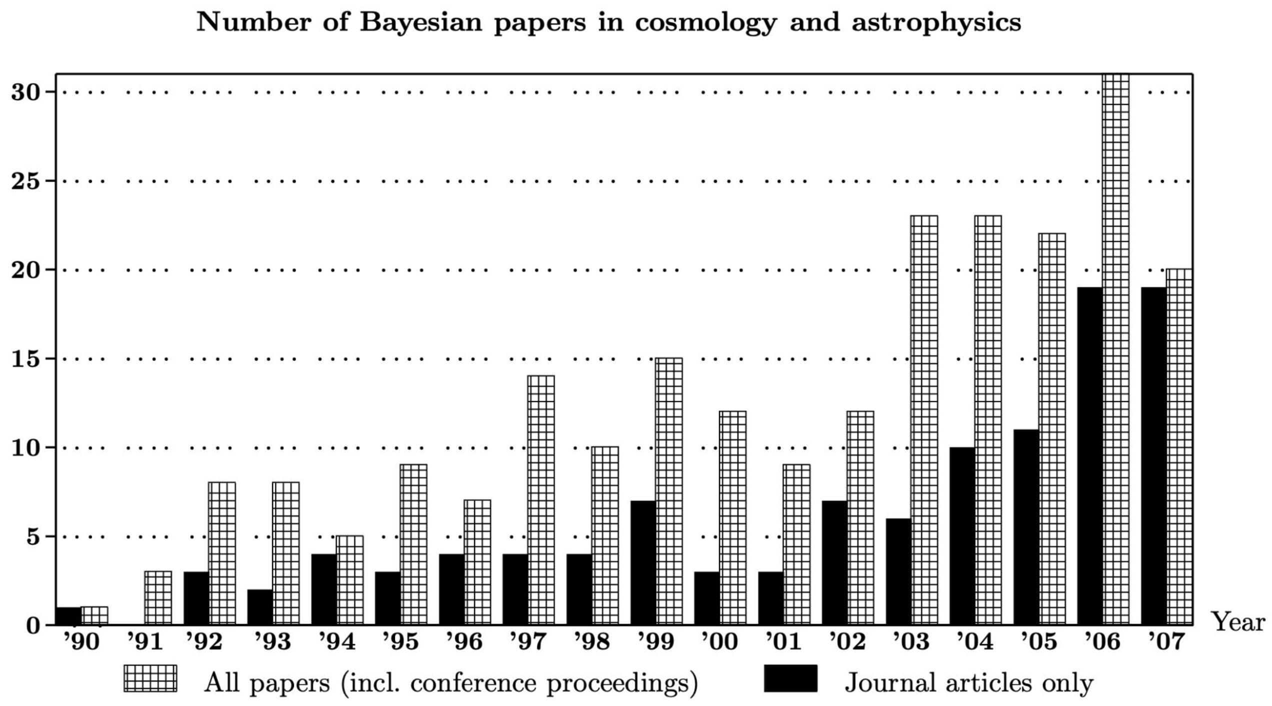
- Guaranteed logically consistent knowledge update supported by data.
- Just Construct the data model and explore the parameter space.

Bayes 1763
Cox 1946

The Promise of Bayesian inference for Cosmology (Trotta 2008)

Why did Bayes become prominent in cosmology?

- **Increased Data Sizes and Complexity**
  - new surveys arrived
- **Algorithmic Advances**
  - MH, HMC
- **Increased Computational Power**
- **Handling of Complex Models**
  - Ease of modeling (e.g. BHM)
- **Nuisance Parameter marginalization**

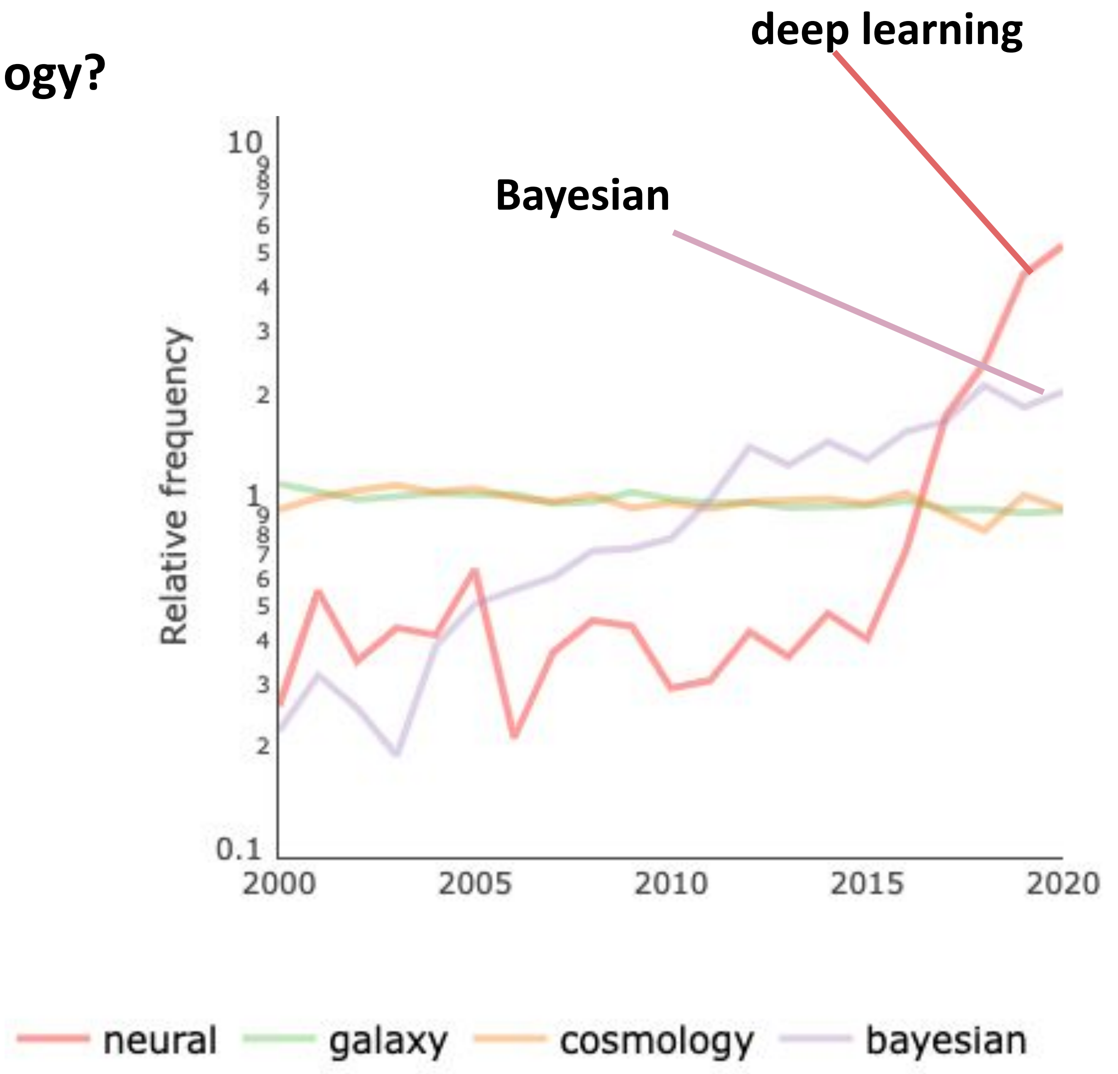


Loredo 1992
Trotta 2008
Trotta 2017

The Promise of Deep Learning for Cosmology (Huertas-Company & Lanusse 2023)

Why did Deep Learning become prominent in cosmology?

- **Increased Data Sizes and Complexity**
  - new are surveys arriving
- **Algorithmic Advances**
  - Backpropagation, CNN
- **Increased Computational Power**
- **Handling of Complex Models**
  - e.g. supervised learning
- **Nuisance Parameter marginalization**

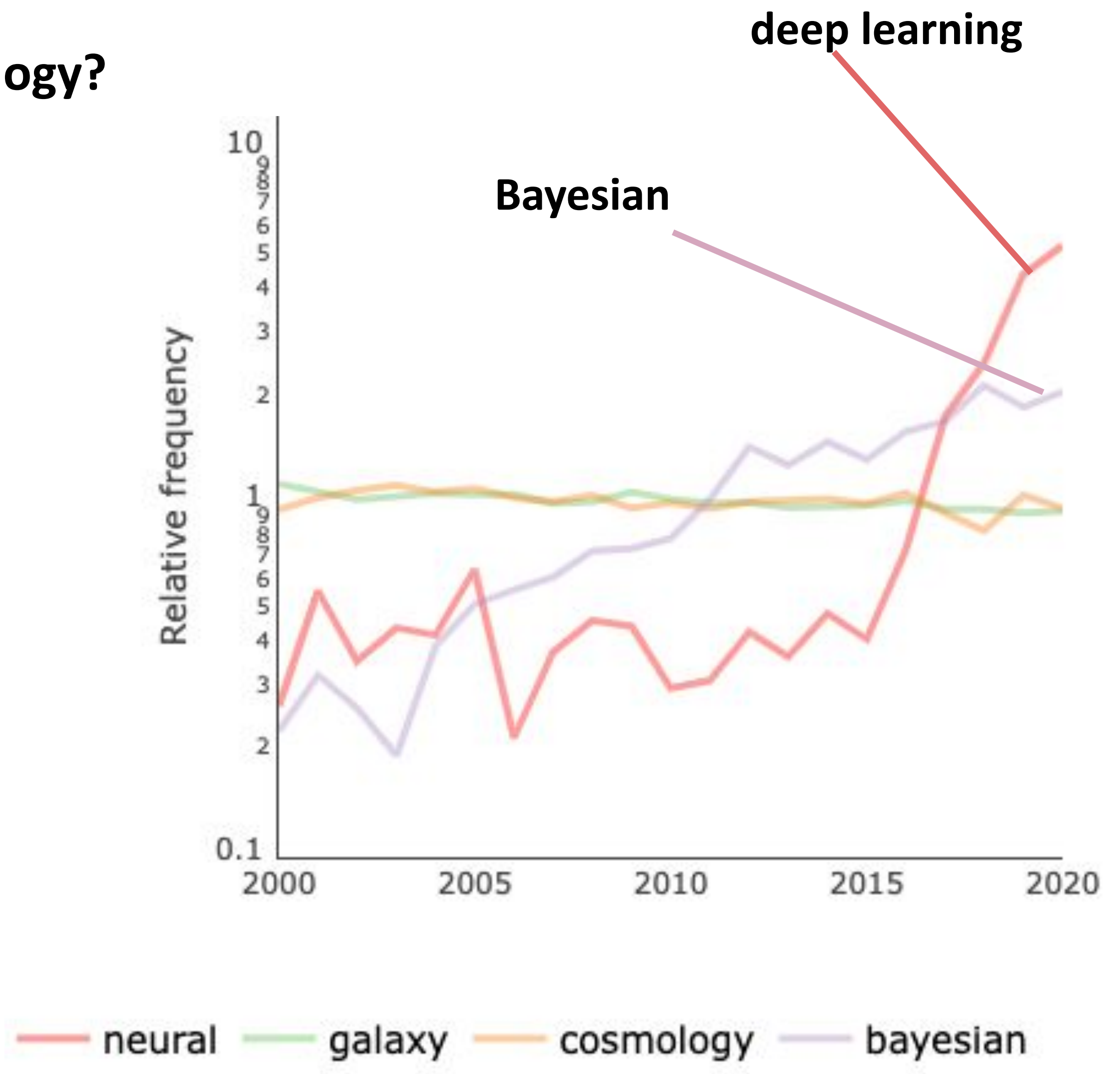


Huertas-Company & Lanusse 2023

The Promise of Deep Learning for Cosmology (Huertas-Company & Lanusse 2023)

Why did Deep Learning become prominent in cosmology?

- **Increased Data Sizes and Complexity**
  - new are surveys arriving
- **Algorithmic Advances**
  - Backpropagation, CNN
- **Increased Computational Power**
- **Handling of Complex Models**
  - e.g. supervised learning
- **Nuisance Parameter marginalization**
- **Enhancing human data modeling**
  - No more manual feature engineering
  - Fully automatic feature learning
  - Data-driven, not algorithm-centered



Huertas-Company & Lanusse 2023

What is gravity?

- Expansion Dynamics

What are the sources of gravity?

- Cosmic Content
- Dynamic Structure Growth

What are the initial conditions?

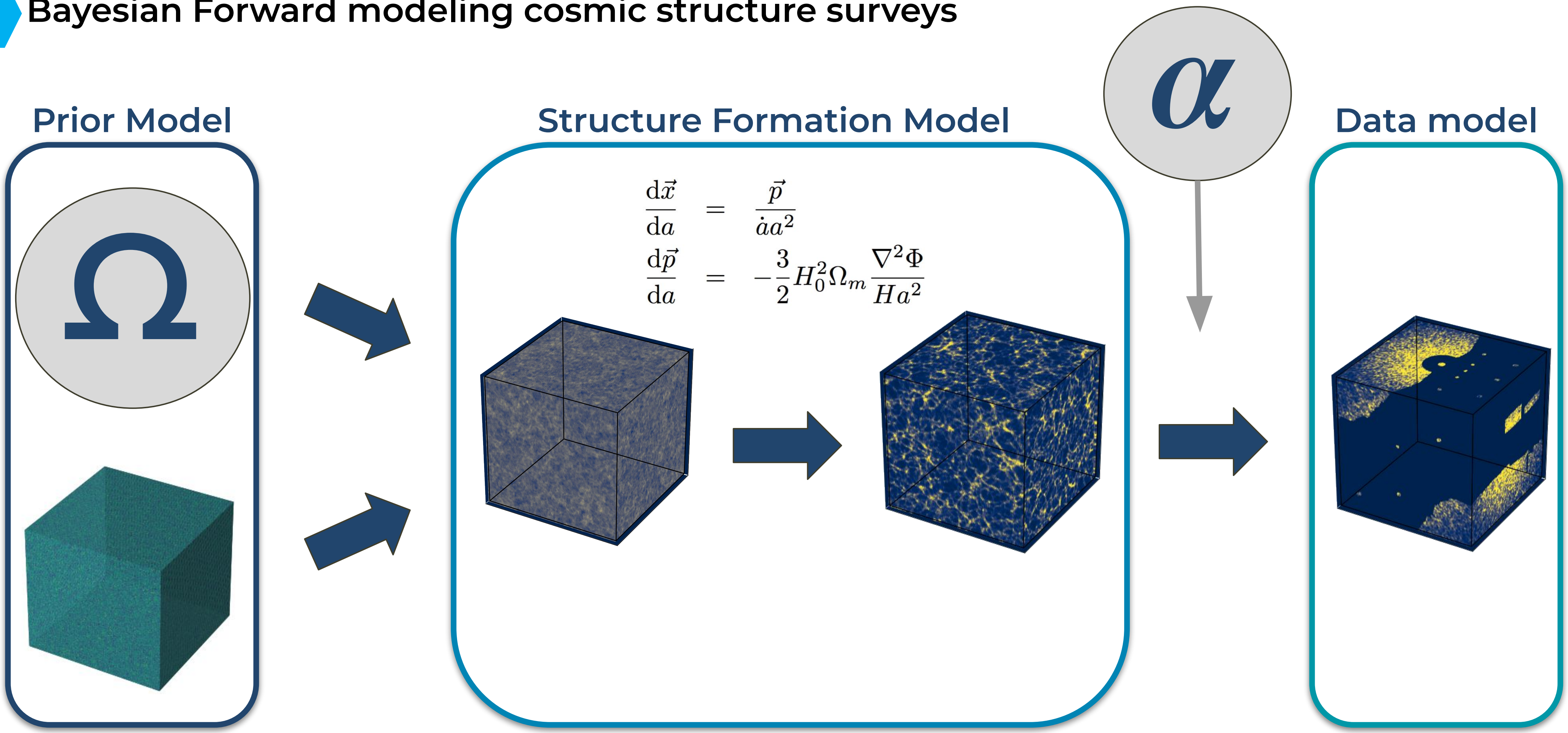
- Early universe physics
- Origin of structure

$$G_{\mu\nu} = \kappa T_{\mu\nu}$$

$$\nabla_\nu T^{\mu\nu} = 0$$

A set of "second order differential equations".
Weinberg (2009)

Bayesian Forward modeling cosmic structure surveys

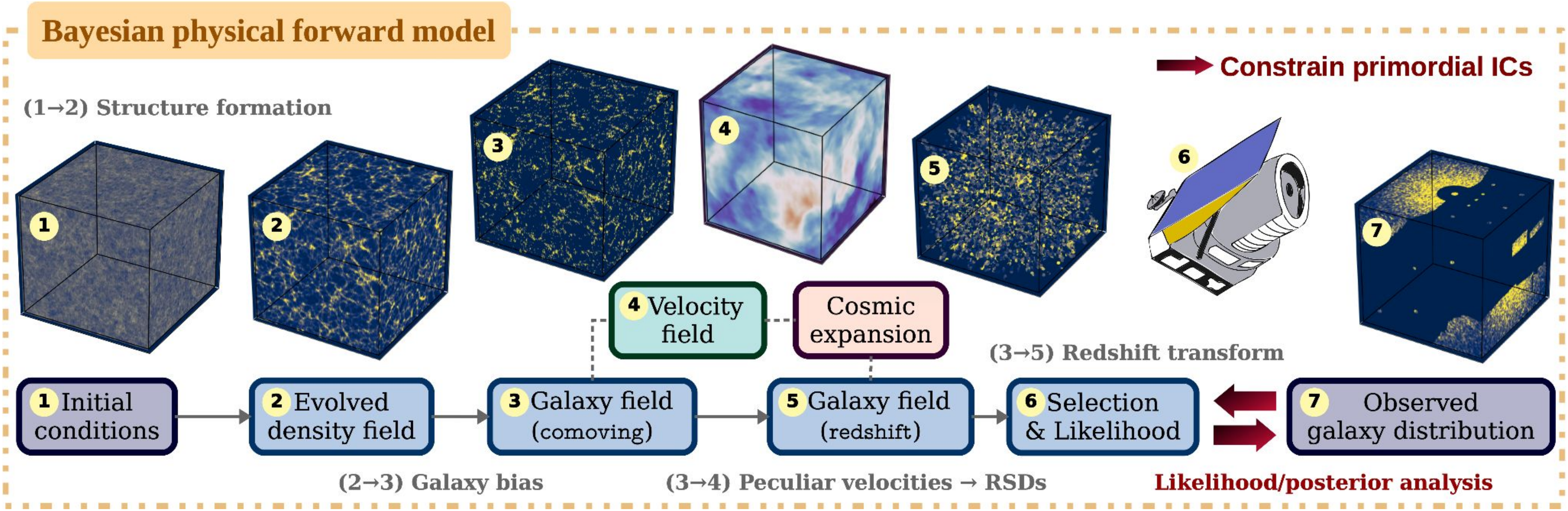


$$\pi\left(\mathbf{x}, \boldsymbol{\Omega}\right)$$

$$\pi\left(\rho_{\mathbf{m}} | \mathbf{x}, \boldsymbol{\Omega}\right)$$

$$\pi\left(\mathbf{N_g} | \rho_{\mathbf{m}}, \alpha, \boldsymbol{\Omega}\right)$$

Bayesian Forward modeling cosmic structure surveys

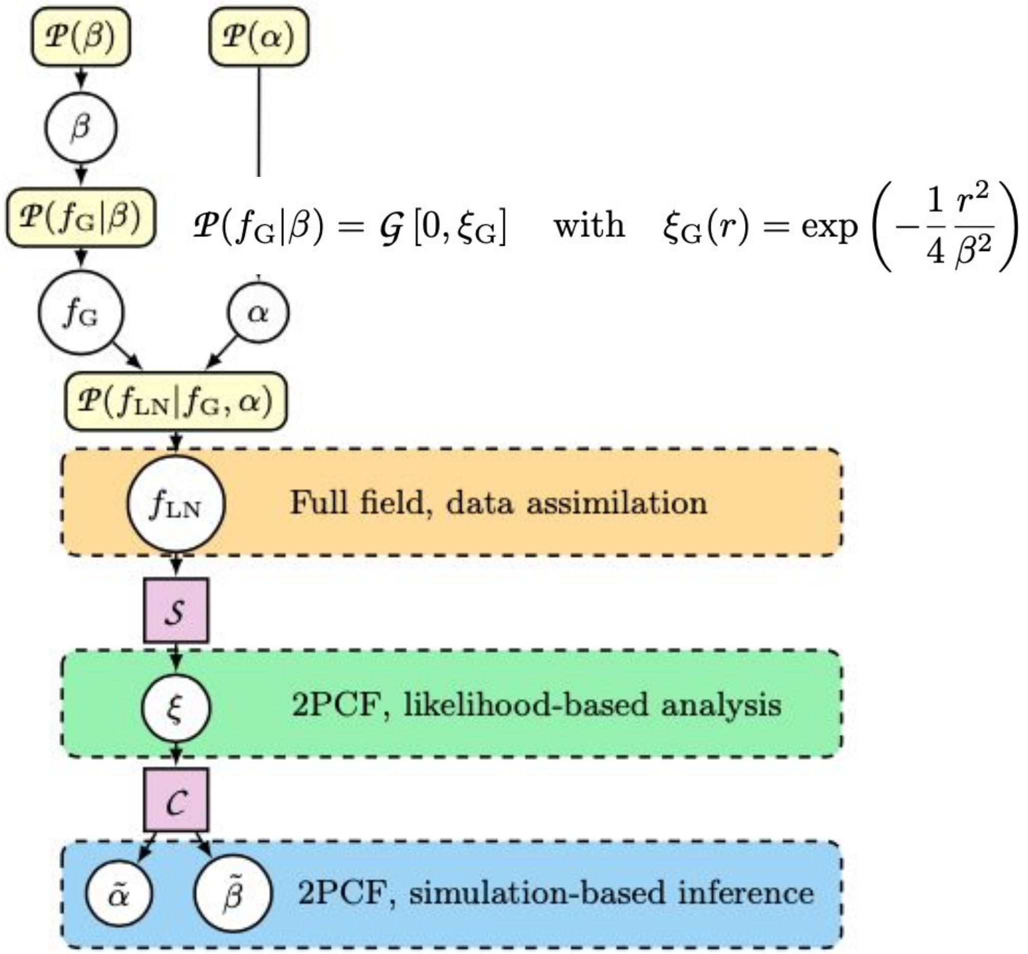


Kostic et al. 2022

How to optimally extract information from data?

