The non-Gaussian Universe workshop - Heraklion June 16-19

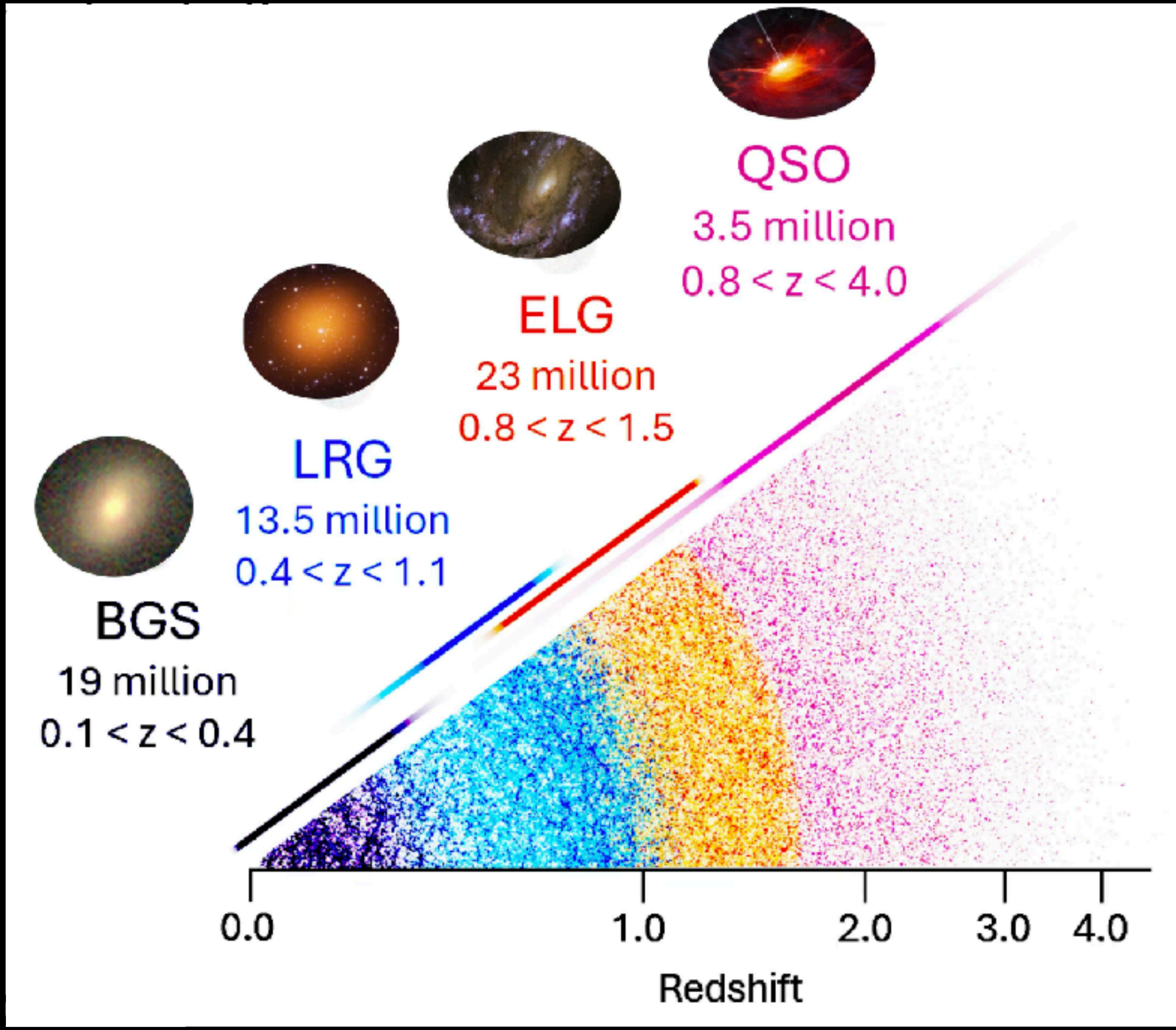
Beyond 2pt galaxy clustering with DESI

Pauline Zarrouk (CNRS-LPNHE, Sorbonne University)

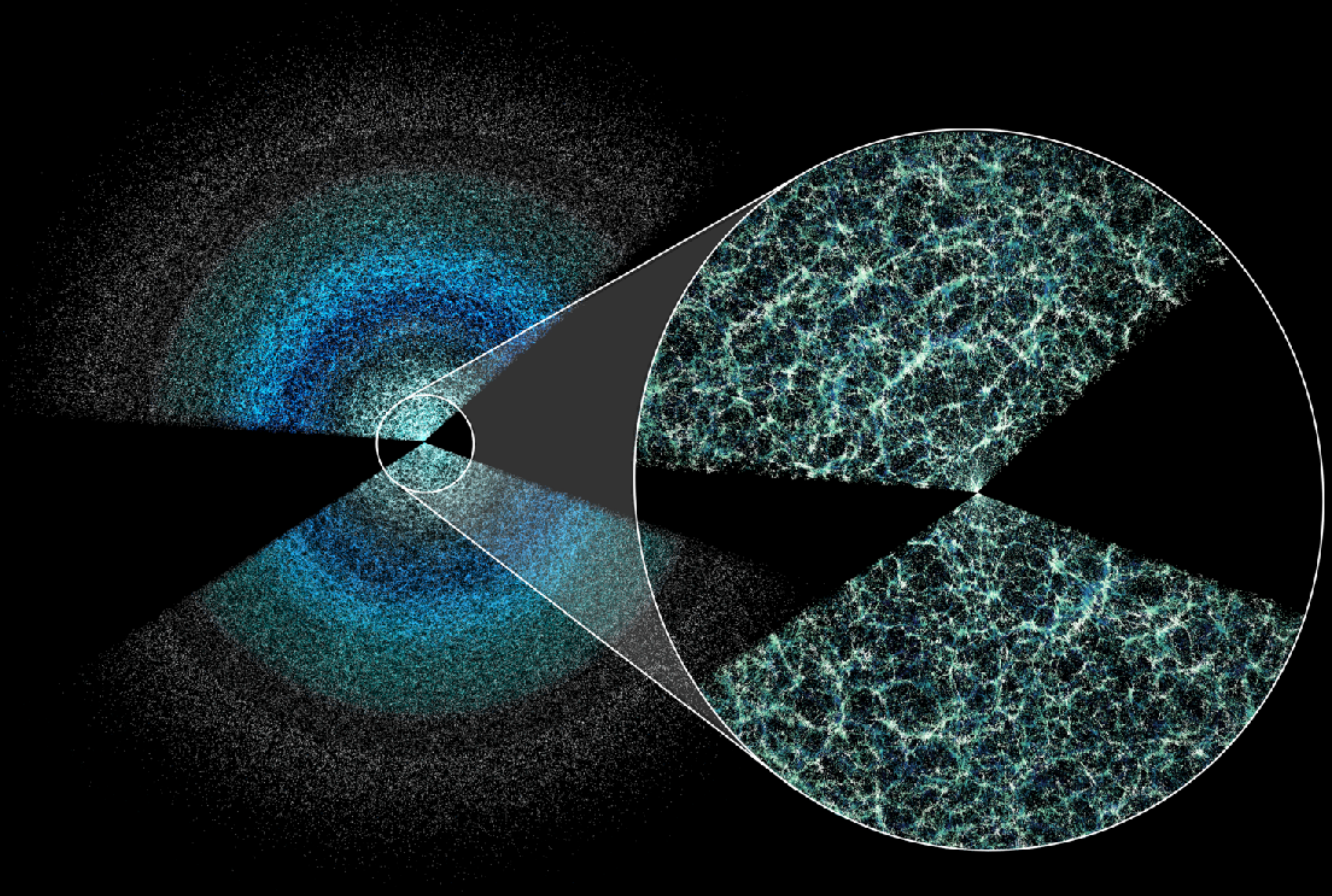
Dark Energy Spectroscopic Instrument (DESI)