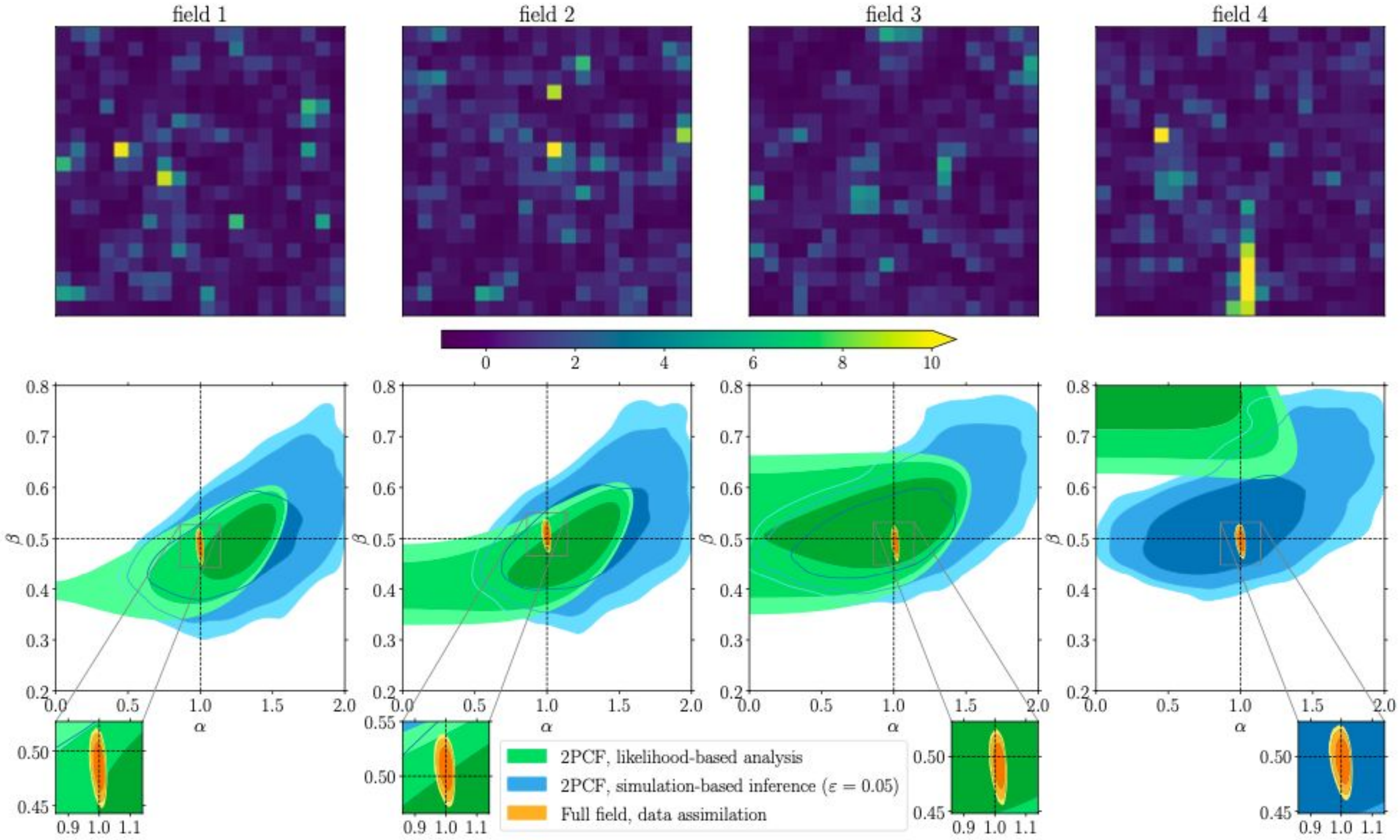
Compare 3 standard approaches:

- **Standard likelihood-based analysis (LBA)** of the two-point correlation function (2PCF), assuming a Gaussian distribution and fixed covariance matrix
- **Simulation-based inference (SBI)**, aka likelihood-free inference, ABC, based on the 2PCF
- **Field level data assimilation (DA)** technique, e.g. Bayesian forward modeling



Leclercq & Heavens 2021

Beyond summary statistics: Field-level inference



Field-level inference leverages additional data constraints, particularly higher-order statistics.

Leclercq & Heavens 2021

02

Field-Level Inference in Cosmic Structure Analysis

A field-level approach to CMB analyses:

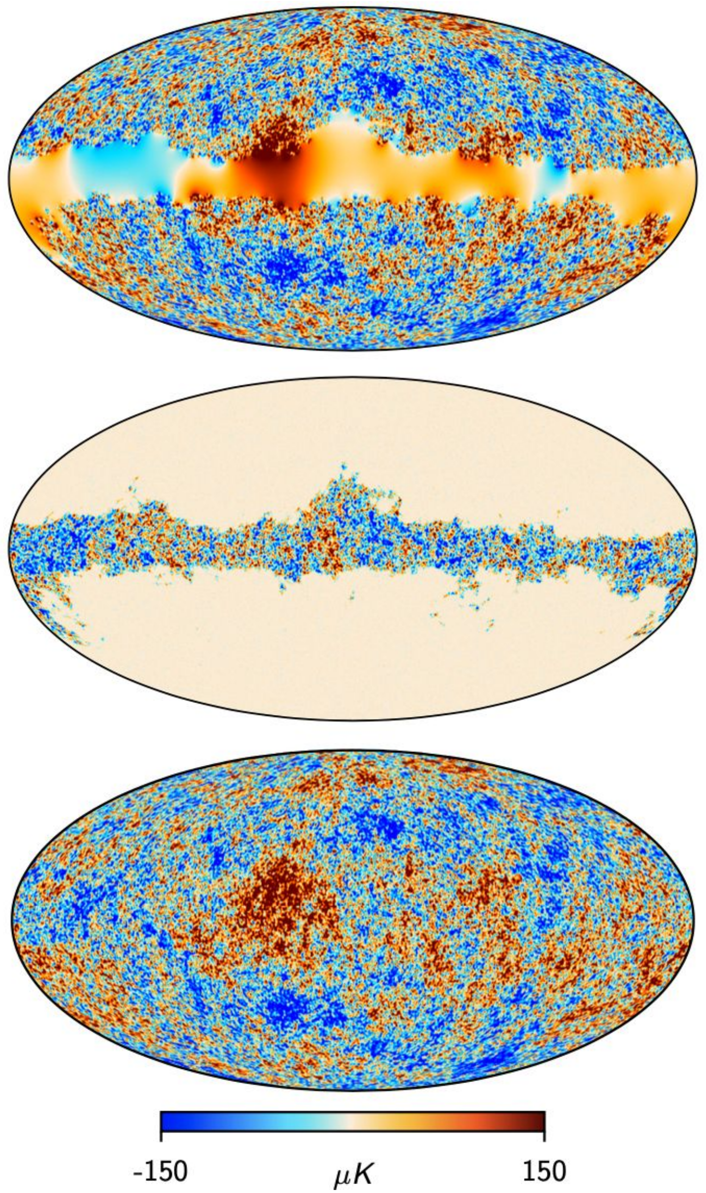
- Wiener Posterior

  (see e.g. Wandelt et al 2004, Jewell al 2004, Eriksen et al 2004, Elsner & Wandelt et al 2012, Thommesen et al 2020)

  - Linear data modeling
  - Gaussian Statistics:

$$P(s|d) \propto \exp\left(-\frac{1}{2}\left(s^{\dagger}S^{-1}s + (d - Rs)^{\dagger}N^{-1}(d - Rs)\right)\right)$$

- Wiener Filtering & Gibbs sampling
  - Handle noise & incomplete data
  - Uncertainty Quantification



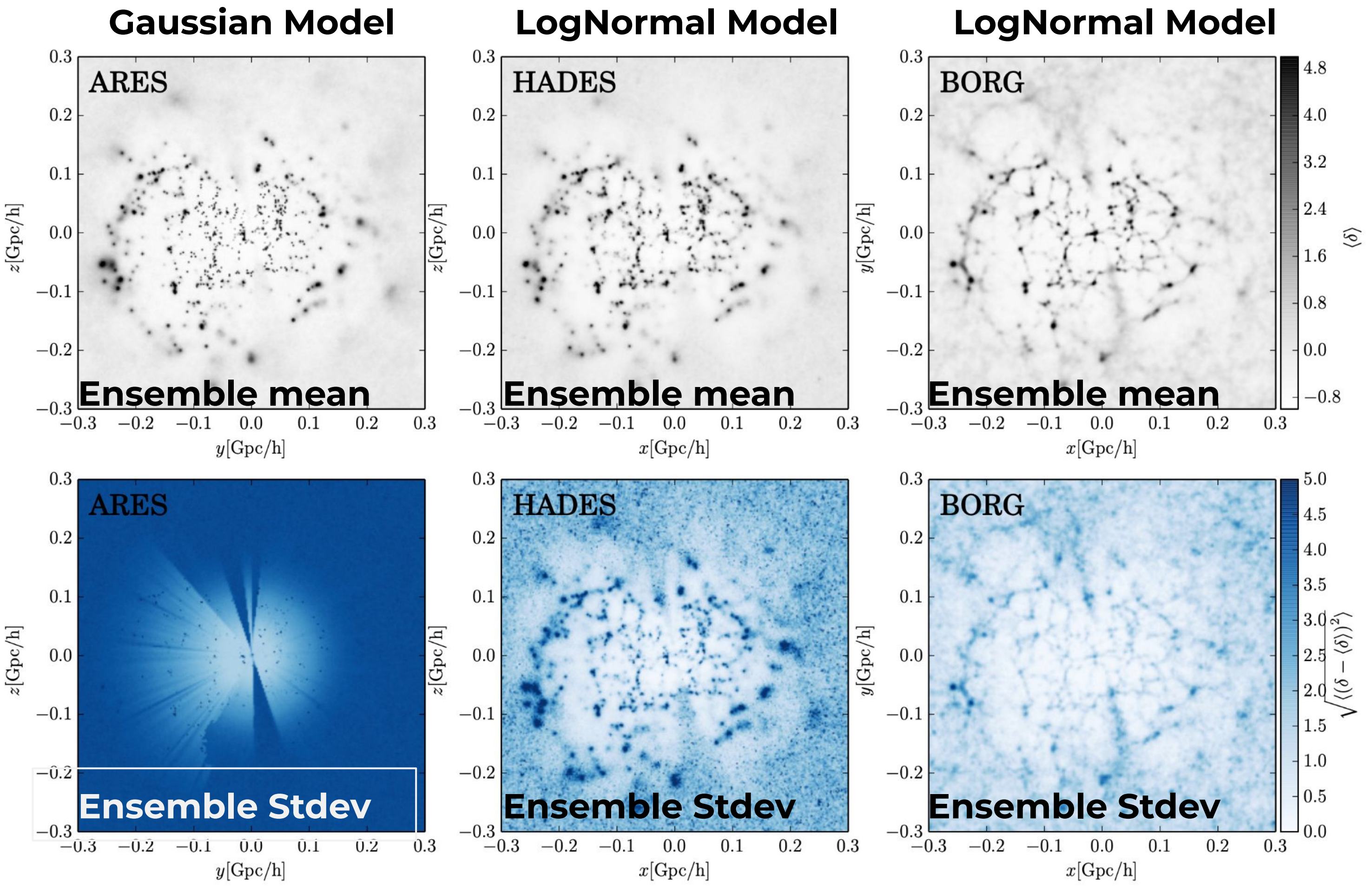
Thommesen et al 2020

(Also see ML Wiener filter implementations in e.g. Münchmeyer & Smith et al 2019, Costanza et al 2023, Pal & Saha 2023)

Modeling cosmic structure data

Searching Cosmic Structure Posteriors:

- Gaussian / Wiener posterior
  - (see e.g. Zaroubi et al 2002, Erdoğdu et al 2004, Kitaura et al 2009, Jasche et al 2010, Grannet et al 2012)
- LogNormal
  - (see e.g. Kitaura et al 2010, Jasche et al 2010, Ata et al 2015, Böhm et al 2017, Böhm et al 2017)
- Physics based model
  - (see e.g. Jasche et al 2013, Wang et al 2013, Kitaura 2013, Ata et al 2022, Modi et al 2023)



Bayesian Origin Reconstruction from Galaxies (BORG)

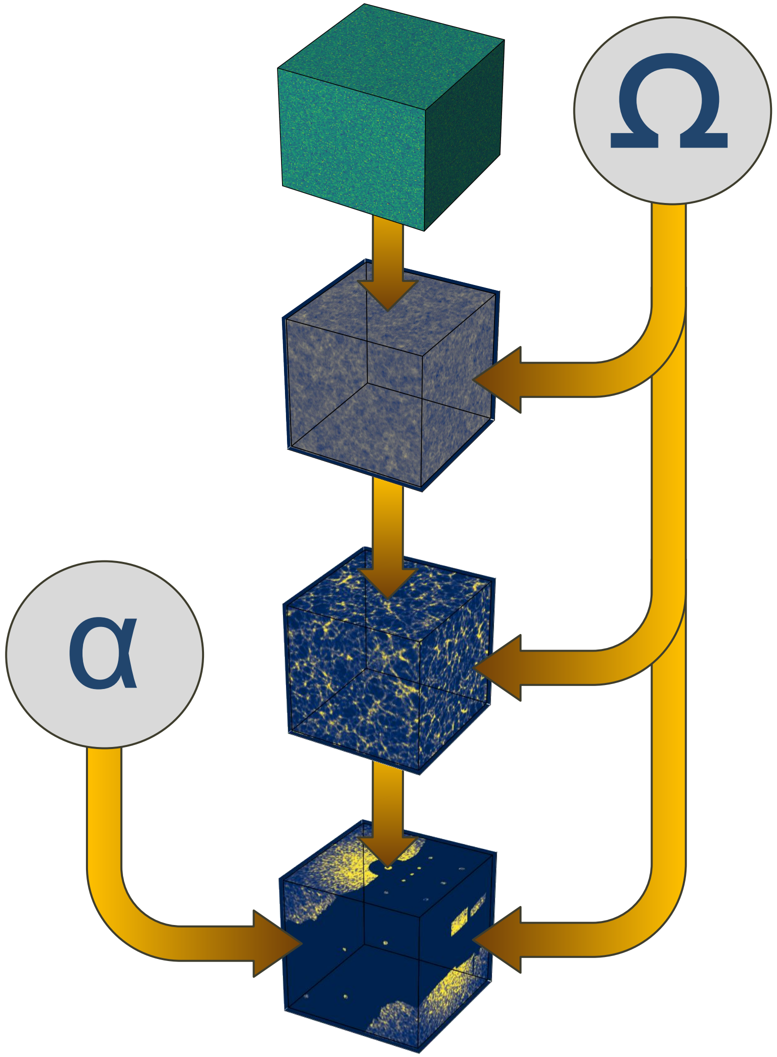
Large scale Structure Posterior via BHM

- "Ab initio" modeling of cosmic structures
  - Gaussian initial conditions to Non-Gaussian final conditions
- Numerical Structure Formation Models
  - e.g. LPT, PM and COLA
- Phenomenological likelihoods and galaxy biases
  - e.g. Poisson distribution, power-law bias

Large scale Bayesian inference via MCMC

- A High-dimensional problem: 10^6 to 10^9 parameter
- Hierarchical Bayes and block sampling
- Efficient Hamiltonian Monte Carlo technique
- Differentiable structure formation model

A BHM of the cosmic structure



Jasche & Wandelt 2014
Jasche & Lavaux 2019

(Also see Wang et al 2013, Kitaura 2013, Ata et al 2022, Modi et al 2023)

Bayesian Origin Reconstruction from Galaxies (BORG)

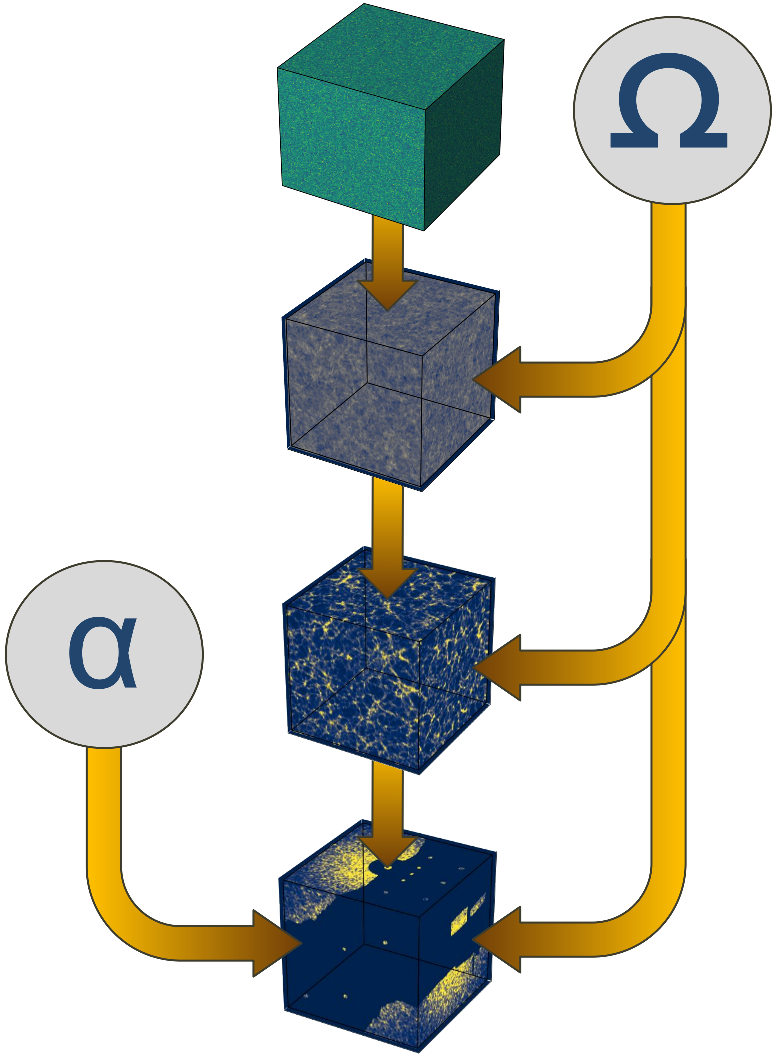
Large scale Structure Posterior via BHM

- "Ab initio" modeling of cosmic structures
  - Gaussian initial conditions to Non-Gaussian final conditions
- Numerical Structure Formation Models
  - e.g. LPT, PM and COLA
- Phenomenological likelihoods and galaxy biases
  - e.g. Poisson distribution, power-law bias

Large scale Bayesian inference via MCMC

- **A High-dimensional problem: 10^6 to 10^9 parameter**
- Hierarchical Bayes and block sampling
- **Efficient Hamiltonian Monte Carlo technique**
- **Differentiable structure formation model**

A BHM of the cosmic structure



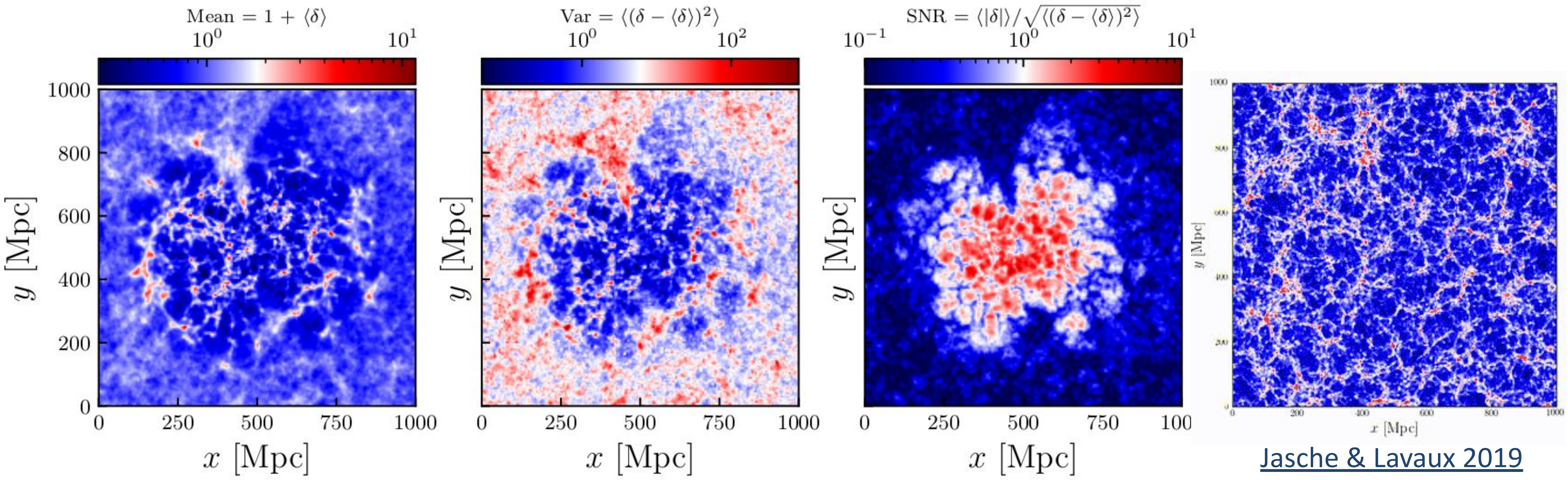
Jasche & Wandelt 2014
Jasche & Lavaux 2019

(Also see Wang et al 2013, Kitaura 2013, Ata et al 2022, Modi et al 2023)

A numerical cosmic structure posterior distribution of the local super volume

- Using the 2M++ Galaxy Compilation (Lavaux & Hudson 2011).

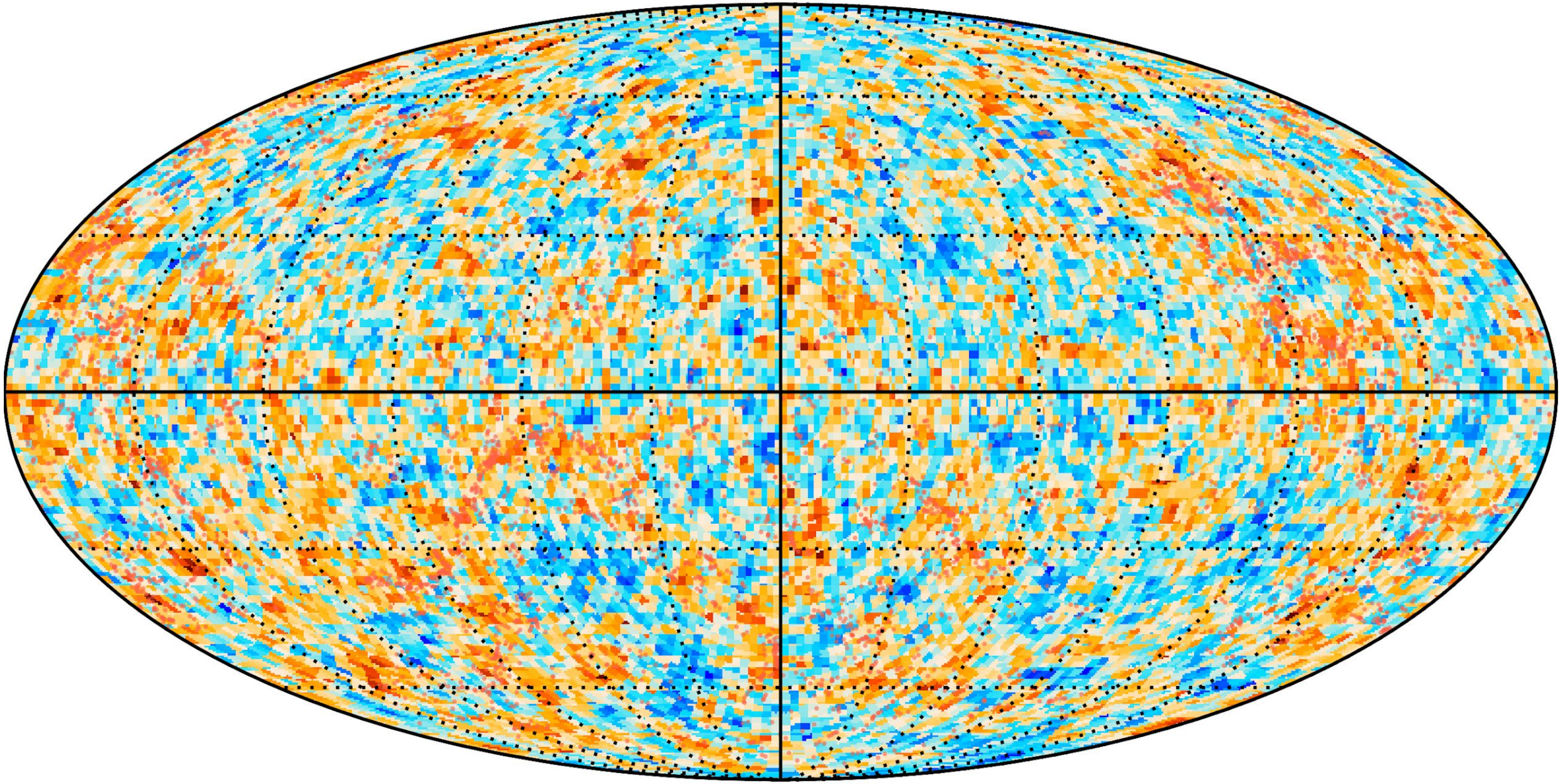
Ensemble mean | **Ensemble standard deviation** | **Signal To Noise** | **Generated Samples**



Jasche & Lavaux 2019
McAlpine (in prep)

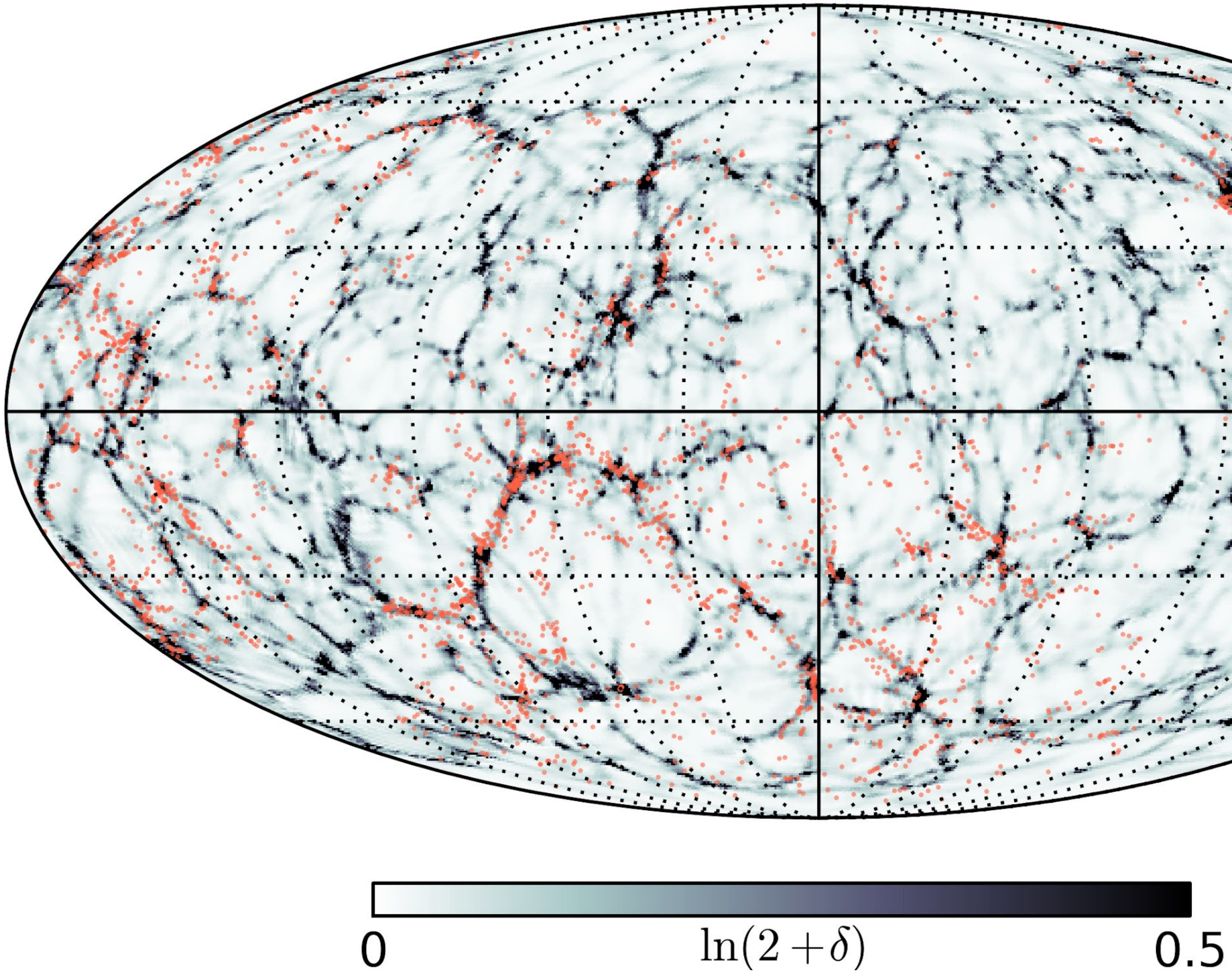
credit: Stuart McAlpine

BORG: A Bayesian Generative Model of the Cosmic Structure



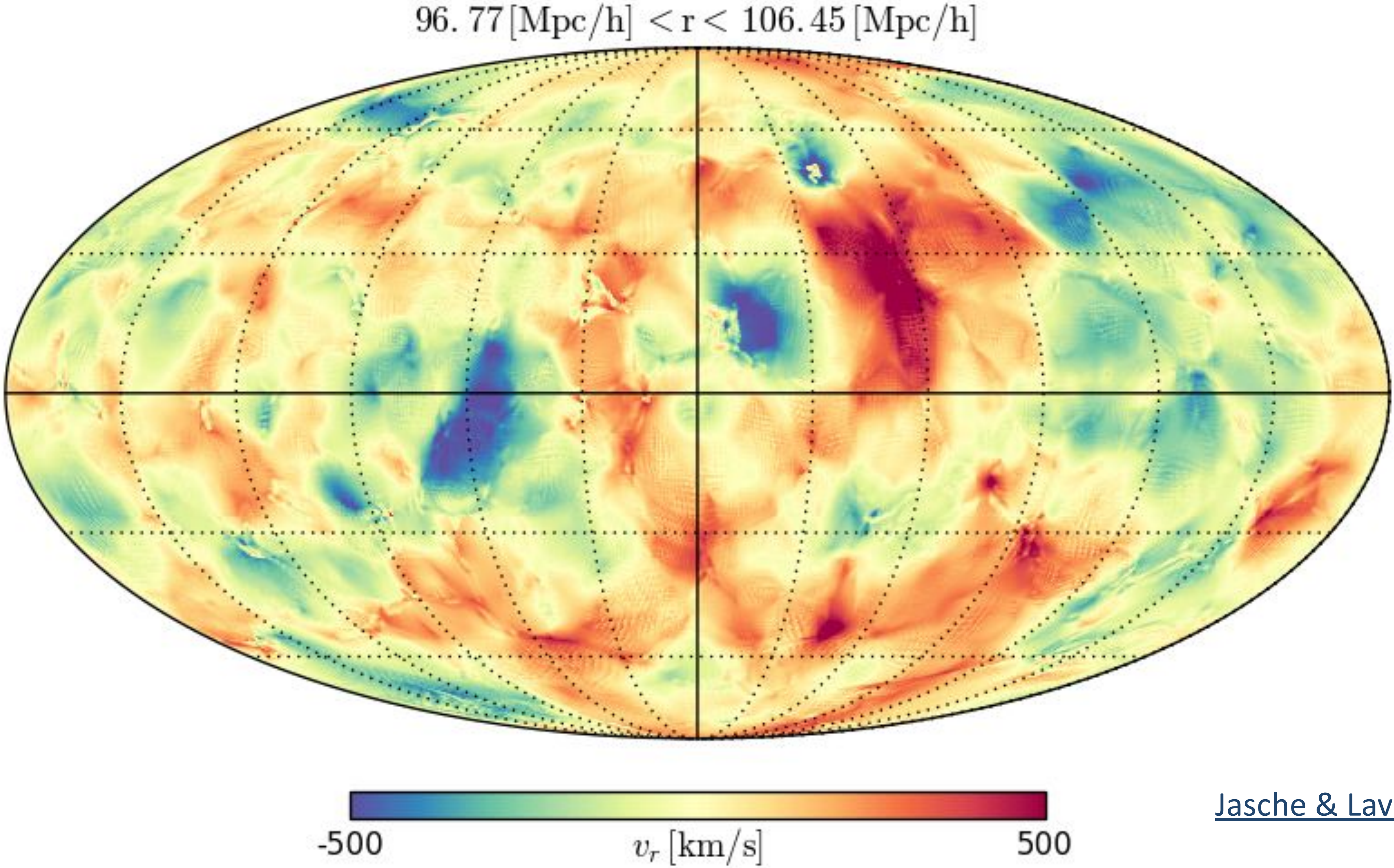
Jasche & Lavaux 2019

BORG: A Bayesian Generative Model of the Cosmic Structure



Jasche & Lavaux 2019

BORG: A Bayesian Generative Model of the Cosmic Structure

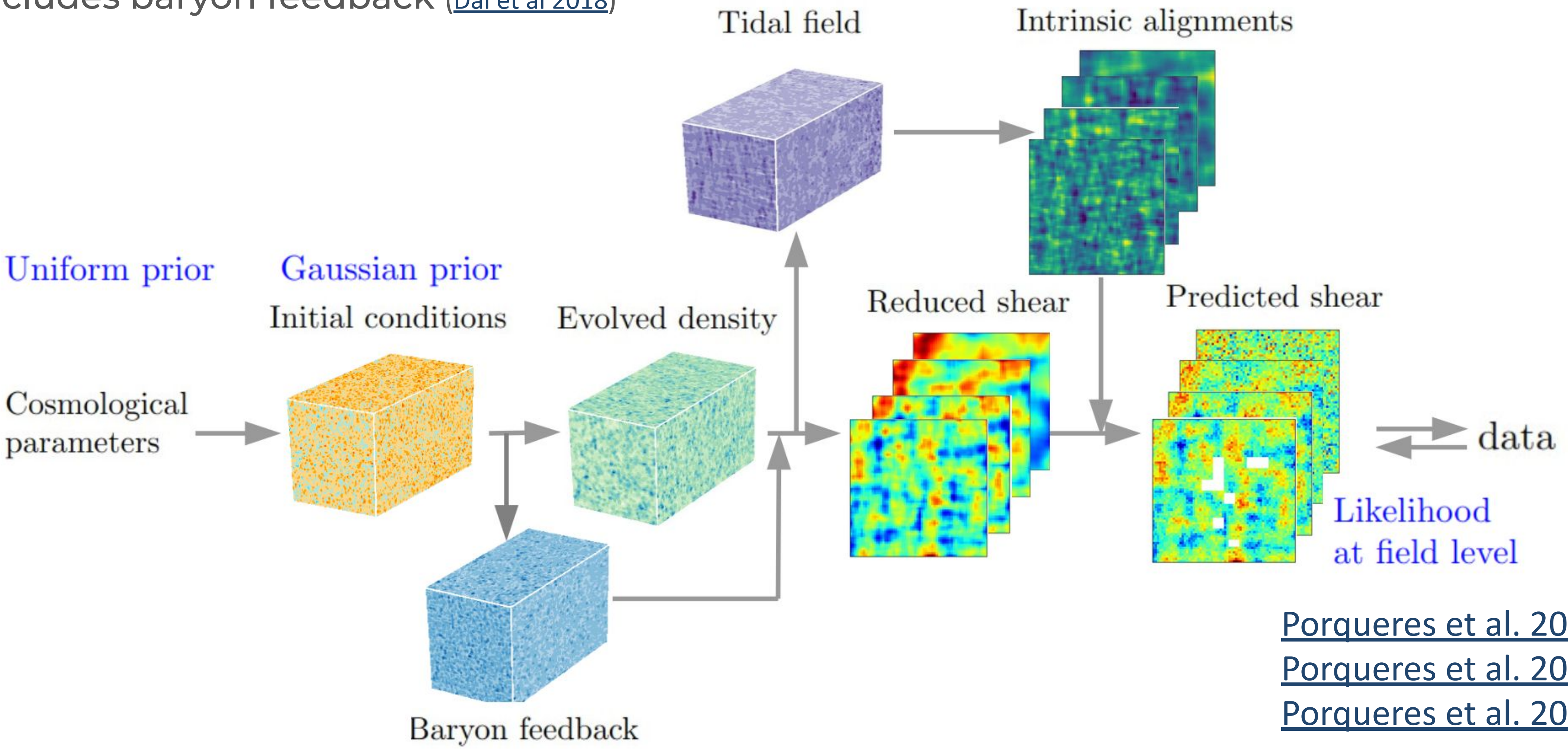


Jasche & Lavaux 2019

Field-level inference of cosmic shear with intrinsic alignments

BORG-WL

- Field-based Bayesian Hierarchical Model for cosmic shear
- Includes baryon feedback (Dai et al 2018)



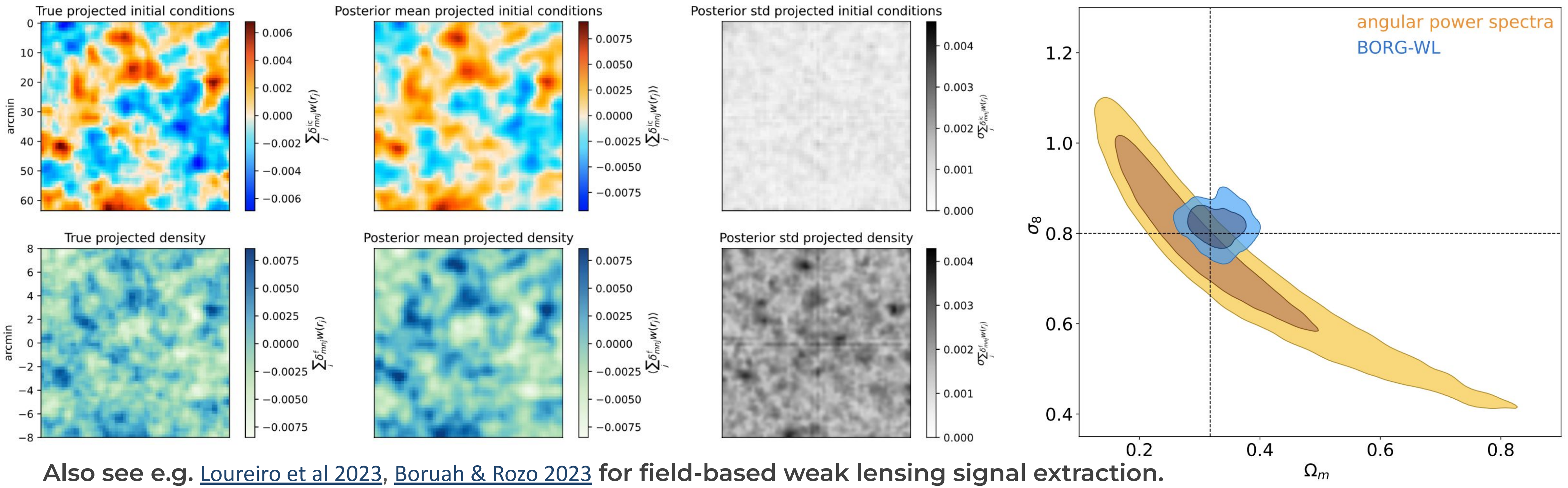
Porqueres et al. 2021
Porqueres et al. 2022
Porqueres et al. 2023

Field-level inference of cosmic shear with intrinsic alignments

Joint inference of fields and cosmological parameter

- Lifting parameter degeneracies by jointly using:
  - matter power spectrum
  - geometry
  - tidal field
  - structure growth
  - distance-redshift relation

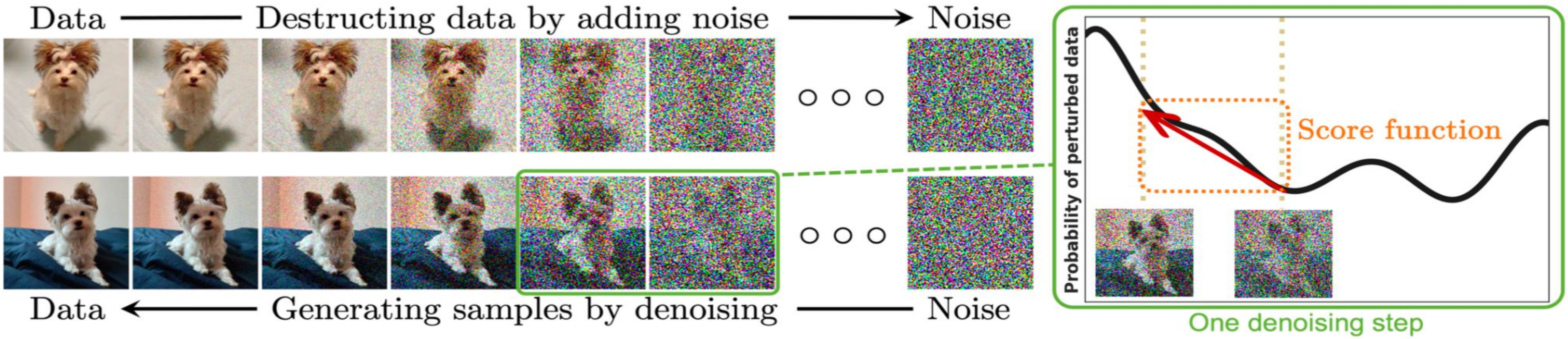
Porqueres et al. 2021
Porqueres et al. 2022
Porqueres et al. 2023



Also see e.g. Loureiro et al 2023, Boruah & Rozo 2023 for field-based weak lensing signal extraction.