DESI galaxy samples

DESI-I (2021-2028)
4m telescope with 5,000 robotic fibres
About 60 million spectra of galaxies and quasars



(Initial) 5-year survey complete since April 2026



How to extract cosmological information?

Sensitivity to cosmology

Inflation and primordial features

Dark matter and neutrinos

Dark energy

redshift

Initial Gaussian and linear density field

How to extract cosmological information?

Sensitivity to cosmology

Inflation and primordial features — Dark matter and neutrinos — Dark energy

redshift

Initial Gaussian and linear density field

⟵ more robust

Two-point statistics (2pt)

- Most mature and robust technique
- Analytical model available up to quasi-linear scales (k_{max}=0.2-0.3 h/Mpc)
- Computationally cheap
- Only Gaussian and linear information

DESI DR1 Galaxy Full-Shape, DESI collaboration
(arXiv:2411.12021 (co-lead), arXiv:2411.12022)
—> 2pt statistics / EFT

How to extract cosmological information?

Sensitivity to cosmology

Inflation and primordial features — Dark matter and neutrinos — Dark energy

redshift

Initial Gaussian and linear density field — *Today **non-Gaussian** and **non-linear** density field*

← more robust — more optimal →

Two-point statistics (2pt)
- Most mature and robust technique
- Analytical model available up to quasi-linear scales (k_{max}=0.2-0.3 h/Mpc)
- Computationally cheap
- Only Gaussian and linear information

Higher-order statistics (HOS)
- Capture some non-Gaussian information
- Mostly simulation-based model up to k_{max}=0.8 h/Mpc
- Computationally cheaper than FLI but still requires thousands of cosmological simulations
- Factor 2 improvement w.r.t. 2pt*

DESI DR2 Full-Shape (papers this Fall)
—> 2pt+3pt statistics / EFT

*Euclid collaboration V 2023

Part 1: Alternative Clustering Methods (ACM) within DESI

ACM group: Gillian Beltz-Mohrmann, Davide Bianchi, Simon Bouchard, Alena Casella, Jean-Marc Colley, Carolina Cuesta-Lazaro, Daniel Forero-Sanchez, Nathan Findlay, Katayoon Ghaemi, Elisabeth Krause, Wei Liu, Sesh Nadathur, Krishna Naidoo, Enrique Paillas, Mathilde Pinon, Alice Pisani, Hernan Rincon, Nico Schuster, Georgios Valogiannis and Pauline Zarrouk

ACM
Alternative Clustering
Methods

Alternative Clustering Methods within the DESI

ACM wishing list

- One single pipeline for all statistics
- One single type of simulation-based model trained on the same suite of N-body simulations
- Standardised pipeline for plug-and-play applications (unit tests, functional tests, automated integration) —> dedicated software engineer is a real asset

External constraints

- Mock-based covariance
- Explicit likelihood inference

Dark matter halo catalogues (Abacus N-body simulations[1])

↓

Galaxy mock catalogues AbacusHOD[2]

↓

Summary statistics

↓

Simulation-based model (SUNBIRD[3])

↓

Parameter inference

↓

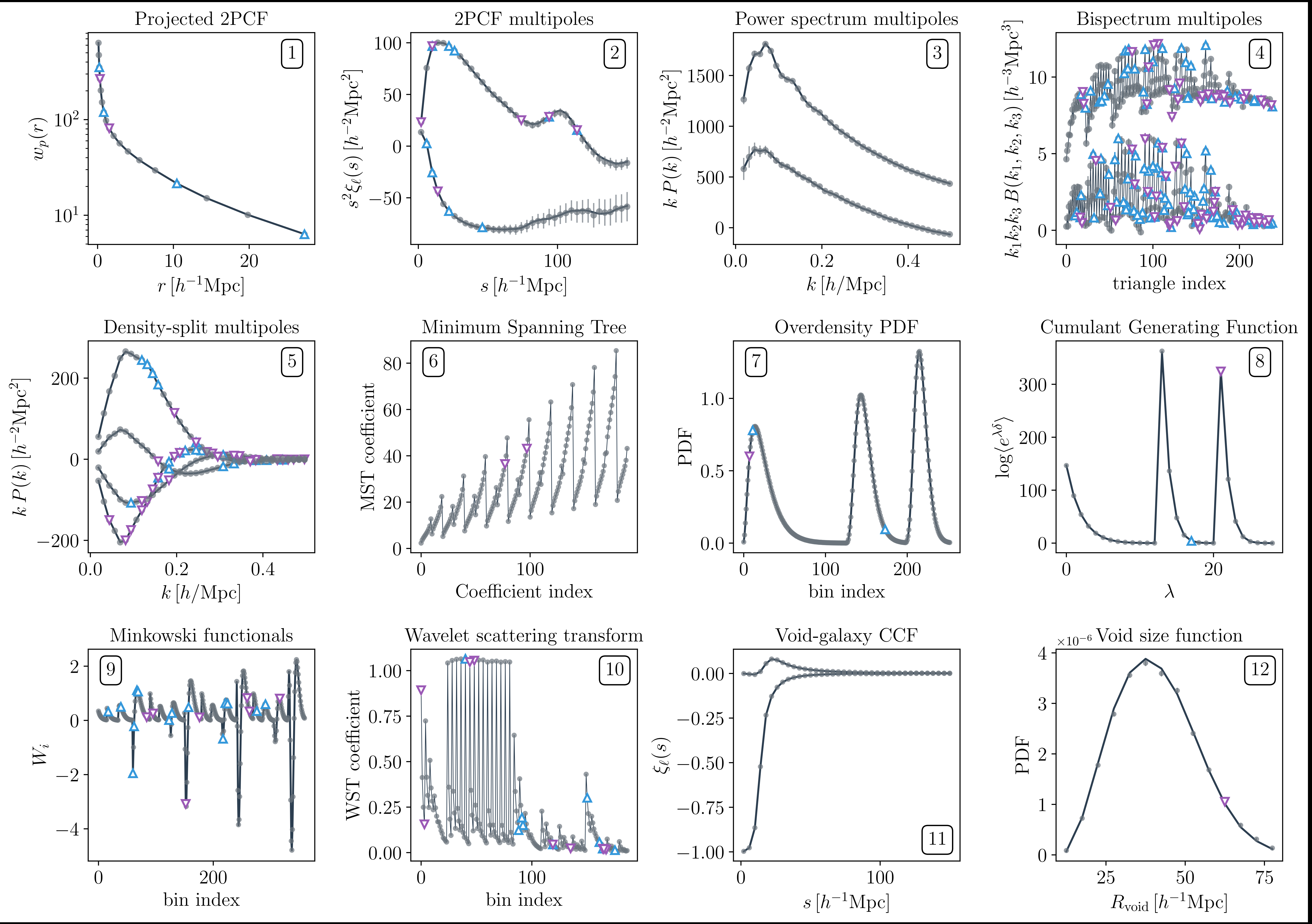
Cosmological constraints

[1] Maksimova et al. 2021

[2] Yuan et al. 2022

[3] Cuesta-Lazaro et al. 2023

Alternative Clustering Methods within the DESI



DESI ACM pipeline, *in prep*

Dark matter halo catalogues (Abacus N-body simulations[1])

[1] Maksimova et al. 2021

↓

Galaxy mock catalogues AbacusHOD[2]

[2] Yuan et al. 2022

↓

Summary statistics

↓

Simulation-based model (SUNBIRD[3])