Inferring Initial conditions without explicit posteriors

Posterior sampling using diffusion models

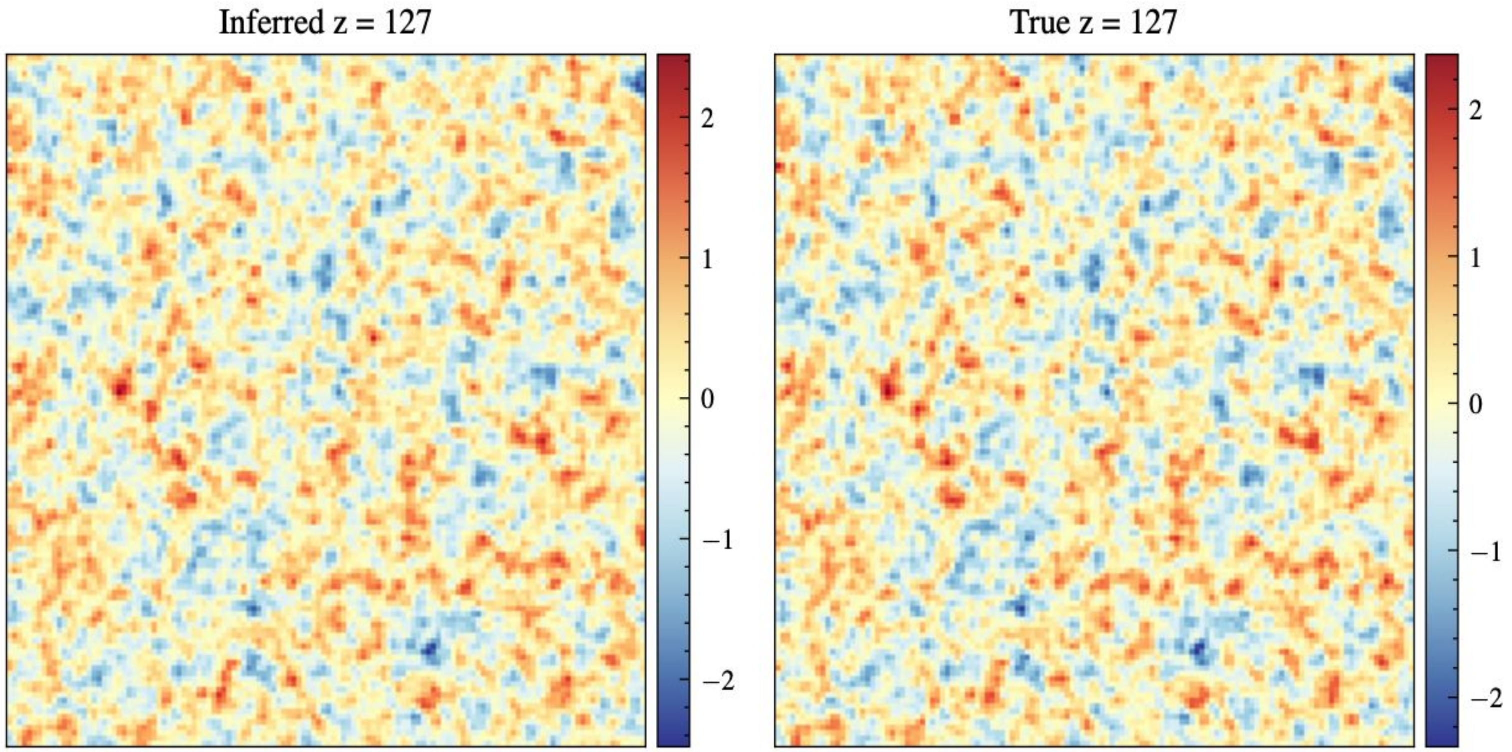


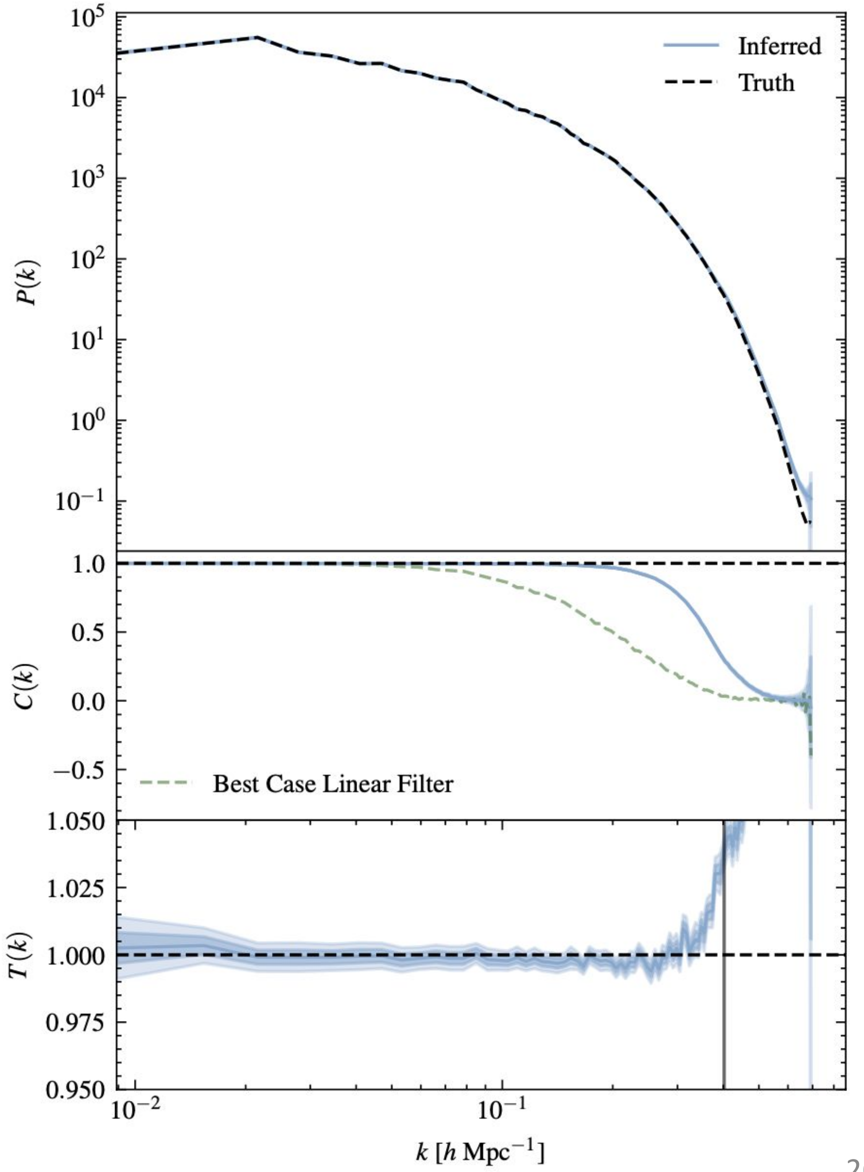
Yang et al 2022

Inferring Initial conditions without explicit posteriors

Posterior sampling using conditional score-based generative models

- No explicit likelihood and gradient needed
- Marginalization over cosmological parameter
- Computationally effective





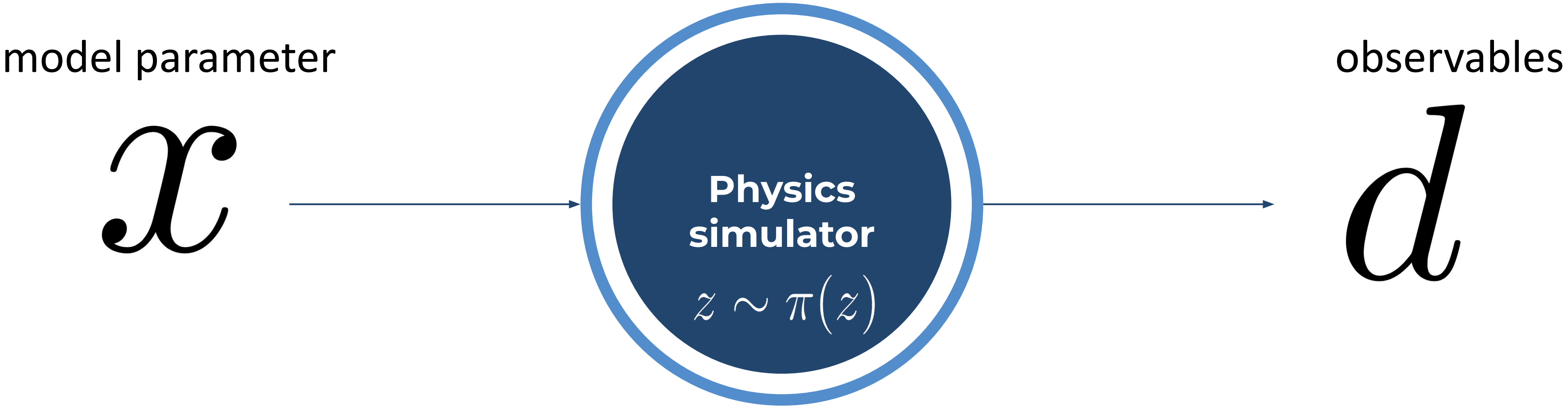
Also see e.g. List et al 2023 for an alternative approach.

Legin et al 2023

03 Implicit Likelihood Inference

Simulation Based Inference: Implicit Likelihoods and Posterior Distributions

Simulators define the likelihood $\pi(d|x)$ as implicit distributions:



$x \sim \pi(x)$

$d \sim \pi(d|x)$

This process generates samples from the joint distribution: $\mathcal{D} = \{x_i, d_i\} \sim \pi(x, d)$

e.g.:
Hahn et al 2022
Alsing et al 2020
Cranmer et al 2020

Constructing posterior distributions $\pi(x|d)$ via variational inference:

- Model the posterior distribution by a parametric conditional distribution $q_\phi(x|d)$
- Match $q \rightarrow \pi(x|d)$ by optimizing the Kullback–Leibler divergence:

$$\max_\phi \sum_i \log\left(q_\phi(x_i|d_i)\right) \qquad \forall x_i, d_i \sim \pi(x, d)$$

- In the simplest case:

$$q \leftarrow \mathcal{N}(\mu,\, \sigma^2) \text{ with } \mu = f_{\phi_1}(d) \text{ and } \sigma^2 = g_{\phi_2}(d)$$

but more generally Normalizing Flows or Mixture Density Networks (MDNs).

e.g.:
Makinen et al 2023
Lemos et al 2023
Hahn et al 2022
Alsing et al 2020
Cranmer et al 2020

Simulation Based Inference: Implicit Likelihoods and Posterior Distributions

Steps of SBI:

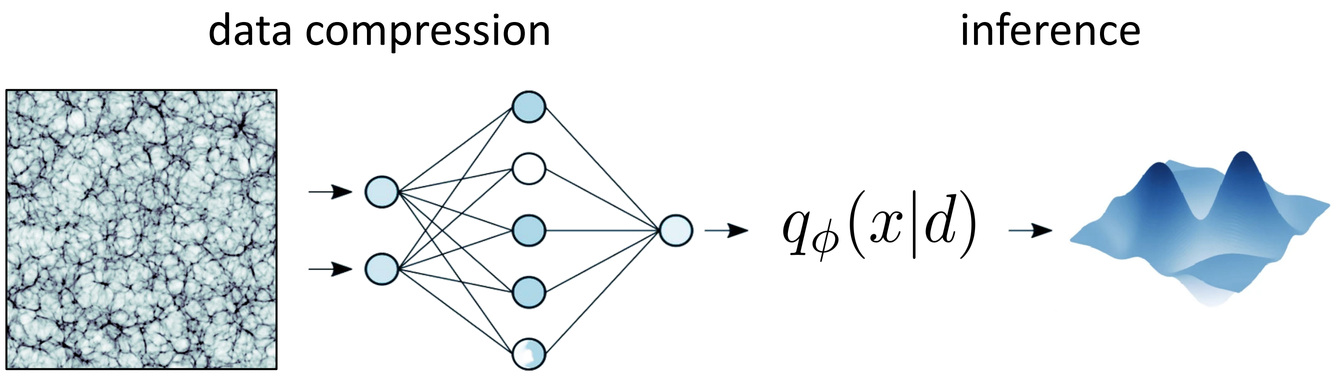
1. Create a training set $\mathcal{D} = \{x_i, d_i\} \sim \pi(x, d)$ (this may include data compression)

2. Choose a parametric model $q_\phi(x|d)$

3. Optimize parameter ϕ :

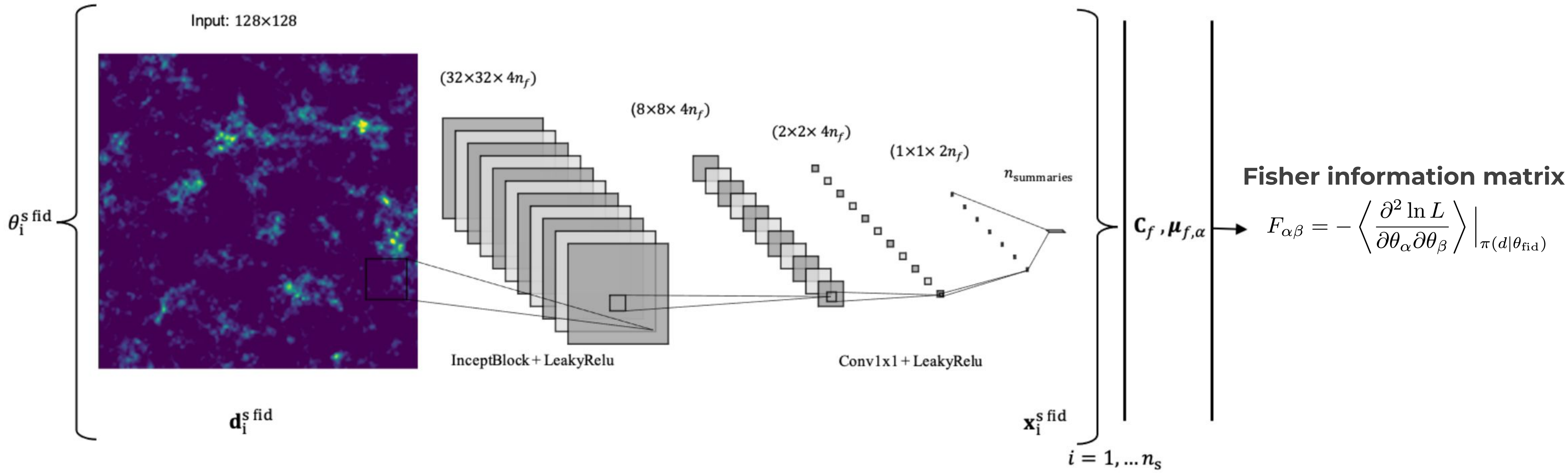
$$\max_\phi \sum_i \log\left(q_\phi(x_i|d_i)\right)$$

4. Perform inference:



e.g.:
Makinen et al 2023
Lemos et al 2023
Hahn et al 2022
Alsing et al 2020
Cranmer et al 2020

Constructing informative Summary Statistics



- **Information Maximizing Networks**
  - Compress data for dimensional reduction
  - Optimal Score compression by optimizing Fisher-Information:

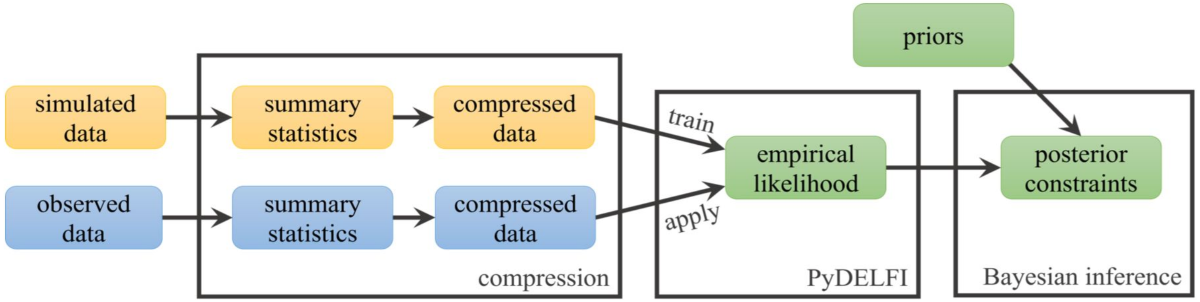
$$\mathcal{L} = -\ln\det\langle F \rangle_{\pi(d|\theta_{\text{fid}})}$$

e.g.:
Alsing et al 2018
Alsing et al 2019
Charnock et al 2018
Makinen et al 2021

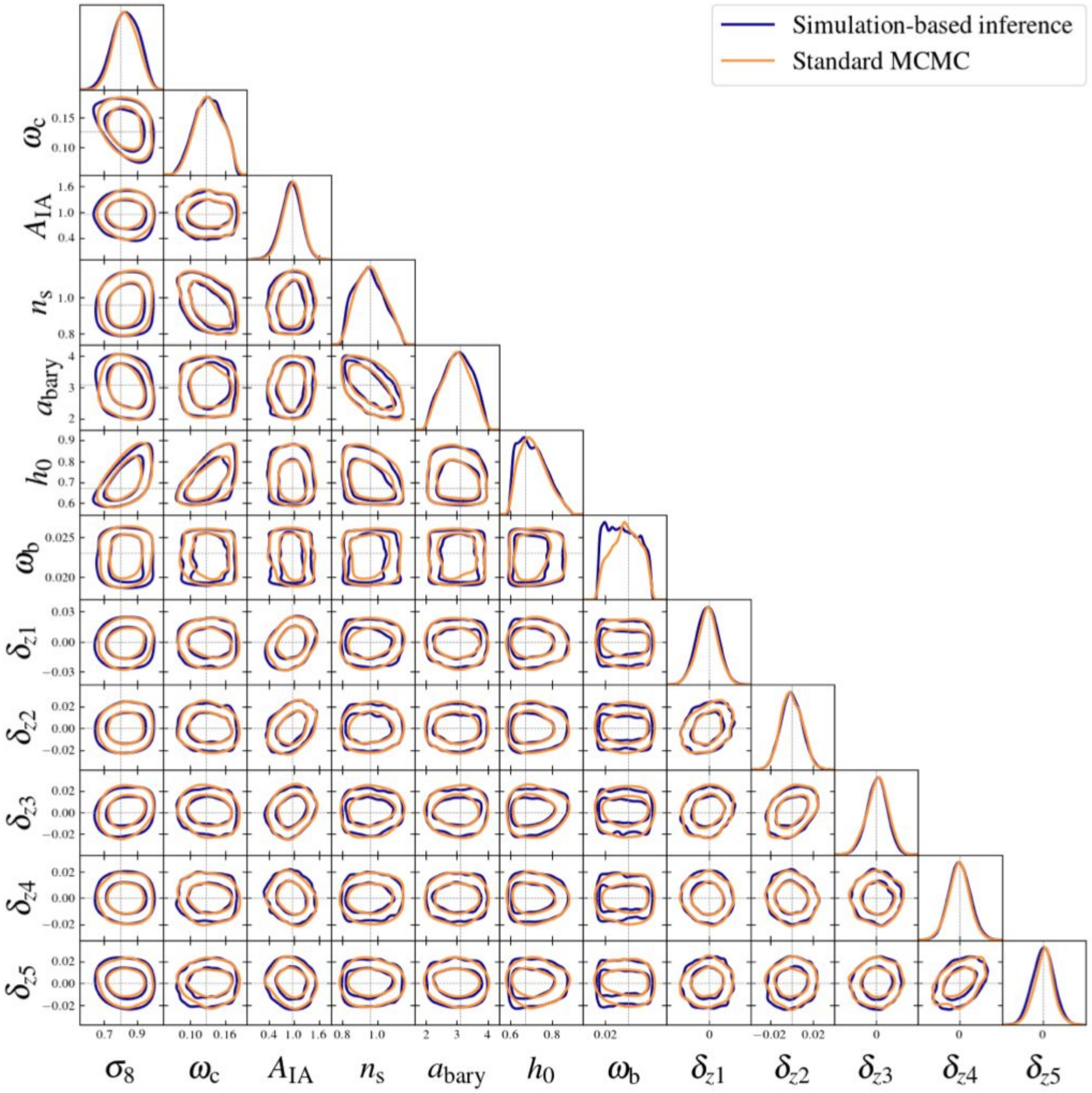
Also see e.g. Heavens et al 2000, Heavens et al 2023, Akhmetzhanova et al 2023