[3] Cuesta-Lazaro et al. 2023

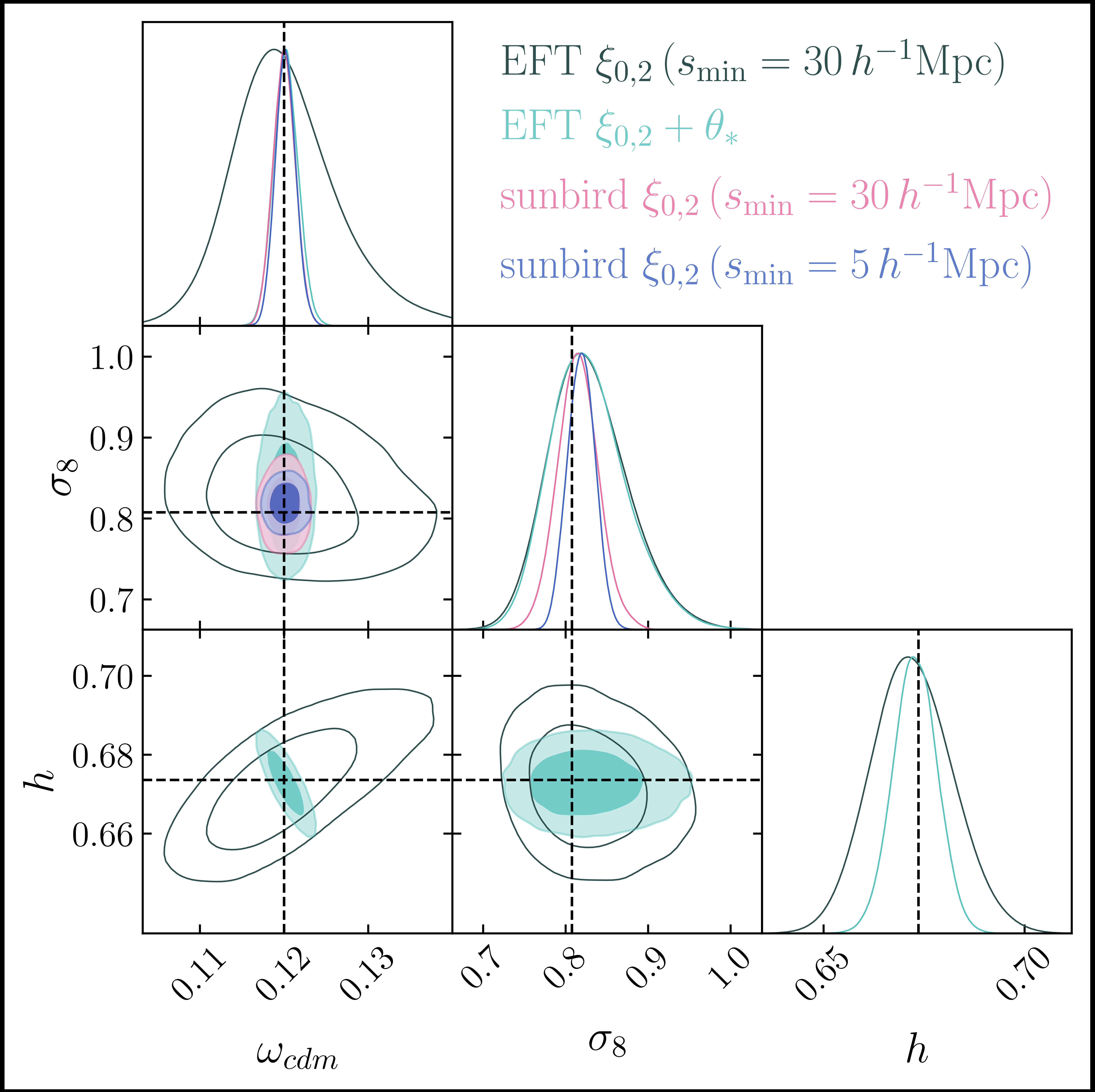
↓

Parameter inference

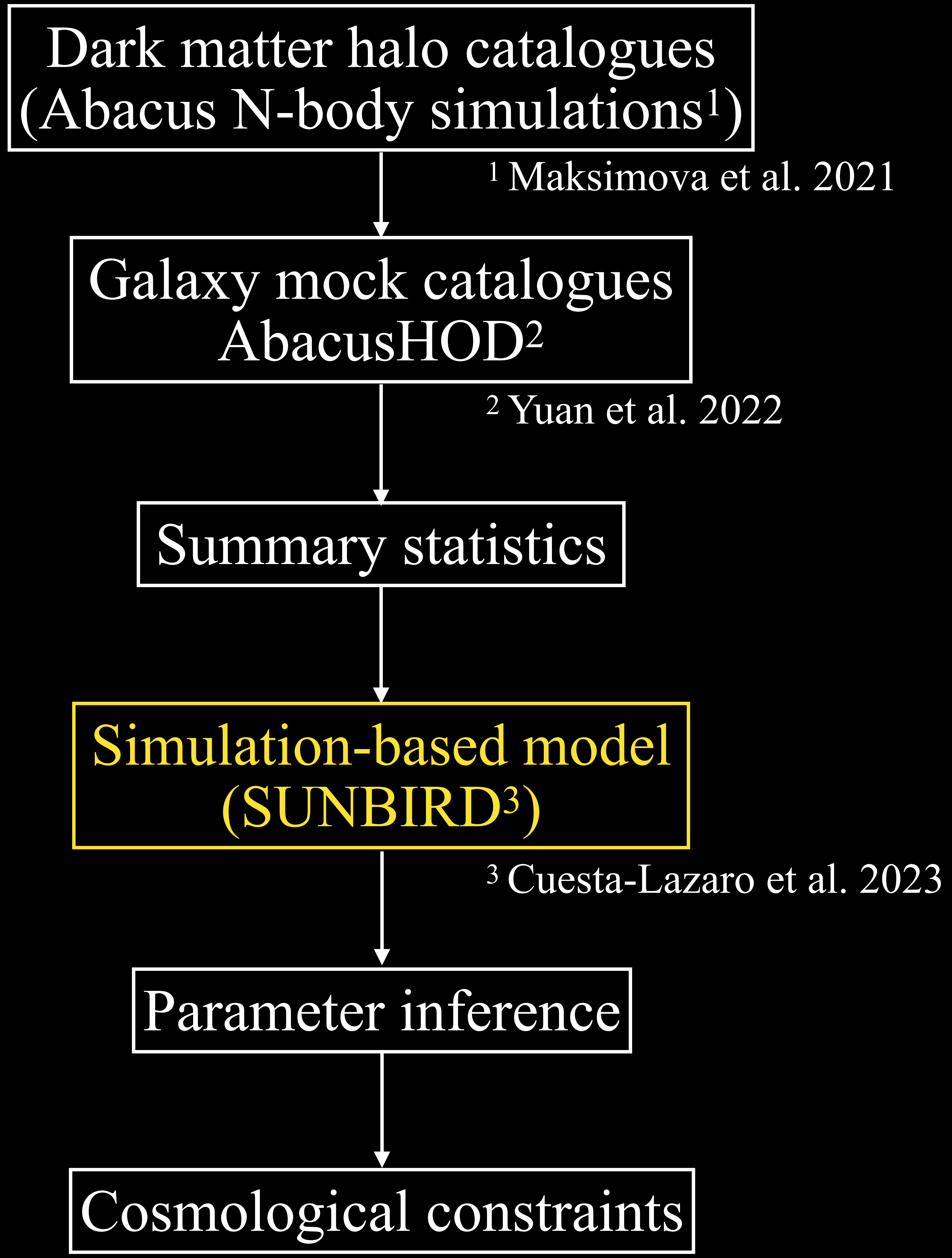
↓

Cosmological constraints

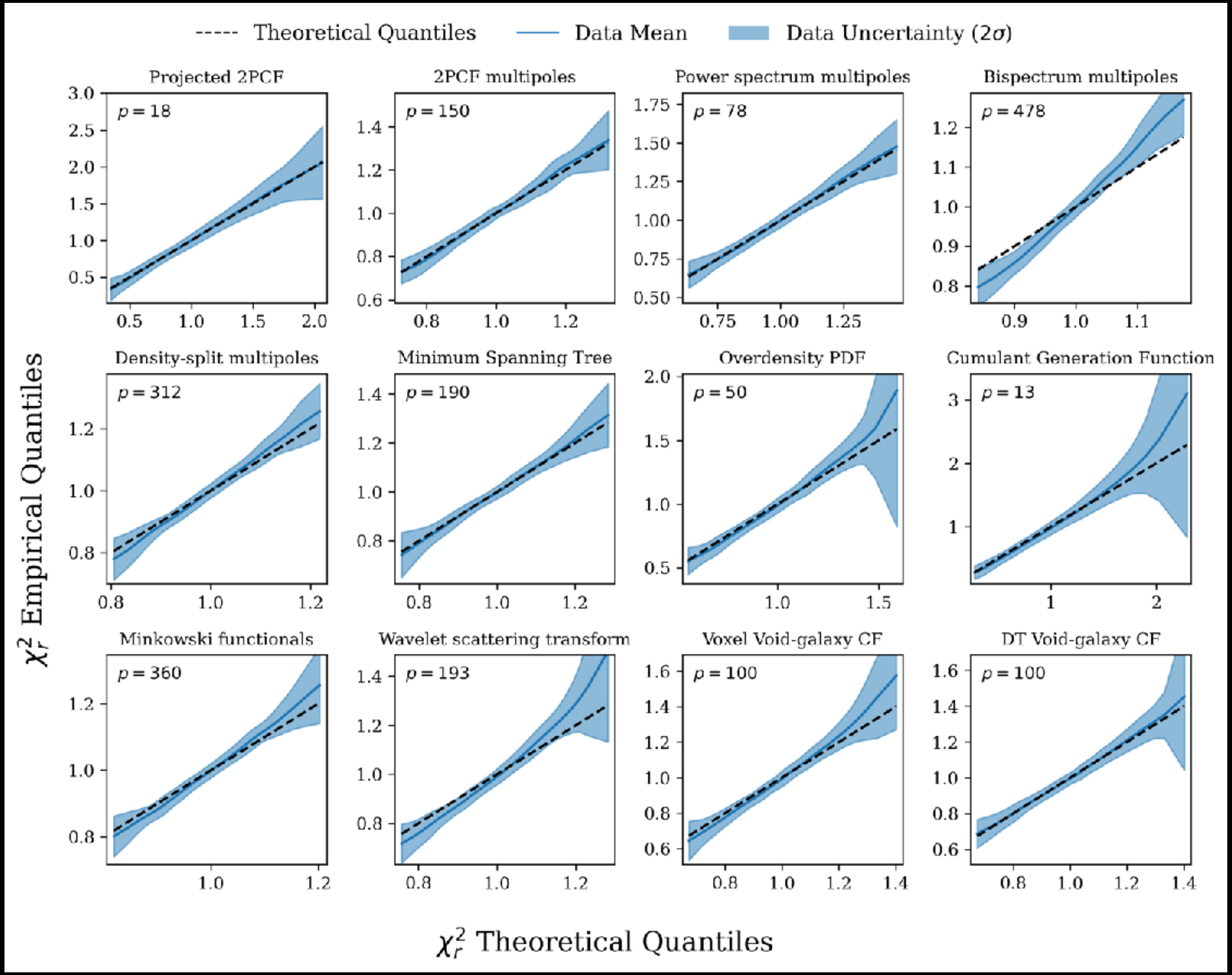
Alternative Clustering Methods within the DESI



DESI ACM pipeline, *in prep*



Alternative Clustering Methods within the DESI



DESI ACM pipeline, *in prep*

Dark matter halo catalogues (Abacus N-body simulations[1])

[1] Maksimova et al. 2021

Galaxy mock catalogues AbacusHOD[2]

[2] Yuan et al. 2022

Summary statistics

Simulation-based model (SUNBIRD[3])

[3] Cuesta-Lazaro et al. 2023

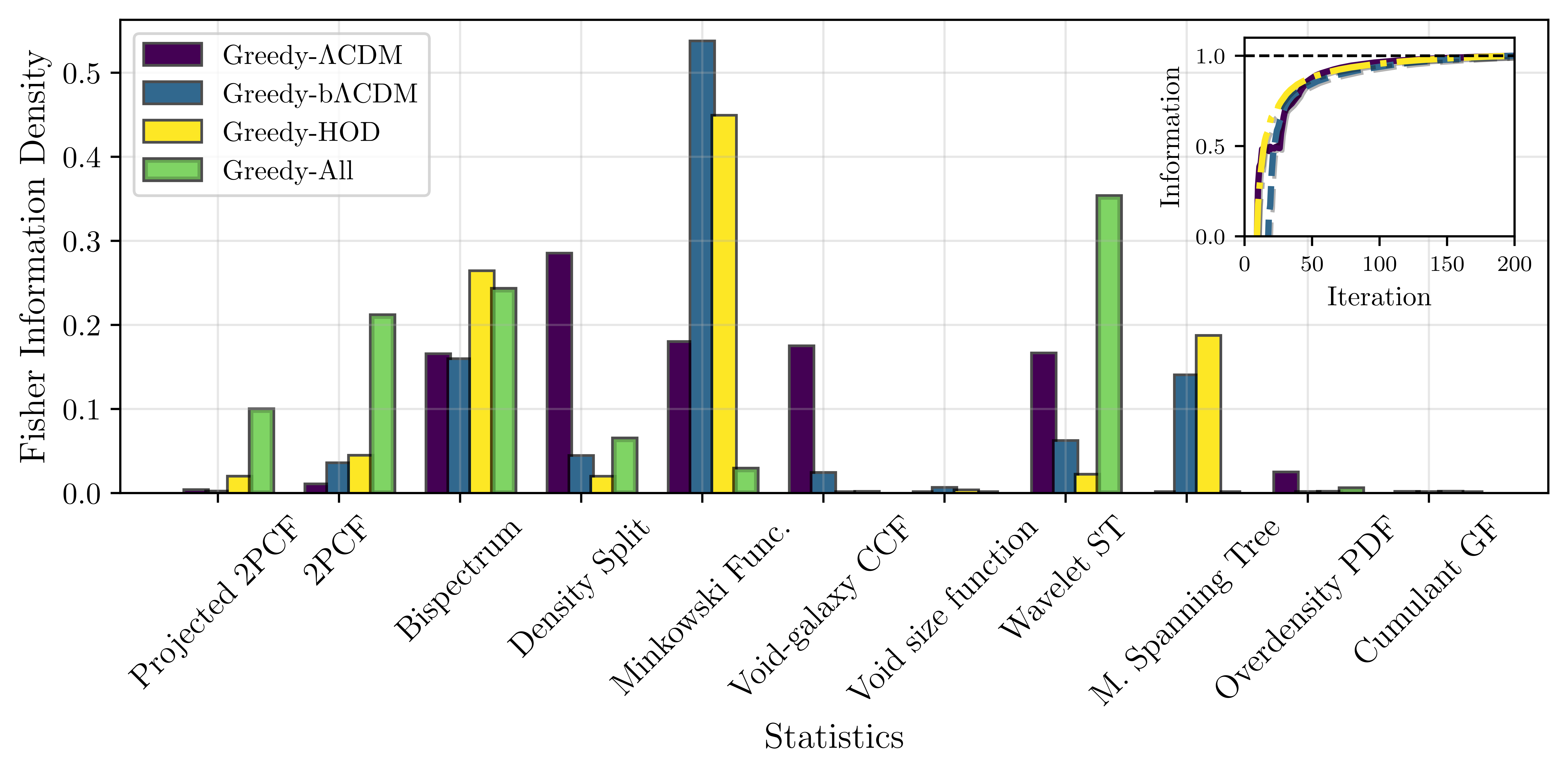
Parameter inference

Explicit likelihood (Gaussian assumption)

Cosmological constraints

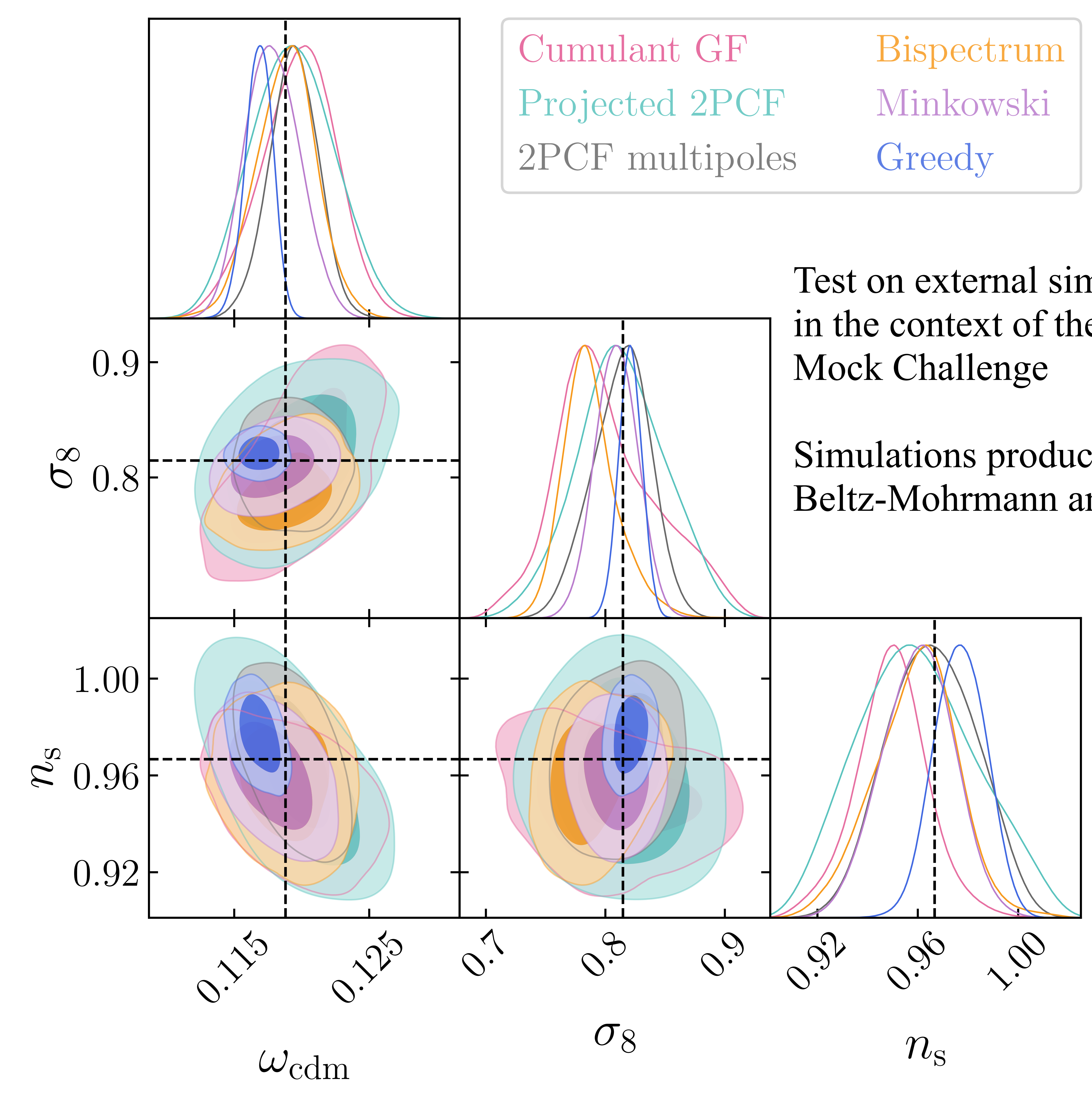
Alternative Clustering Methods within the DESI