Simulation-based inference for cosmic shear with KiDS

Test SBI vs MCMC:



- Recovers the 12-dimensional KiDS posterior.
- Requires under 10,000 simulations.
- SBI is competitive to MCMC



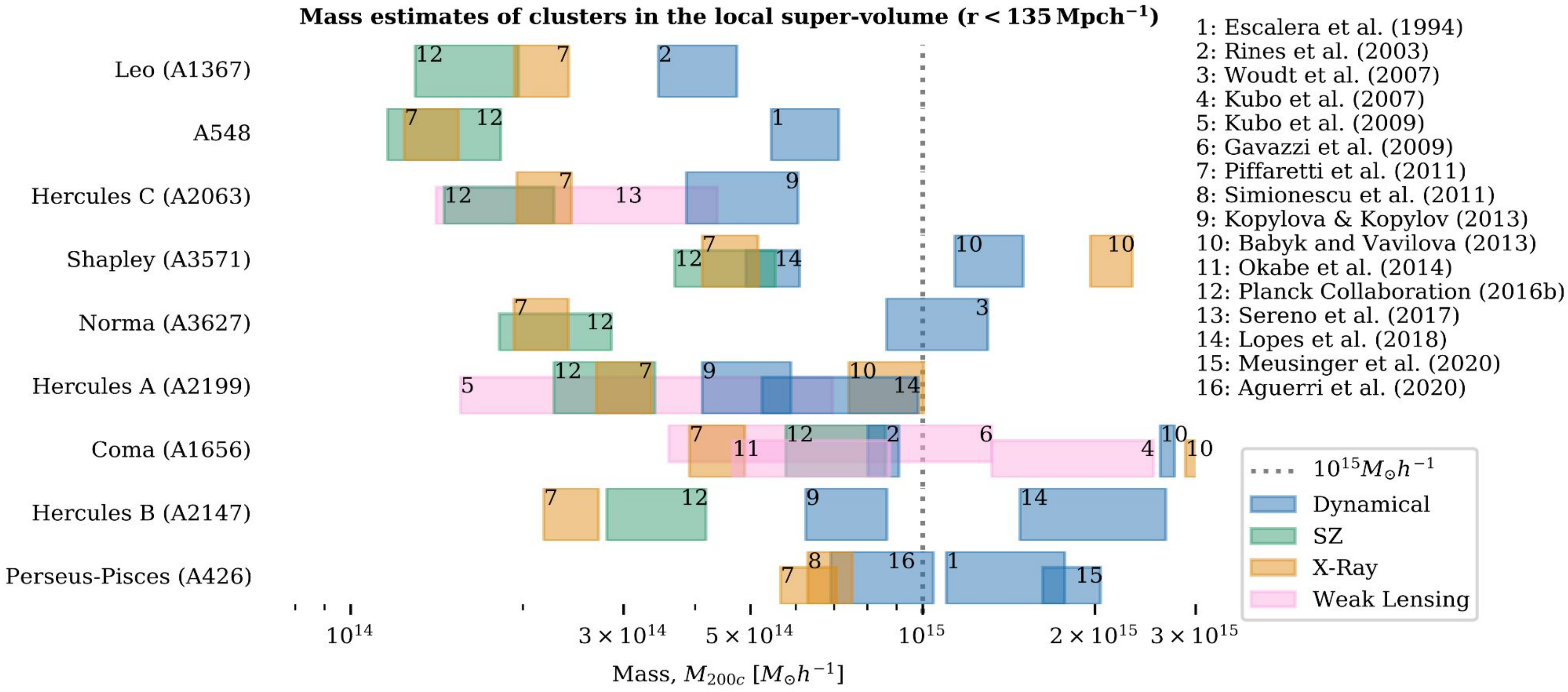
Lin et al 2023

Also see e.g. Zhao et al 2022, Lemos et al 2023, Makinen et al 2023, Modi et al 2023

04

Fast Emulation of Cosmological Simulations

Massive Galaxy Clusters in the Nearby Universe

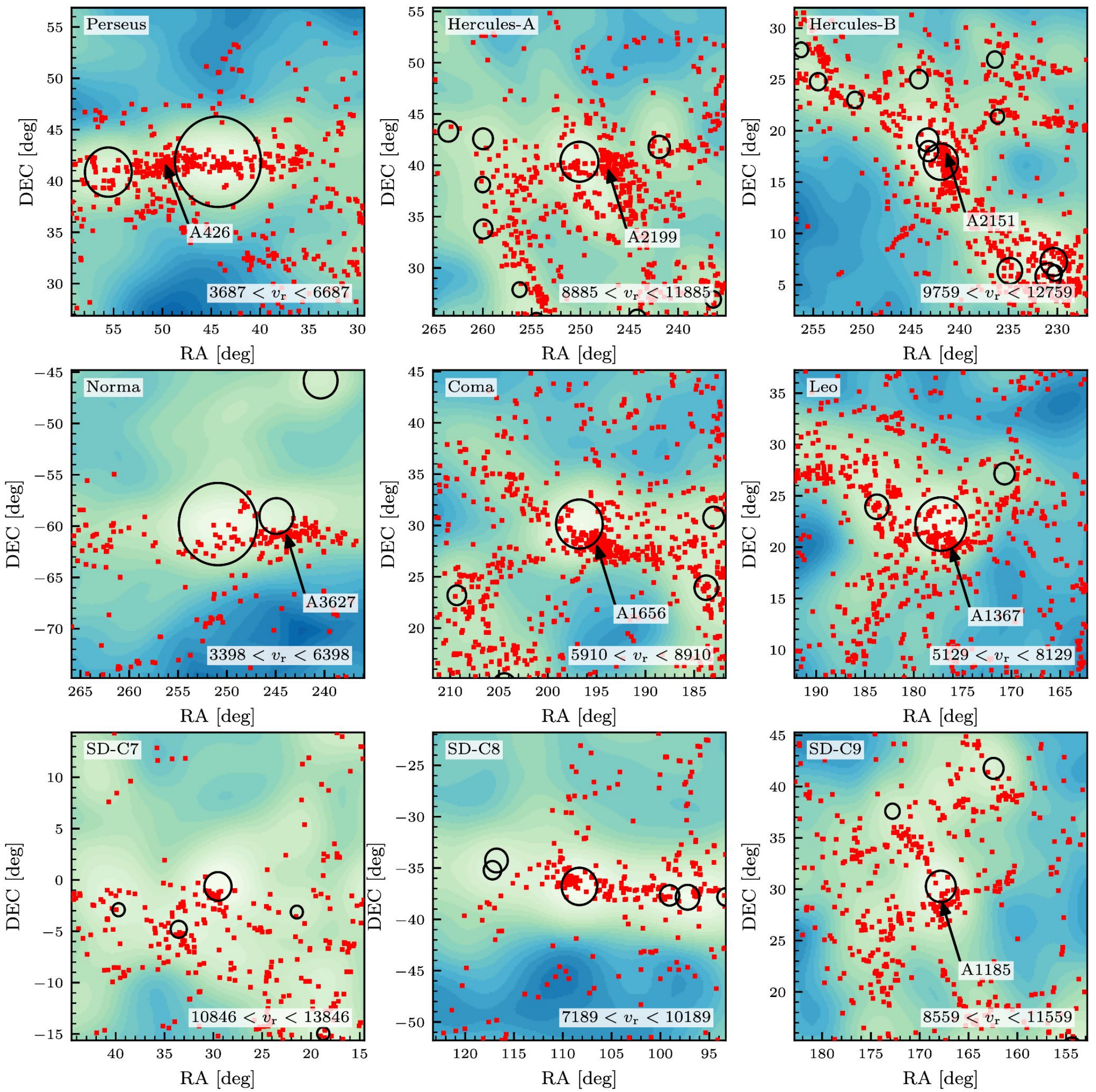


Stopyra et al 2021

Massive Galaxy Clusters in the Nearby Universe

- **Significance of High-Mass Objects:**
  - Small volume contribution
  - Large impact on cosmology
- **Challenges in Cluster Mass Estimation:**
  - Intricate details
  - complex cosmic environments
  - Simplifying assumptions do not apply (e.g., sphericity, virial theorem)
- **Requires accurate modeling:**
  - Non-linear structure formation
  - Fast and scalable models

Local Universe Cluster Mugshots

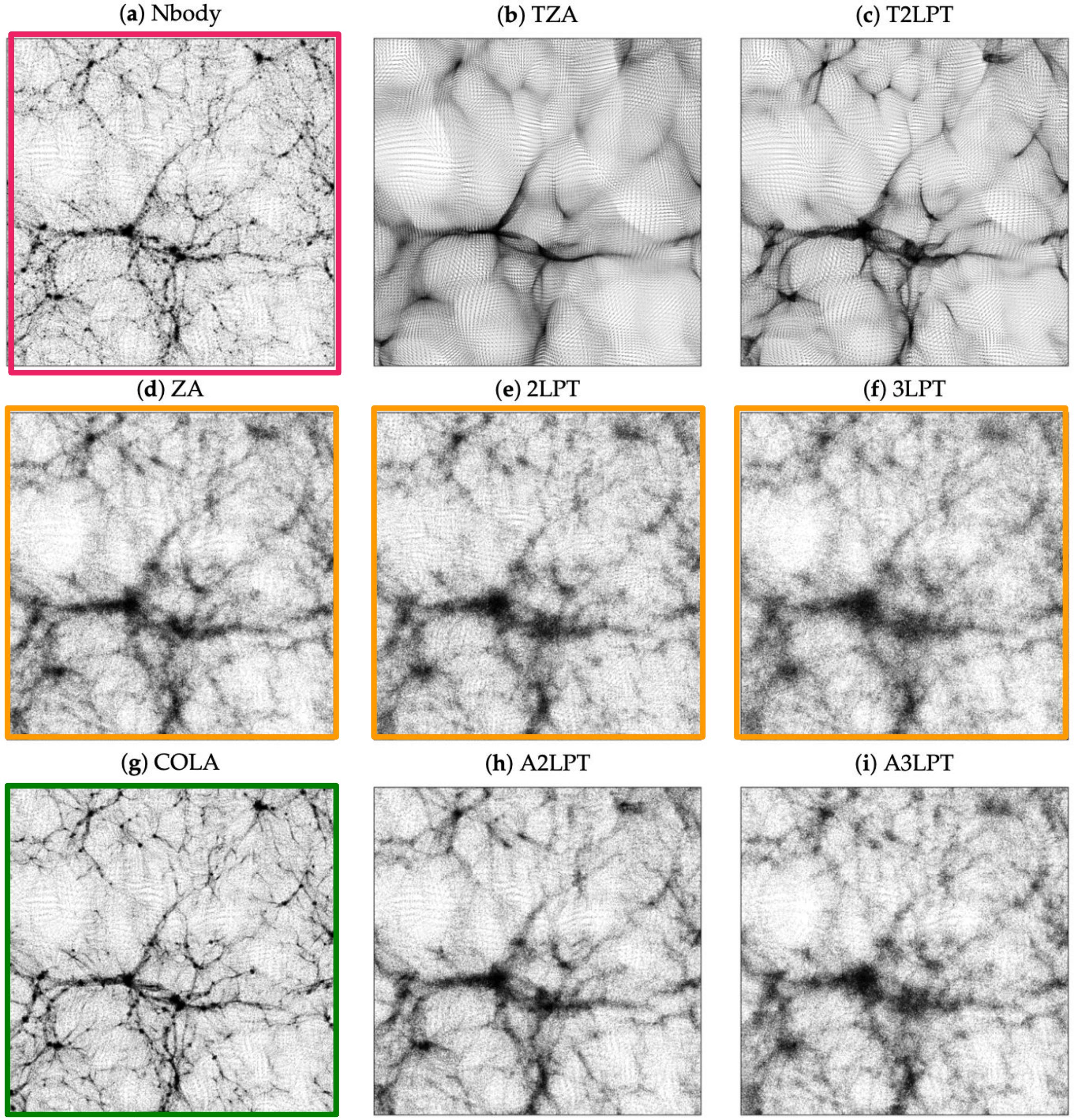


McAlpine et al 2022

Fast Field-Level Emulation of Cosmological Simulations

A multitude of approximate models:

- Lagrangian / Eulerian perturbation theory
  (see e.g. Zel'dovich 1970, Bernardeau 2002)
- Adhesion / Schrödinger mode
  (see e.g. Gurbatov et al 1985, Coles & Spencer et al 2002)
- **E**ffective **F**ield **T**heory of Cosmic Structures
  (see e.g Carrasco et al 2012, Schmidt et al 2018)
- For more see e.g. Monaco 2016



Monaco 2016

Fast Field-Level Emulation of Cosmological Simulations

A multitude of approximate models:

- Lagrangian / Eulerian perturbation theory
  (see e.g. Zel'dovich 1970, Bernardeau 2002)
- Adhesion / Schrödinger mode
  (see e.g. Gurbatov et al 1985, Coles & Spencer et al 2002)
- **E**ffective **F**ield **T**heory of Cosmic Structures
  (see e.g Carrasco et al 2012, Schmidt et al 2018)
- For more see e.g. Monaco 2016

Residual Correction Technique:

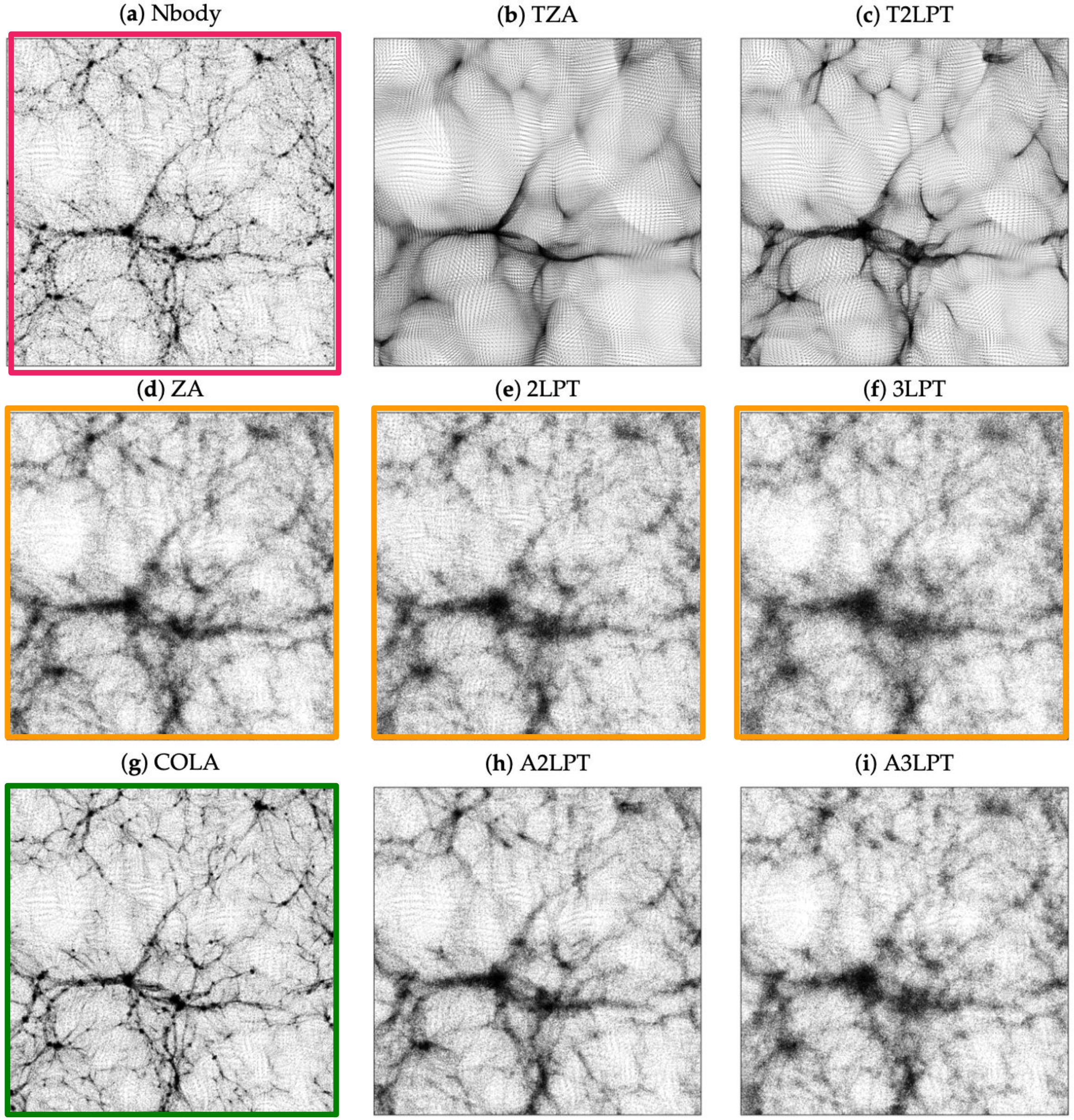
- **CO**moving **L**agrangian **A**cceleration (COLA)
  (Tassev et al 2013)

$$\boldsymbol{x} = \boldsymbol{x}_{\mathrm{LPT}} + \underline{\boldsymbol{x}_{\mathrm{res}}}$$

$$\boldsymbol{v} = \dot{\boldsymbol{x}}_{\mathrm{LPT}} + \underline{\boldsymbol{v}_{\mathrm{res}}}$$

$$\boldsymbol{F}(\boldsymbol{x}) = \ddot{\boldsymbol{x}}_{\mathrm{LPT}} + \underline{\boldsymbol{F}_{\mathrm{res}}(\boldsymbol{x})}$$