Which set of summary statistics maximises the Fisher information?
Constraint: Keep a data vector of length 200 for mock-based covariance matrix



DESI ACM pipeline, *in prep*

Alternative Clustering Methods within the DESI



Test on external simulation (Diffsky) in the context of the DESI Emulator Mock Challenge

Simulations produced led by Gillian Beltz-Mohrmann and Andrew Hearins

DESI ACM pipeline, *in prep*

Dark matter halo catalogues (Abacus N-body simulations[1])

[1] Maksimova et al. 2021

Galaxy mock catalogues AbacusHOD[2]

[2] Yuan et al. 2022

Summary statistics

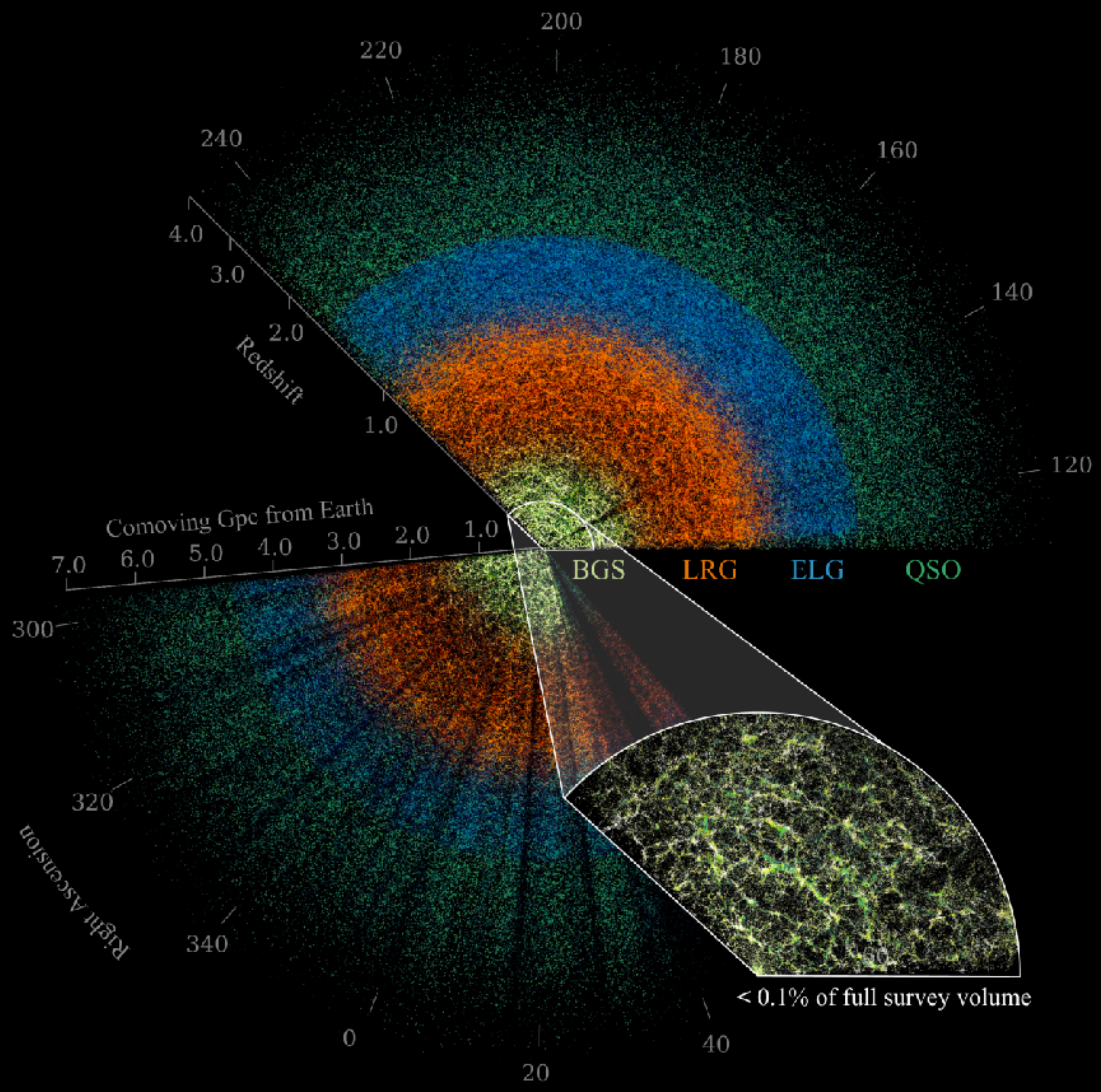
Simulation-based model (SUNBIRD[3])