- ≈ 100 times faster than N-body simulations



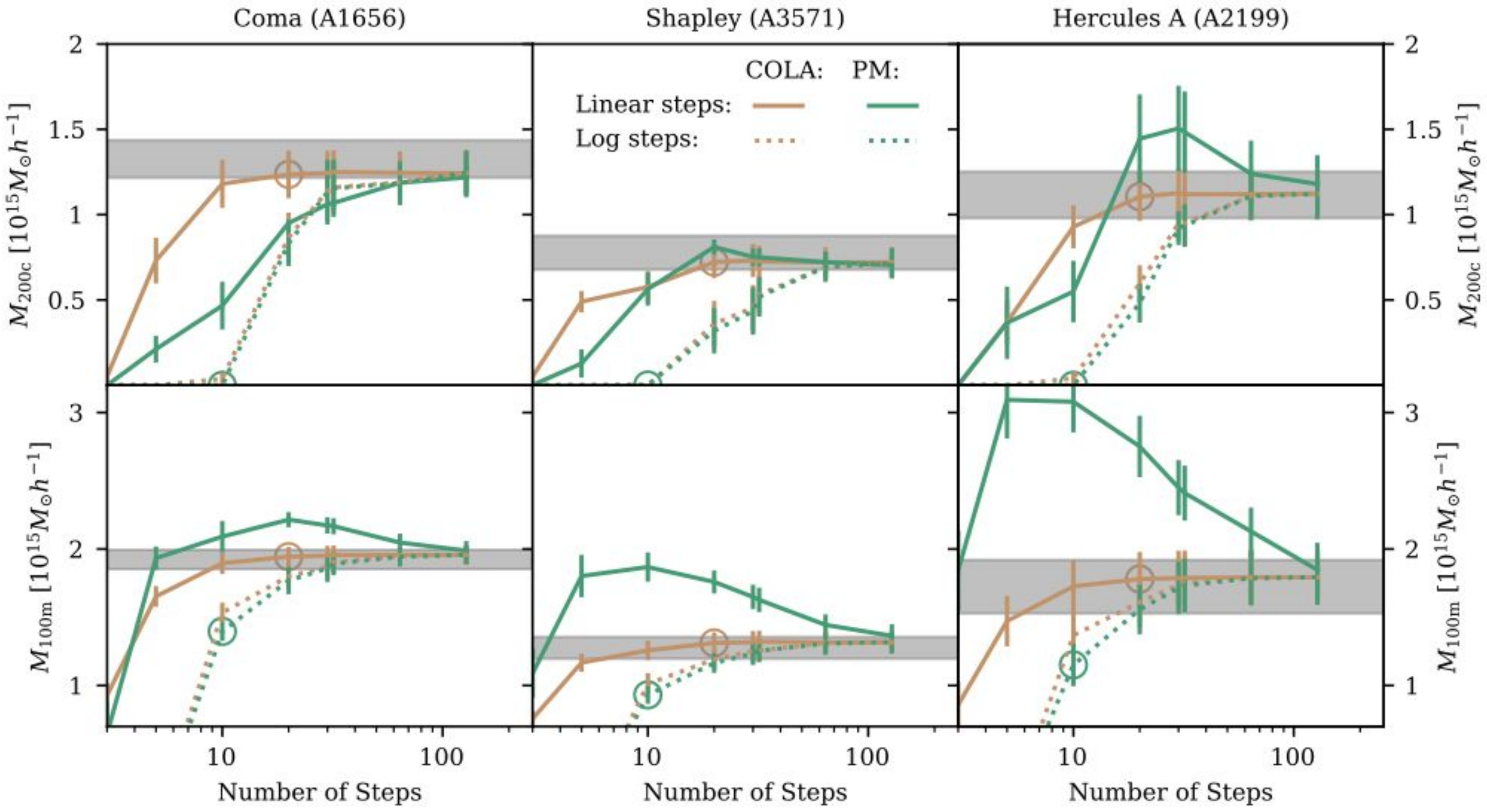
Monaco 2016

Also see e.g. Kitaura & Hess 2013, Feng et al 2016, Leclercq et al 2020

Fast Field-Level Emulation of Cosmological Simulations

Required approximation accuracy for field-level inference

- required COLA time steps to re-simulate cluster masses



Also see e.g. Ding et al 2023

Stopyra et al 2023

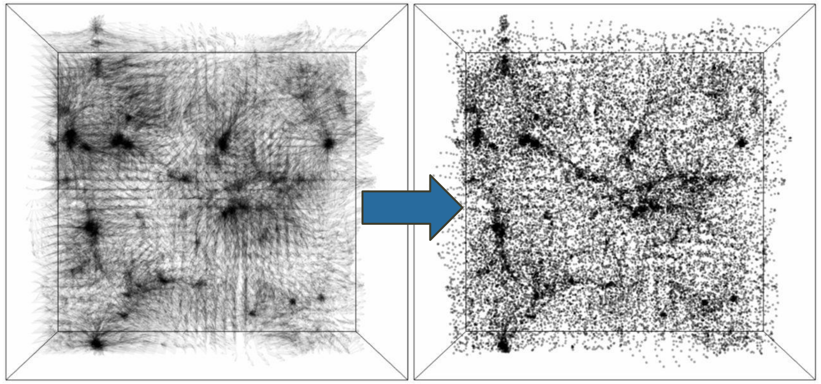
Fast Field-Level Emulation of Cosmological Simulations

Modeling residual displacements

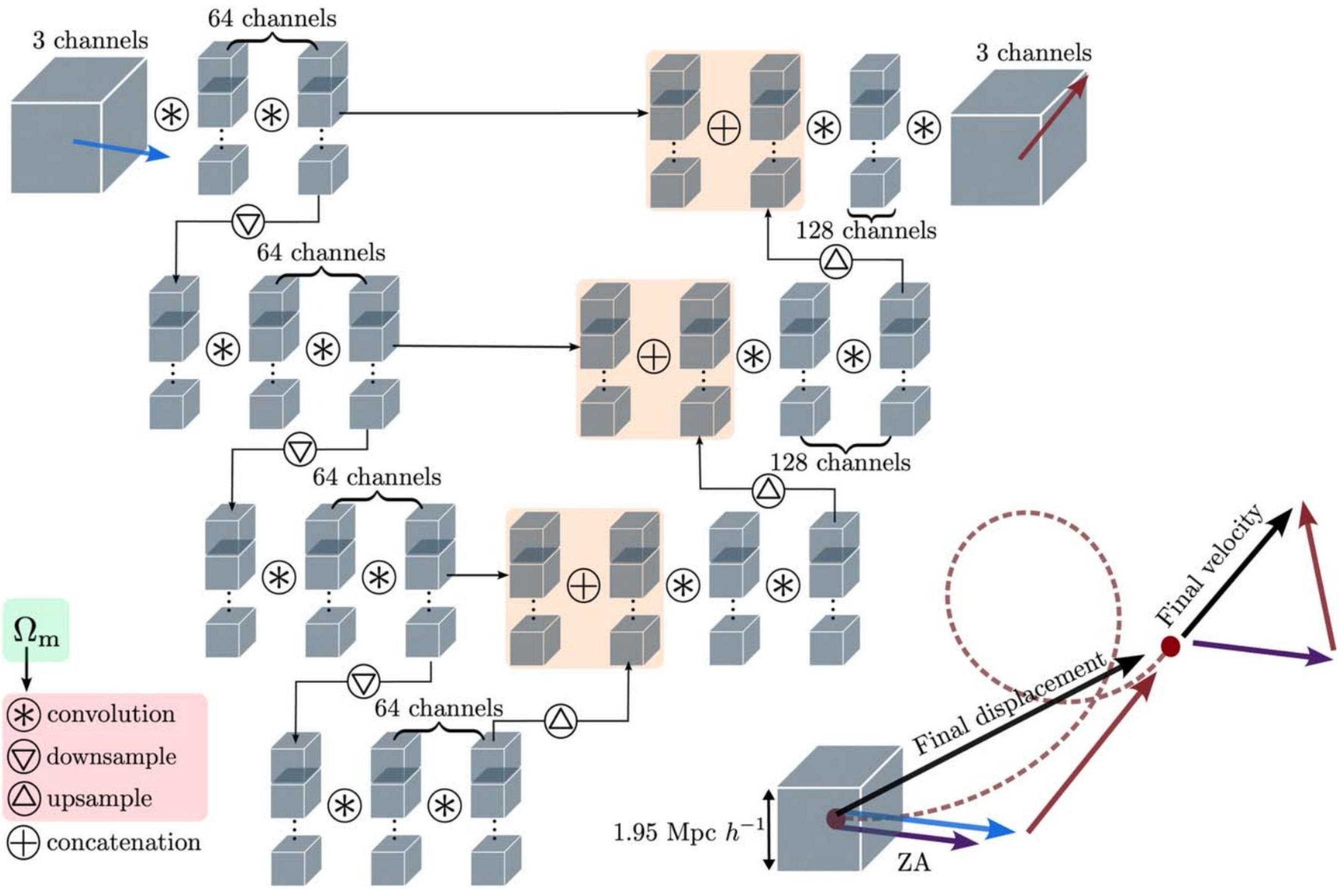
- Learns LPT residuals from paired simulations.

Deep neural network field-level emulator

- V-net (convolutional with skip connections)
- Cosmology encoded as style parameters
- Neural network enables differentiability
- Accurate alternative to 3D simulation.



Deep Density Displacement Model (D3M).



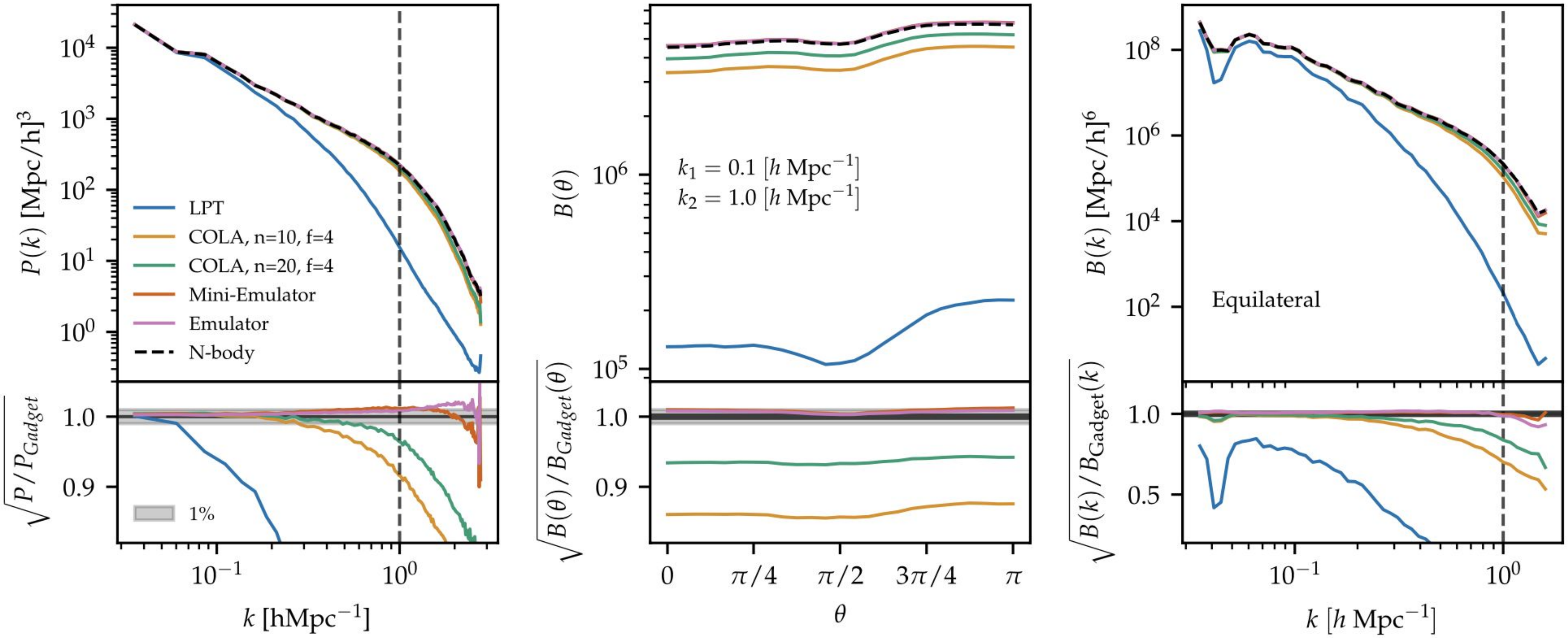
He et al 2019
Jamieson et al 2023

Also see e.g. Dai et al 2018

Fast Field-Level Emulation of Cosmological Simulations

Field-level Neural Network Emulator outperform traditional approximate models

- Power- and Bi-spectra

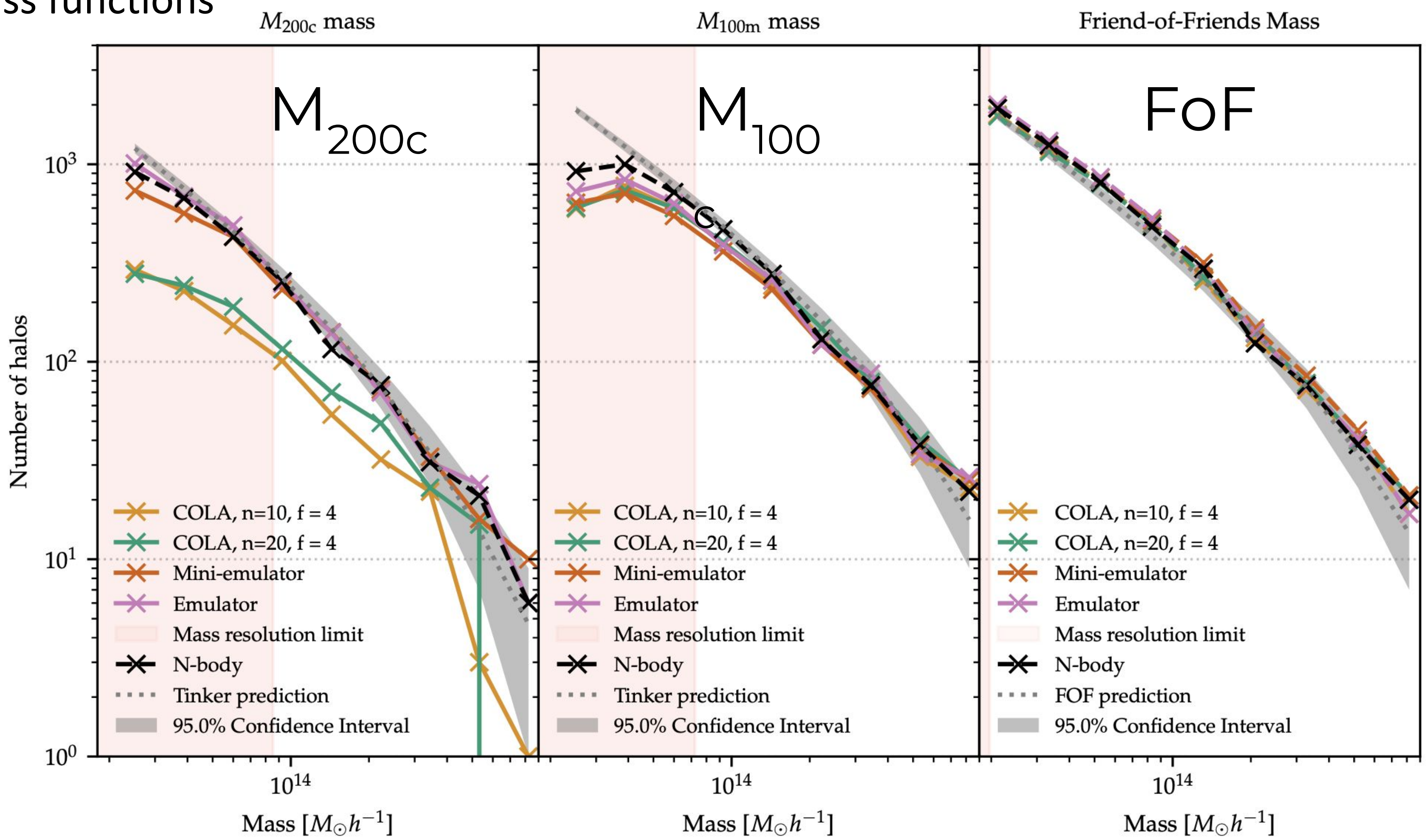


He et al 2019
Jamieson et al 2023
Doeser et al (in prep)

Fast Field-Level Emulation of Cosmological Simulations

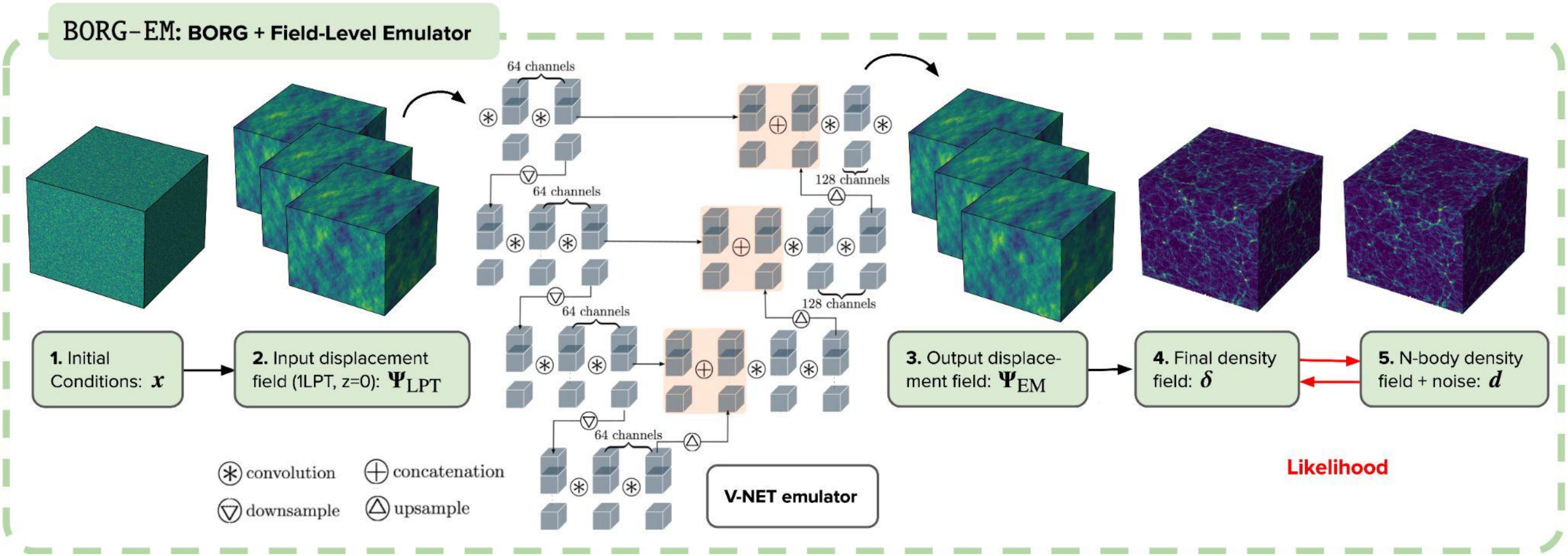
Field-level Neural Network Emulator outperform traditional approximate models

- Halo mass functions



Doeser et al (in prep)

Enabling efficient Bayesian Inference with Neural Emulators

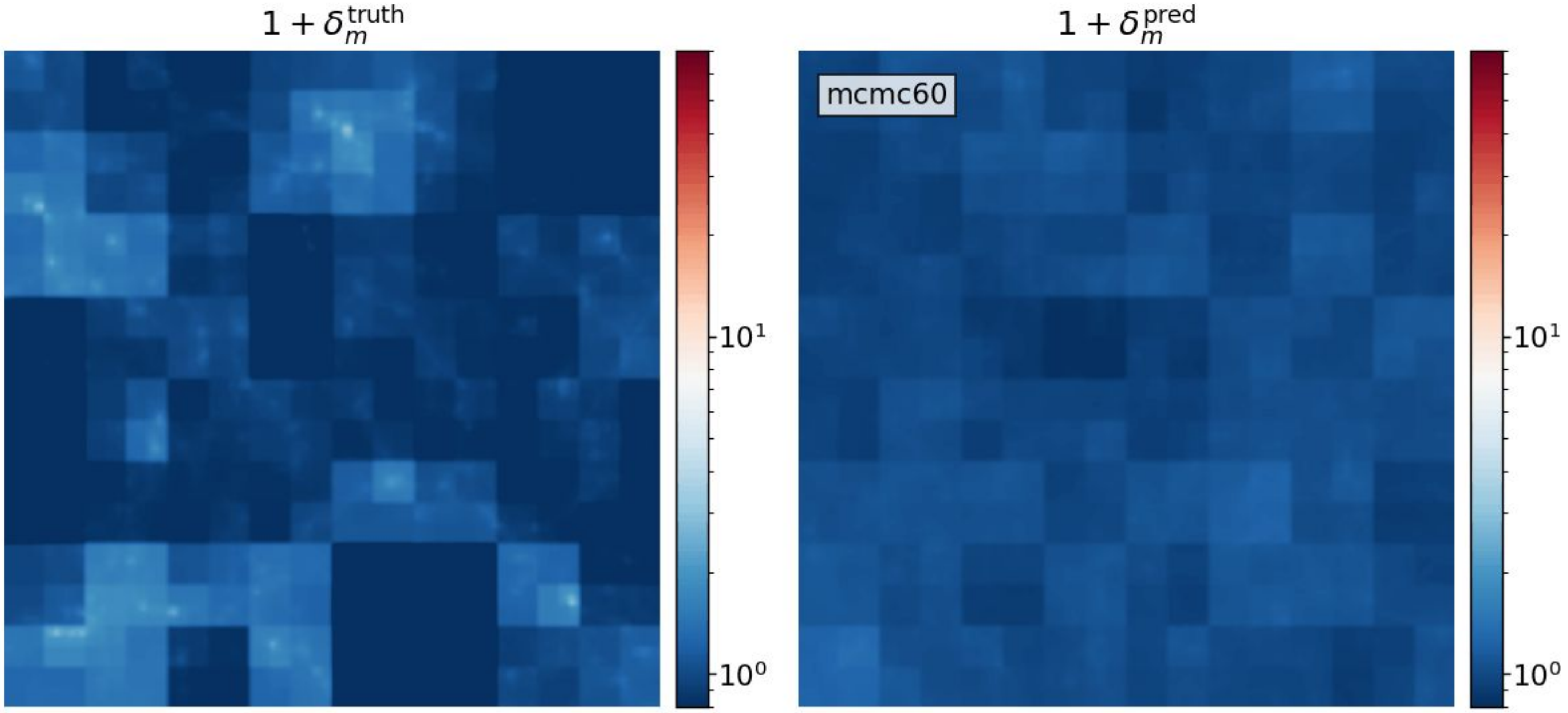


Bayesian Origin Reconstruction from Galaxies (BORG) Image Design: D.K.Ramanah et al. (2019)

slide credit: Ludvig Doeser

Doeser et al (in prep)

Enabling efficient Bayesian Inference with Neural Emulators



slide credit: Ludvig Doeser

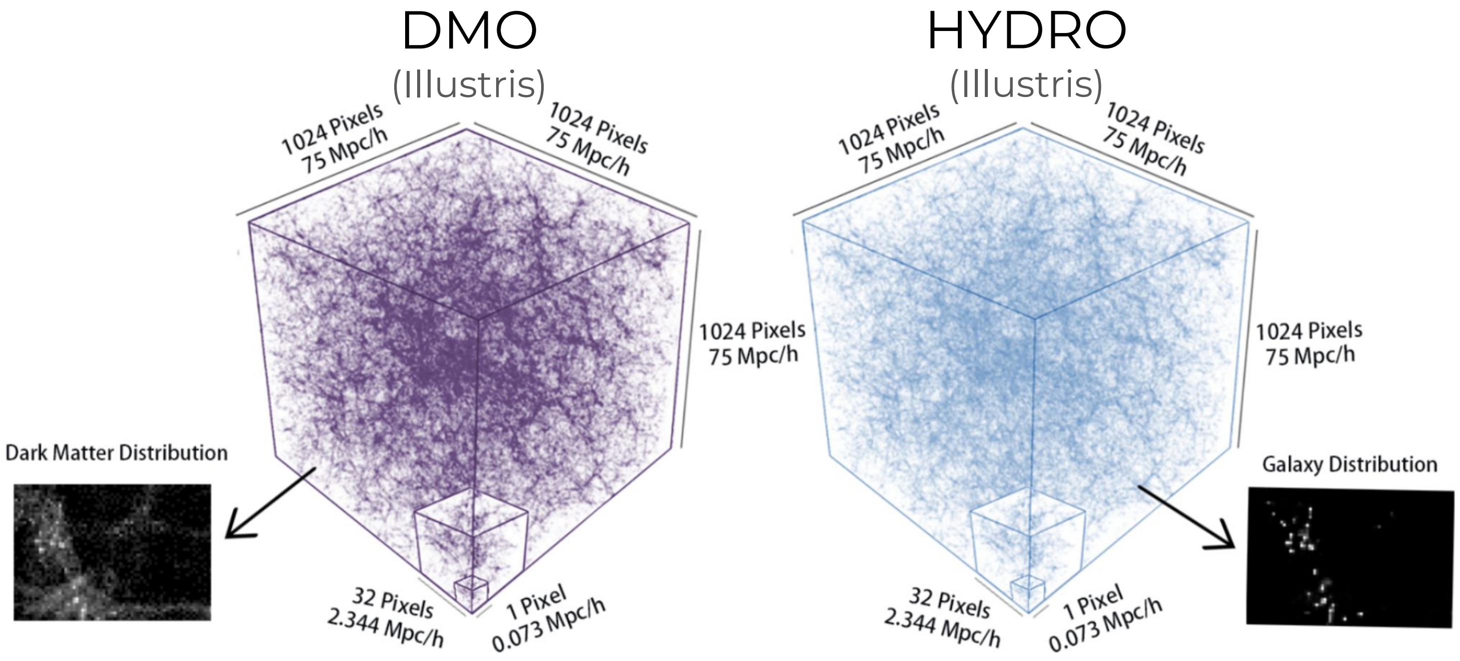
Doeser et al (in prep)

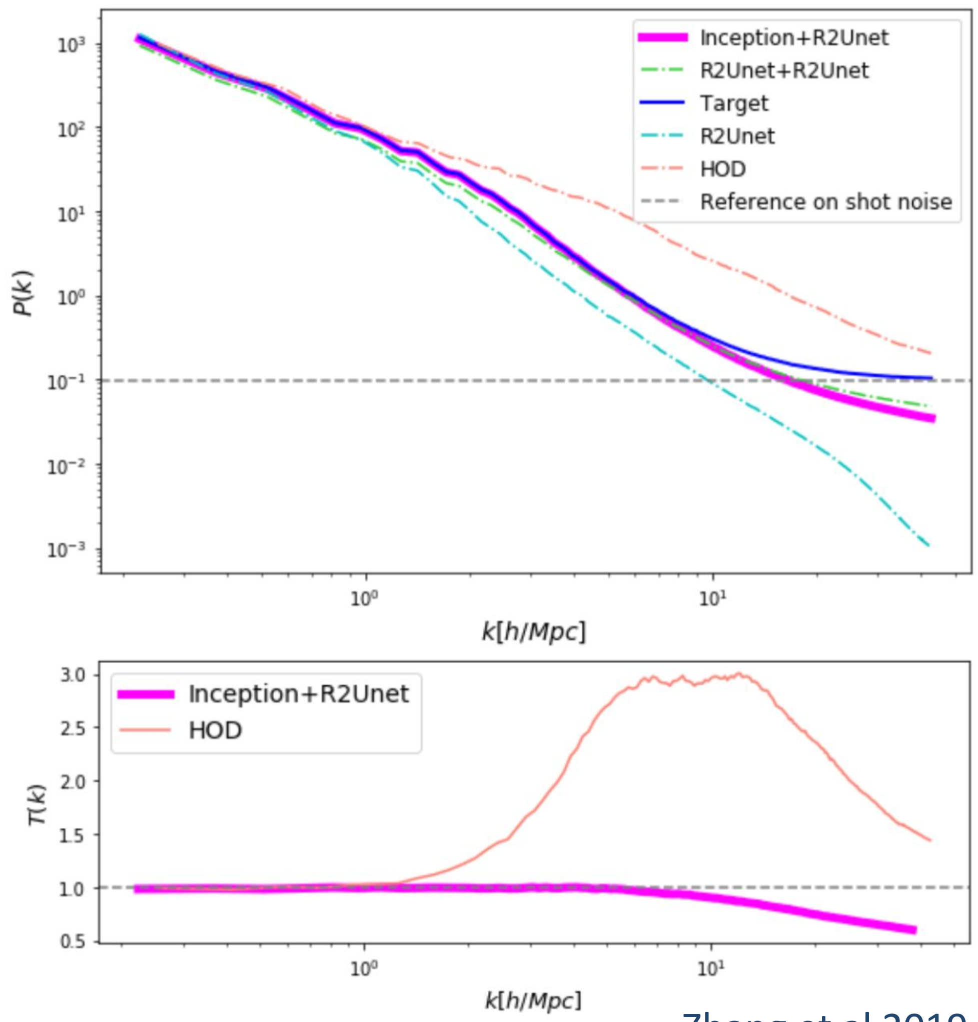
05

Galaxy Bias, the Biggest Bottleneck in Cosmological Studies

Translate Eulerian DM density to galaxy fields

- Map DMO 3D matter to galaxy fields in full hydro simulations
- Two-step learning process due to zero-inflation of data
- Unet achieves 10% level accuracy up to k = 10 *h*Mpc^{-1}



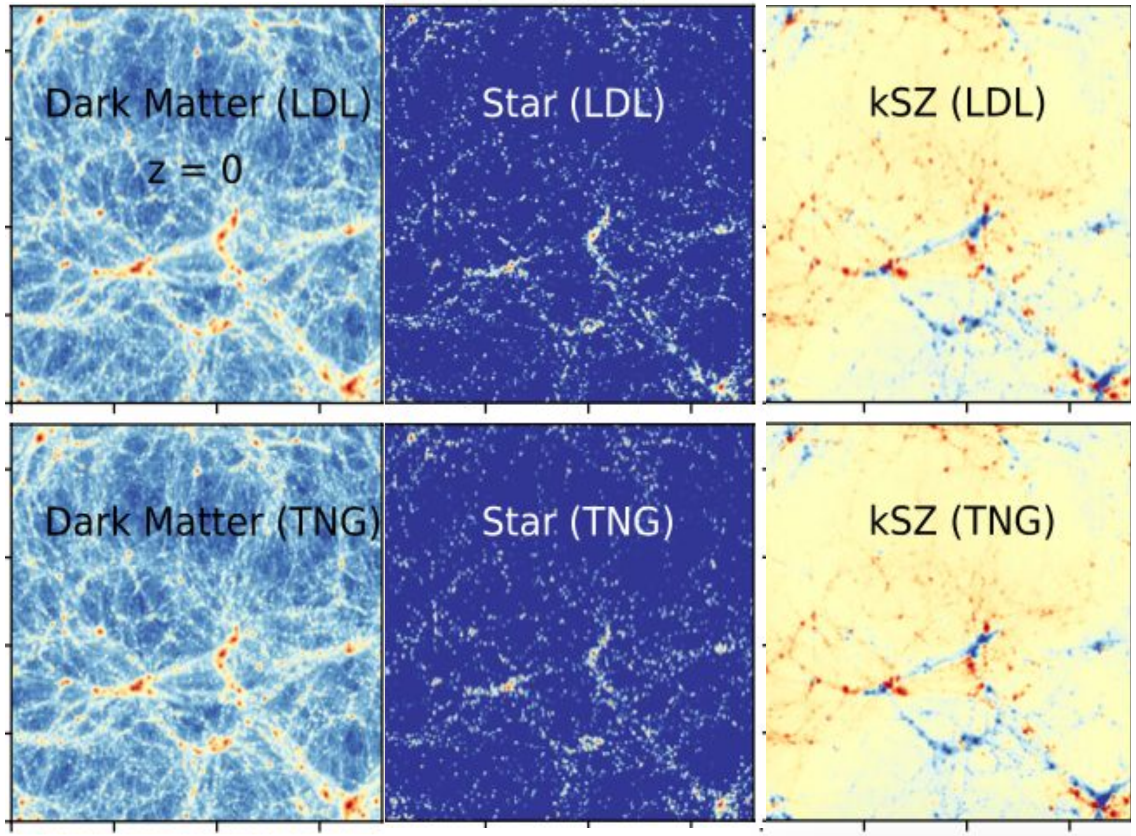


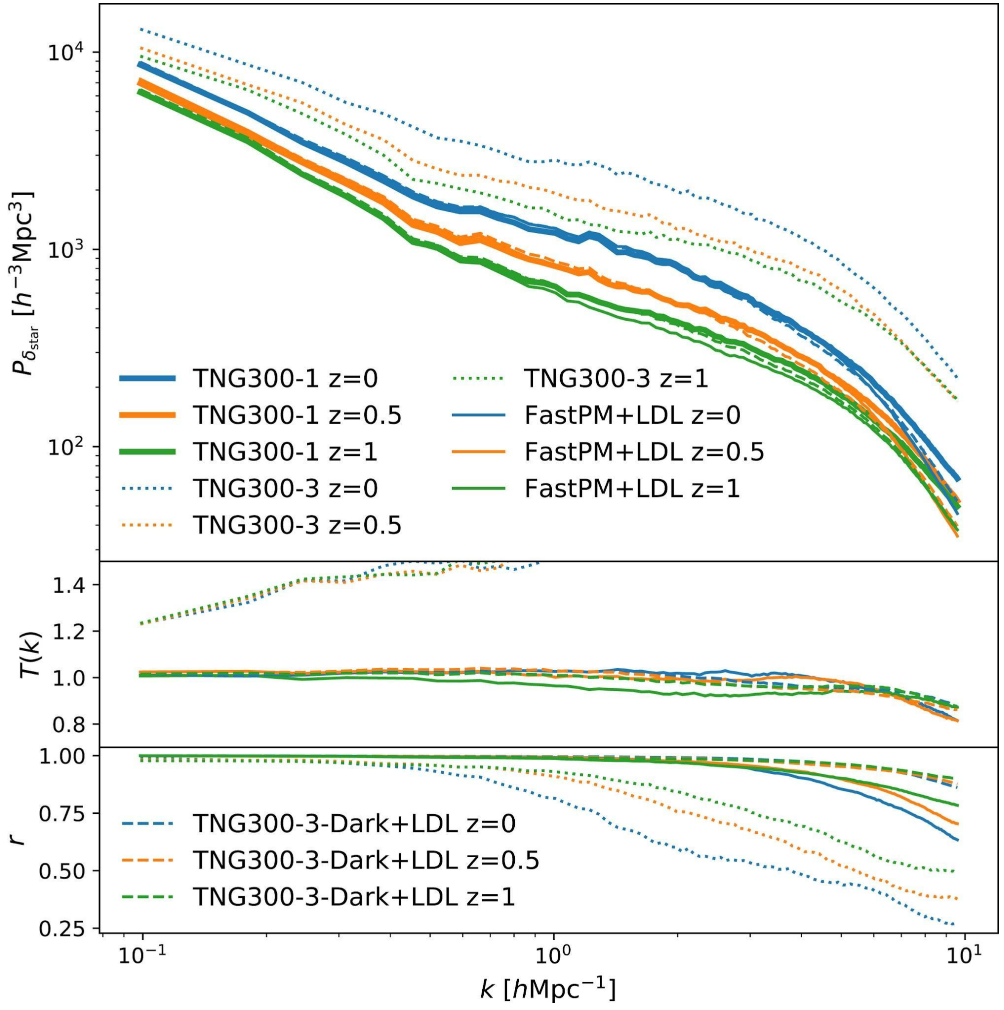
Zhang et al 2019

Also see e.g. Thiele et al 2020, Wadekar et al 2021, Bernardini et al 2022

Generate astrophysical observables by displacing particles

- Leverage translational and rotational symmetries
- Stable, easy to train, and explicitly differentiable.
- **Order 10 parameters**



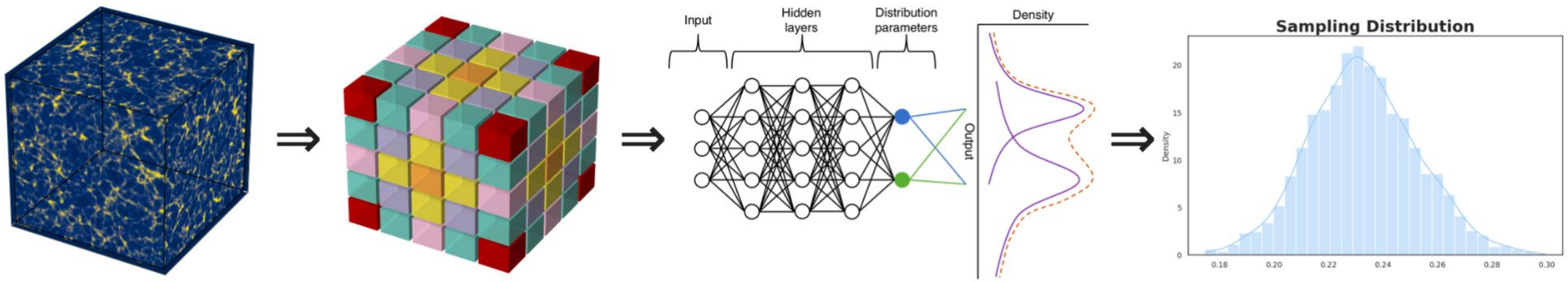


Dai & Seljak 2021

Translate Eulerian DM density to halo fields

- Encode Physical Symmetries
  - Translational invariance
  - Local rotational invariance
  - Locality

- Use non-local information
- Use isotropy
- Model non-linearity
- Generative process



$\delta_m(x) \rightarrow$ **Convolve** $\rightarrow$ **Transform** $\rightarrow$ **Sample** $\rightarrow$ $n(M|\delta_m(x))$

Ding et al in prep
Charnock et al 2020

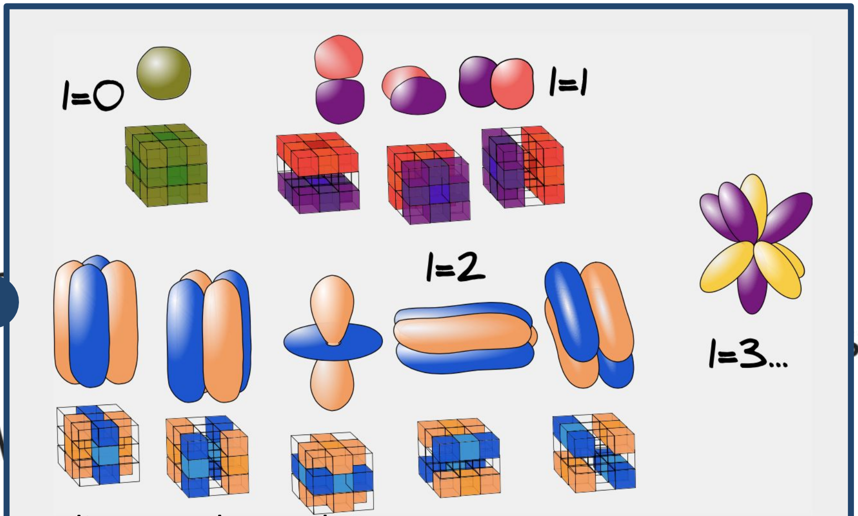
Also see Dai & Seljak 2021 for an analogous Lagrangian approach.

credit for slide: Simon Ding

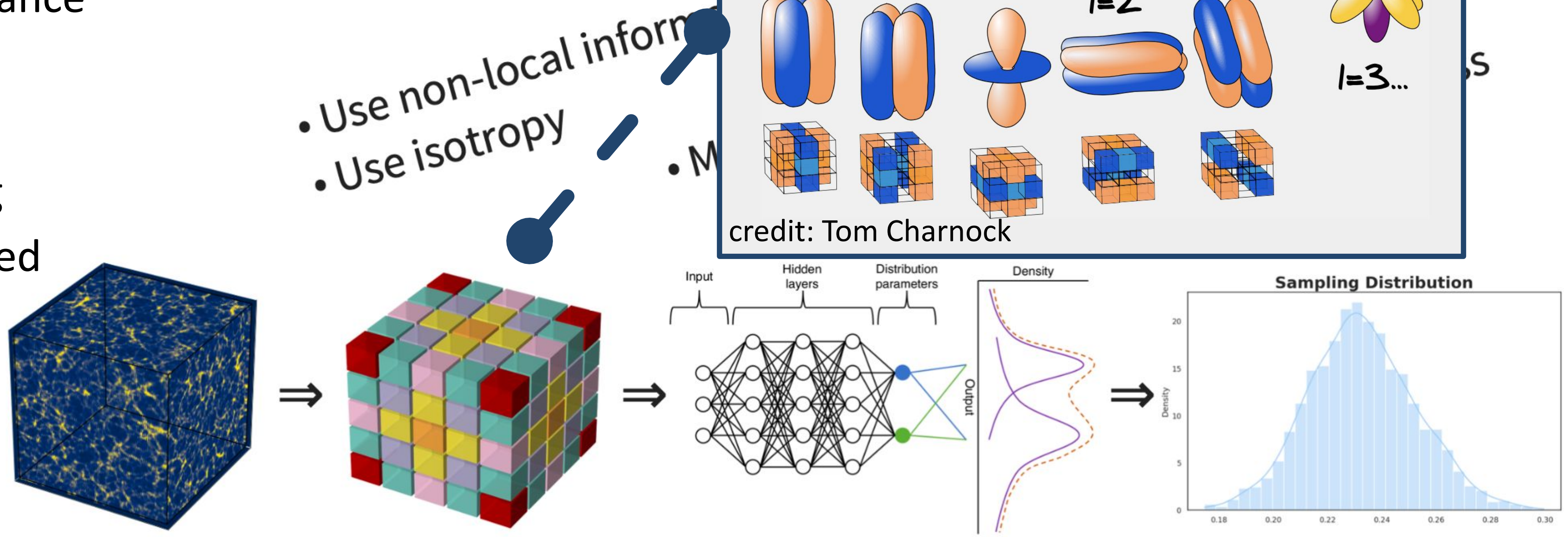
Modeling Galaxy bias with ML: Neural Physical Engine

Translate Eulerian DM density to halo fields

- Encode Physical Symmetries
  - Translational invariance
  - Local rotational invariance
  - Locality
- **Only 17 parameters**
- Enables zero-shot learning
  - no training data needed



credit: Tom Charnock



$\delta_m(x) \to$ **Convolve** $\to$ **Transform** $\to$ **Sample** $\to$ $n(M|\delta_m(x))$

Ding et al in prep
Charnock et al 2020

Also see Dai & Seljak 2021 for an analogous Lagrangian approach.