[3] Cuesta-Lazaro et al. 2023

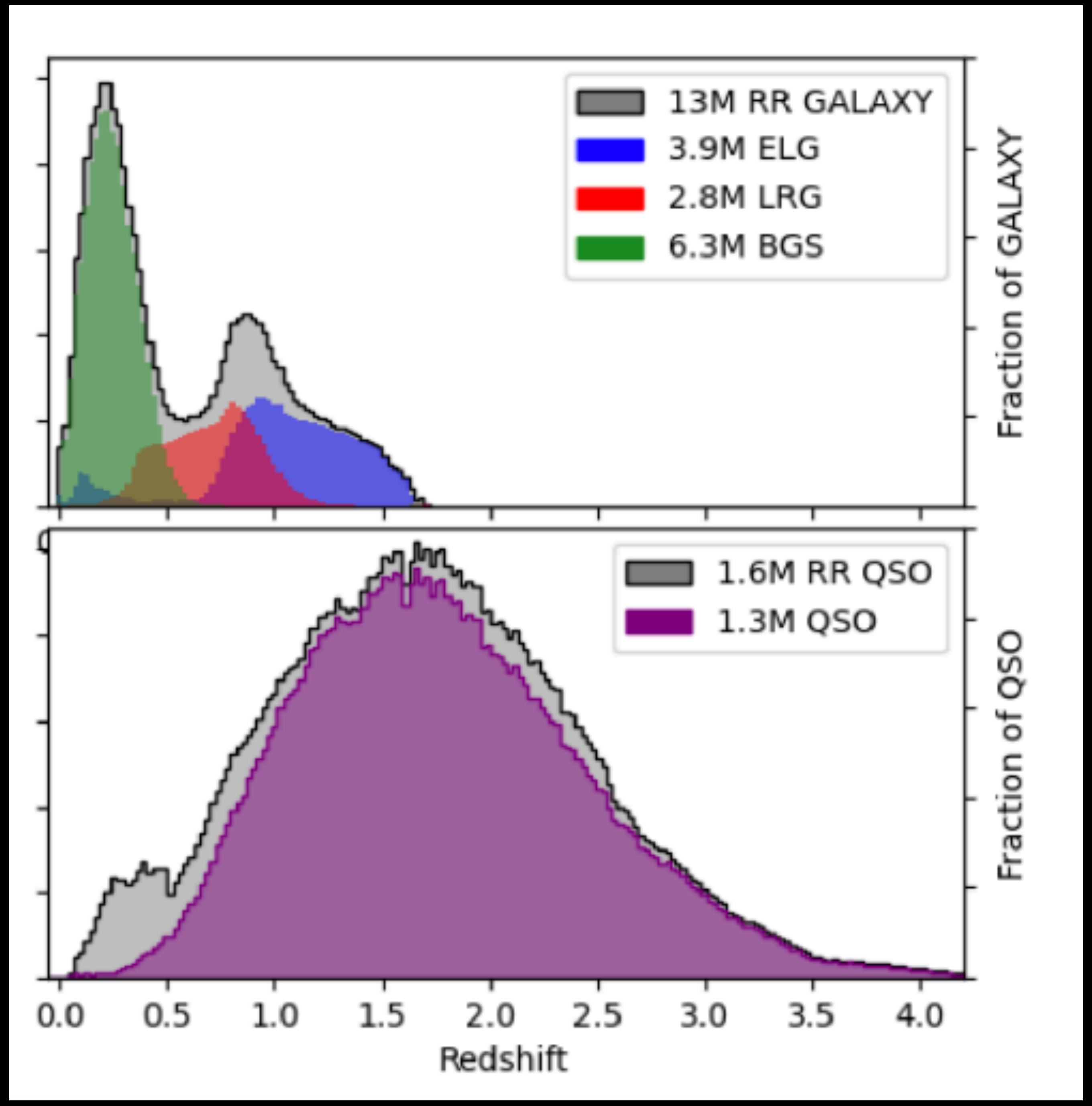
Parameter inference

Cosmological constraints

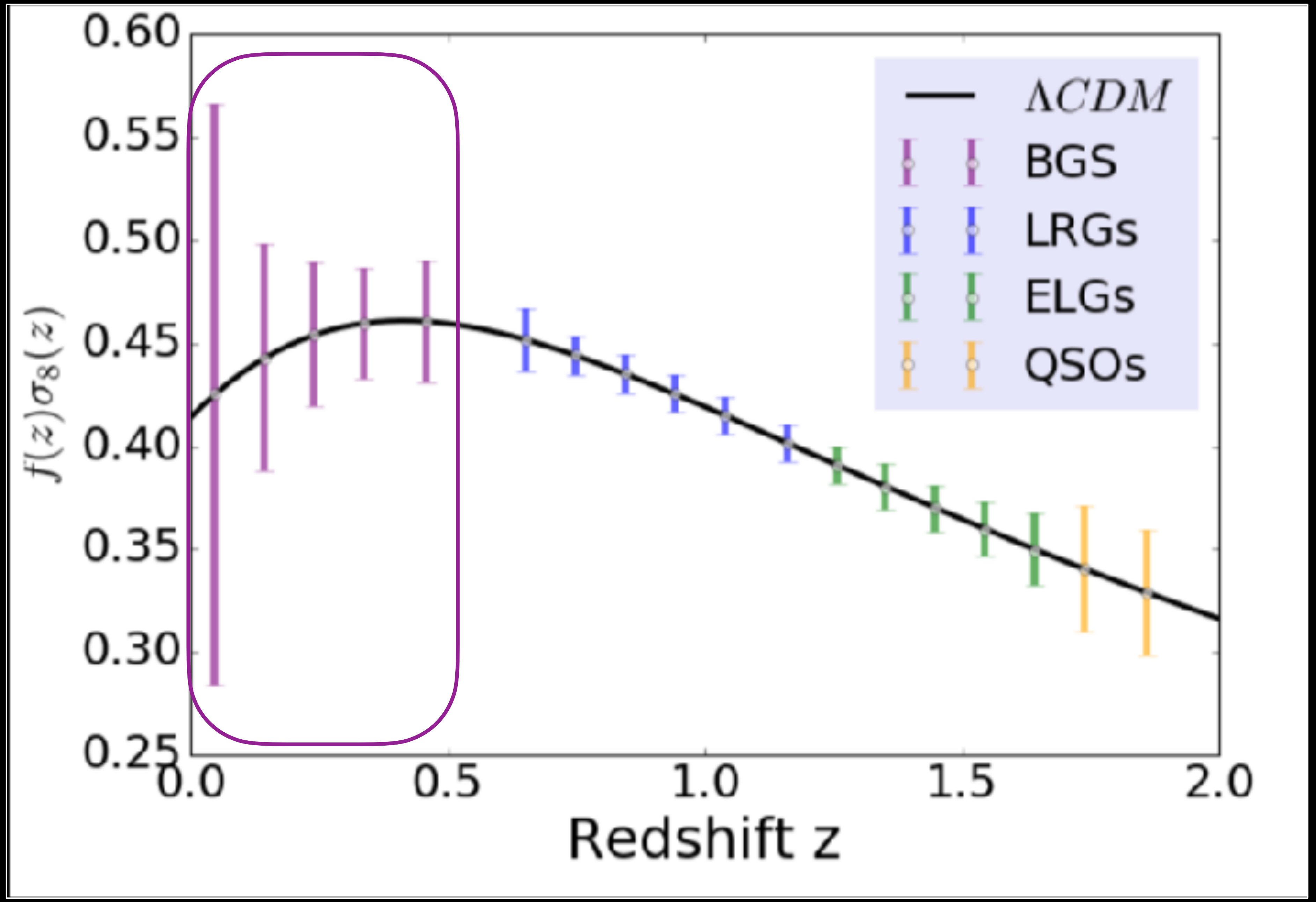
Special effort for the DESI BGS



DESI DR1, DESI collaboration 2025



DESI DR1, DESI collaboration 2025