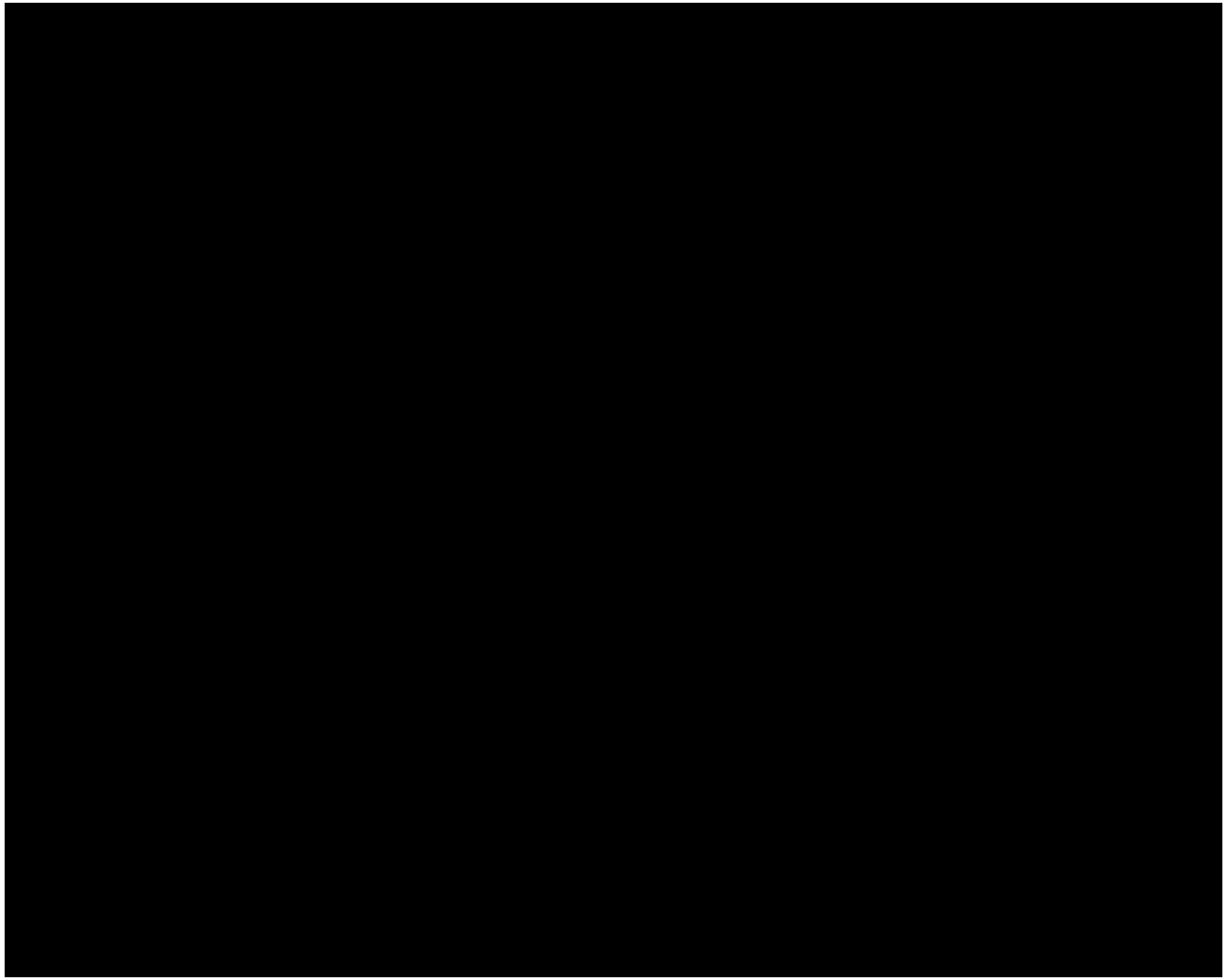
credit for slide: Simon Ding

Modeling Galaxy bias with ML: Neural Physical Engine

Simultaneous Inference of Initial Conditions and Halo Fields

DM field at z=0

Data = Halo Catalog



Inferred NPE halo model

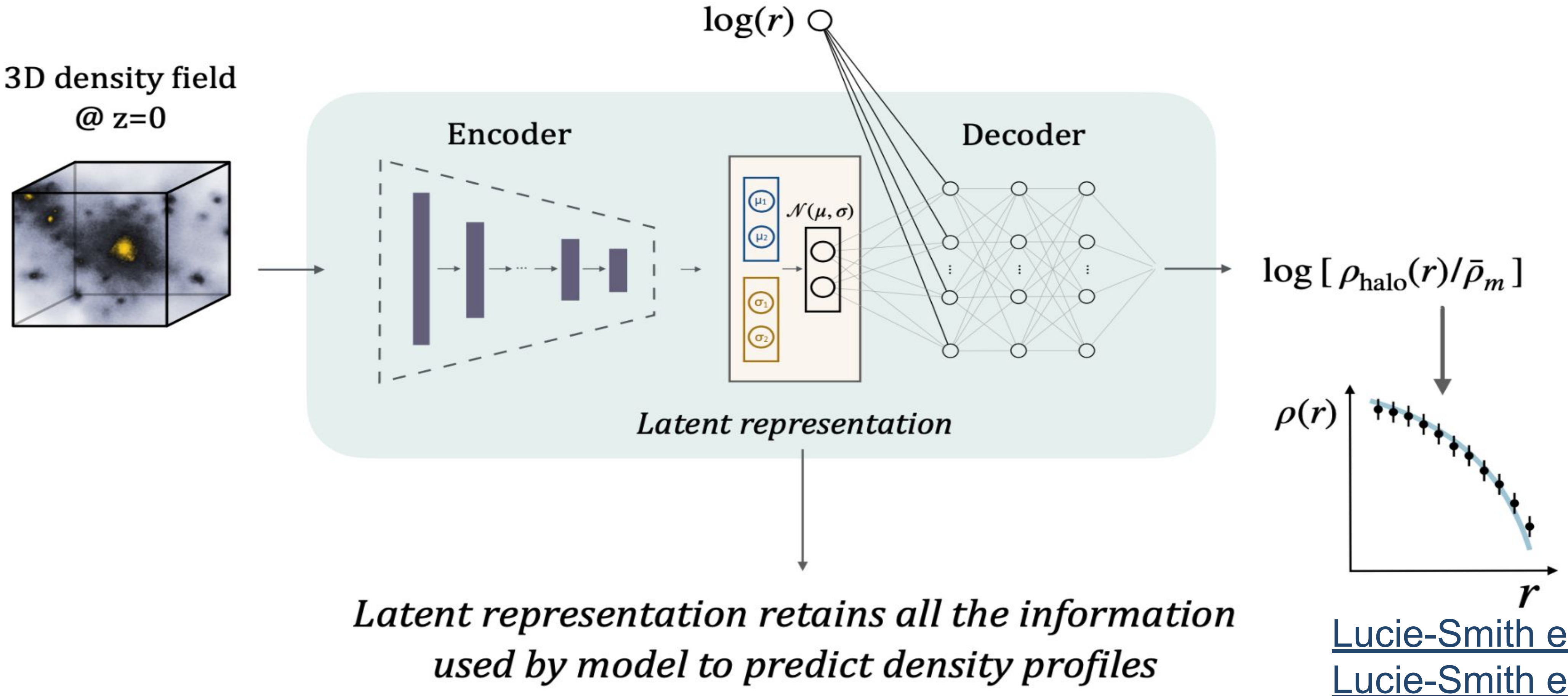
video credit: Tom Charnock

06

Learning to Learn from ML

Learning Dark Matter Halo Density Profiles:

- Supervised encoder-decoder framework
- Compresses inputs into a low-dimensional latent representation.
- Outputs ρ(r) for any desired value of radius r.

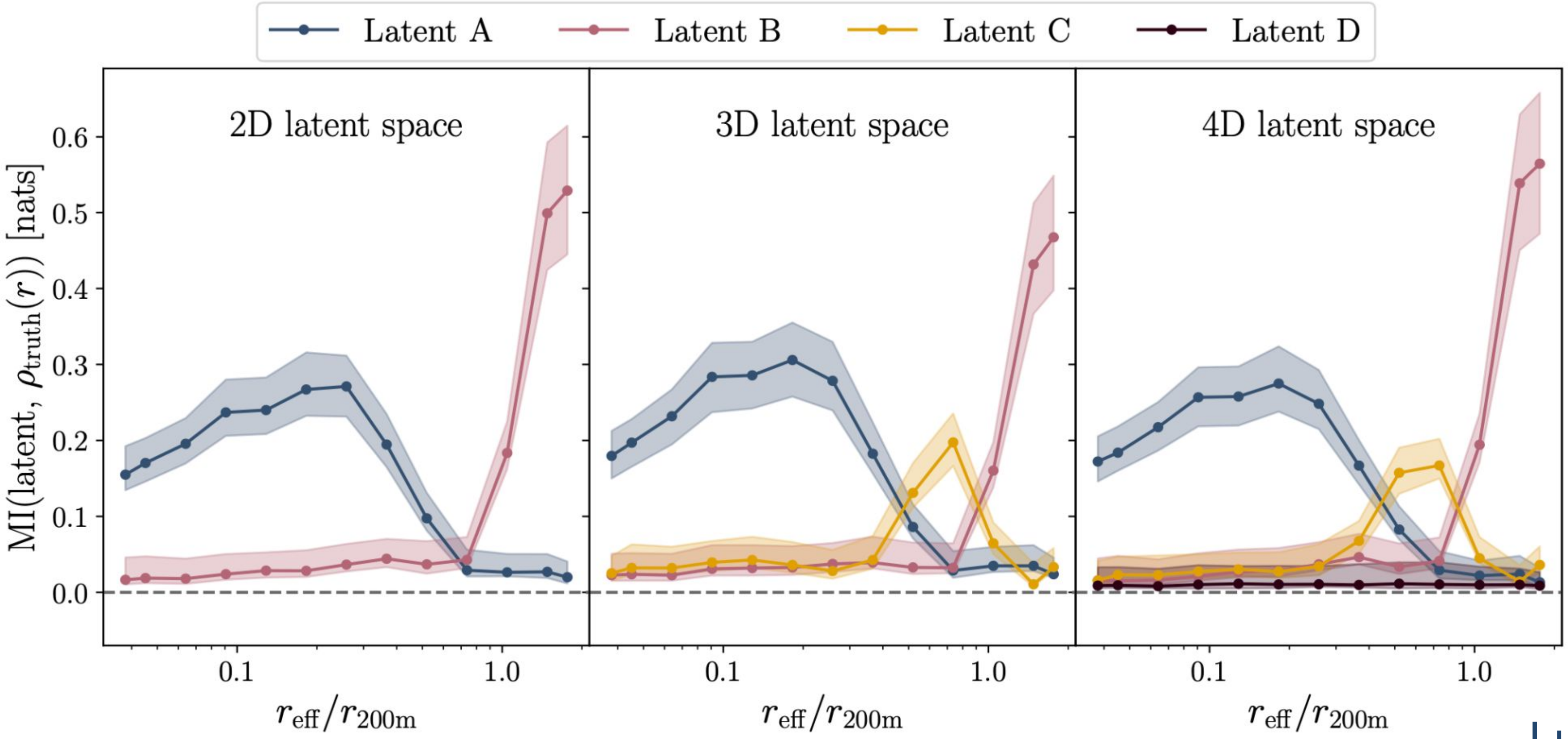


Lucie-Smith et al 2023
Lucie-Smith et al 2022

Explainable AI: Dark Matter Halo density profiles

Interpreting the latent representation using mutual information

- 2D: Accurately models density profiles up to the virial radius.
- 3D: Required for describing outer profiles beyond the virial radius.
- 4D+: Reveals infalling material in outer profiles and splashback boundaries.



[Lucie-Smith et al 2023](#)
[Lucie-Smith et al 2022](#)

Optimal machine-driven acquisition of future cosmological data

Predict remaining information content of 3D cosmic structures

- Optimally answer physics questions
- Propose optimal regions for follow-up observations

Fisher information:

Parameter Of Interest (POI)

$$\mathcal{I}(\theta|\phi) = \mathbb{E}_{(d|\theta,\phi)}\left[\left(\frac{\partial \ln(\mathcal{L}(d|\theta,\phi))}{\partial\theta}\right)^2\right]$$

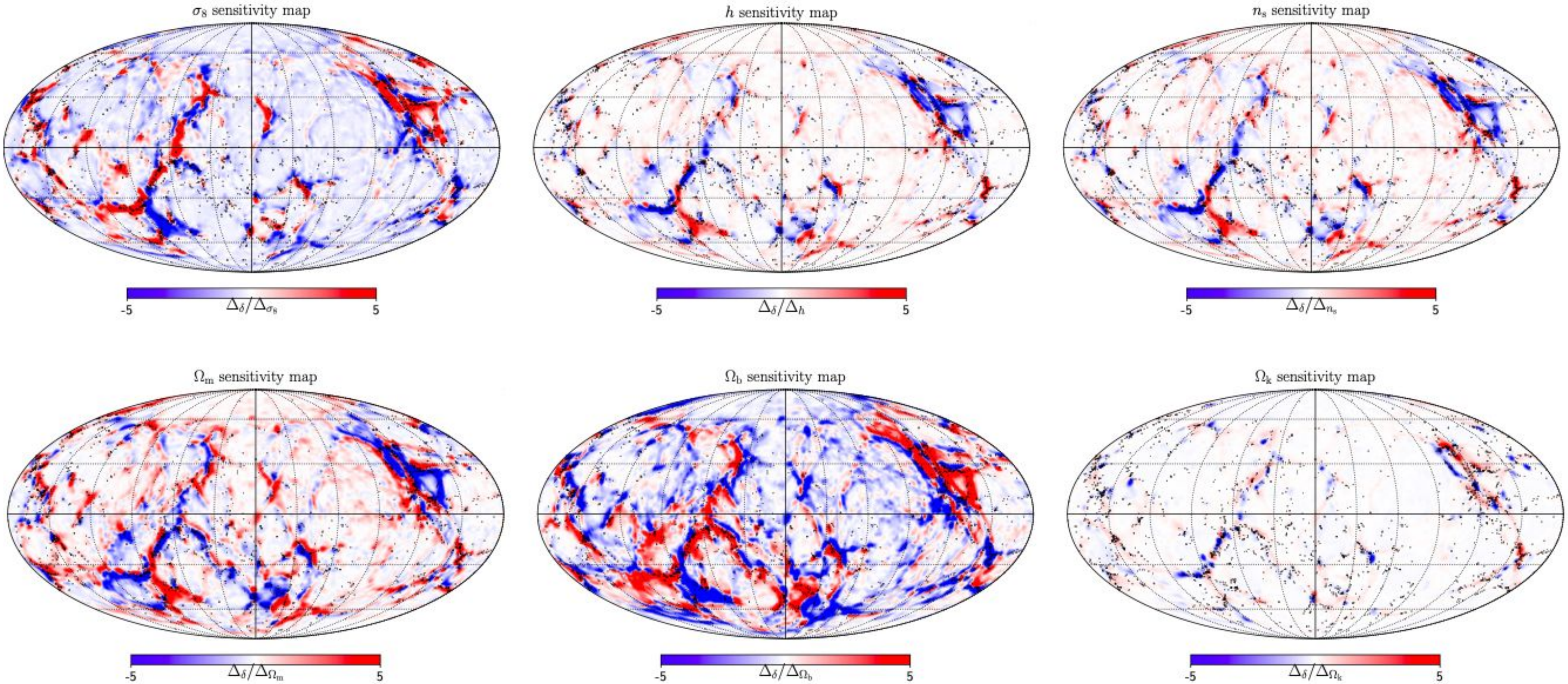
Inferred Initial conditions

$$= \int \mathcal{D}d \left(\frac{\partial \ln(\mathcal{L}(d|\theta,\phi))}{\partial\theta}\right)^2 \mathcal{L}(d|\theta,\phi)$$

Parameter Sensitivity

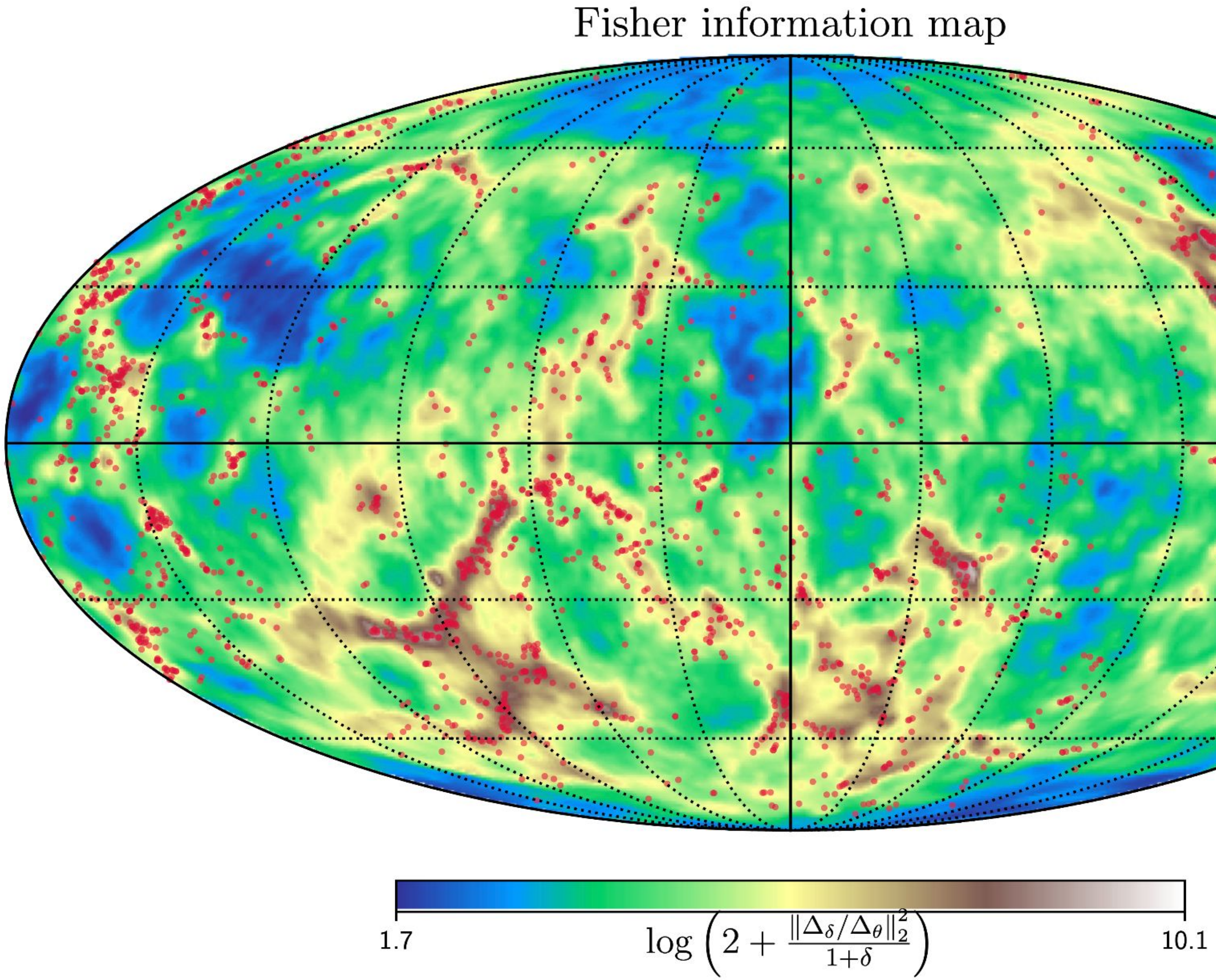
Kostic et al 2022

Parameter Sensitivities



Kostic et al 2022

A Fisher Information Map of the Universe



Kostic et al 2022

07

Summary & Conclusion

Summary & Conclusions

1

Cosmological data will no longer be scarce:

- Our capability to analyse data will limit knowledge gains

2

Inference Technology:

- **Field-level Inference:**
  - Information Optimality
  - The complete characterization of cosmic structure (no compression)
- **Implicit Likelihood Inference:**
  - Simulation Based Inference (no explicit likelihood)
  - Information Optimal Data Compression
  - A viable alternative to classical MCMC

3

ML enhanced data modeling

- Accurate and fast Physics Emulators
- Automatic feature design (e.g. information maximizing summaries)

4

Current limitations:

- Handling of small-scale baryonic physics and galaxy biasing
- Training data requirements
  - Accuracy, Size, Diversity, Representativeness