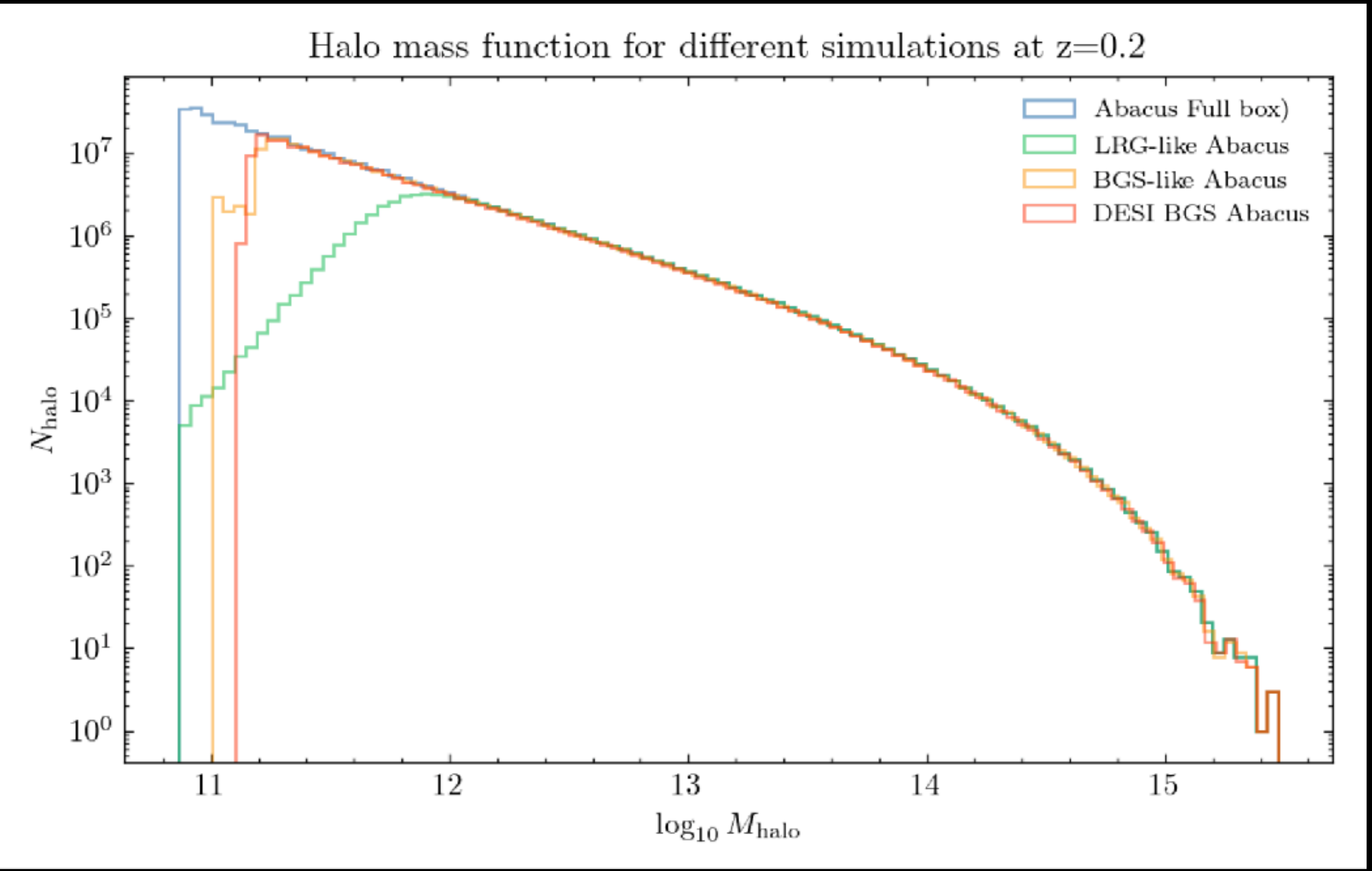
DESI collaboration 2016a

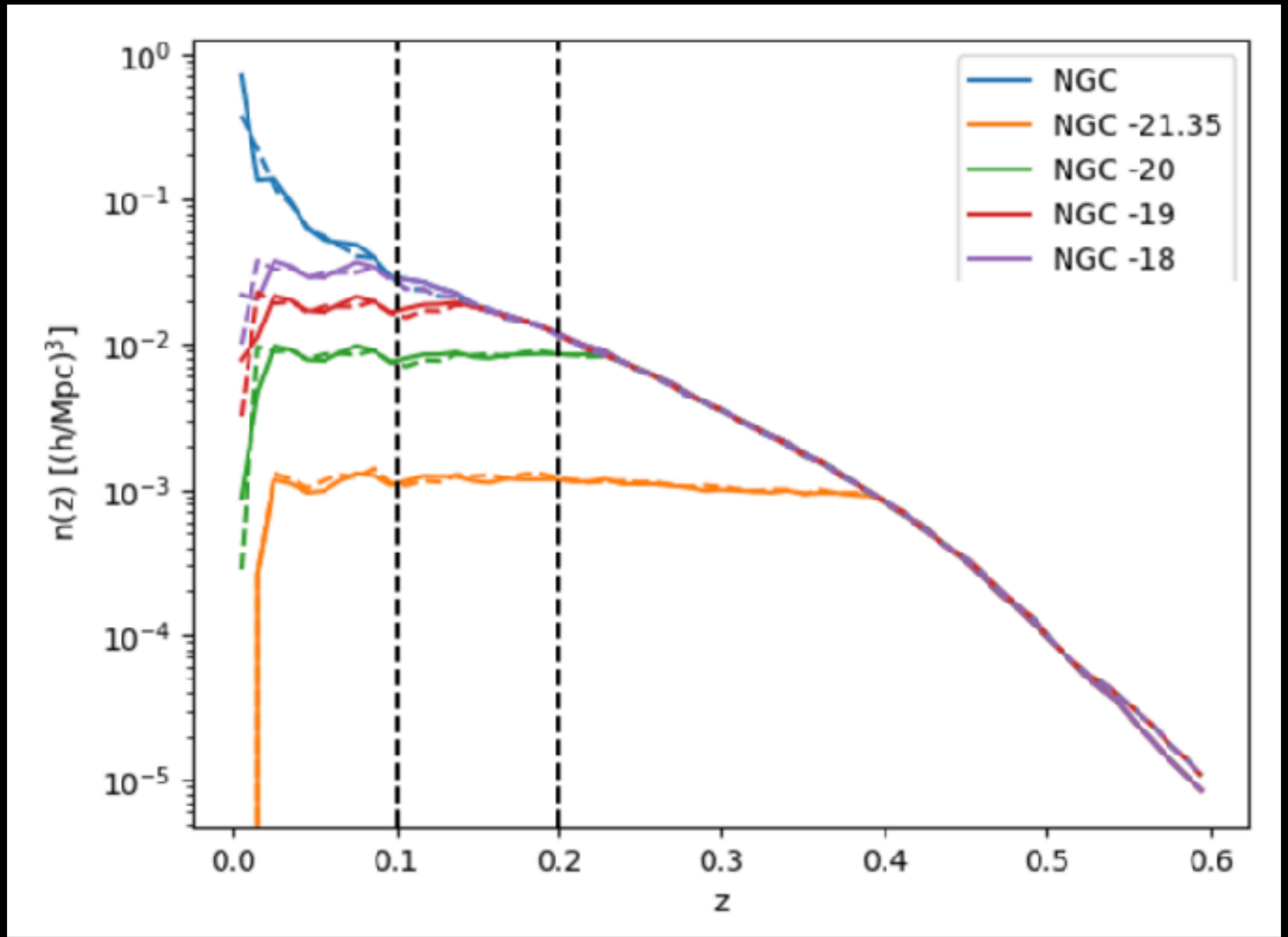
Alternative Clustering Methods with the DESI BGS

—> Specific needs for the **Bright Galaxy Survey**

Dark matter halo catalogues
(Abacus N-body simulations[1])

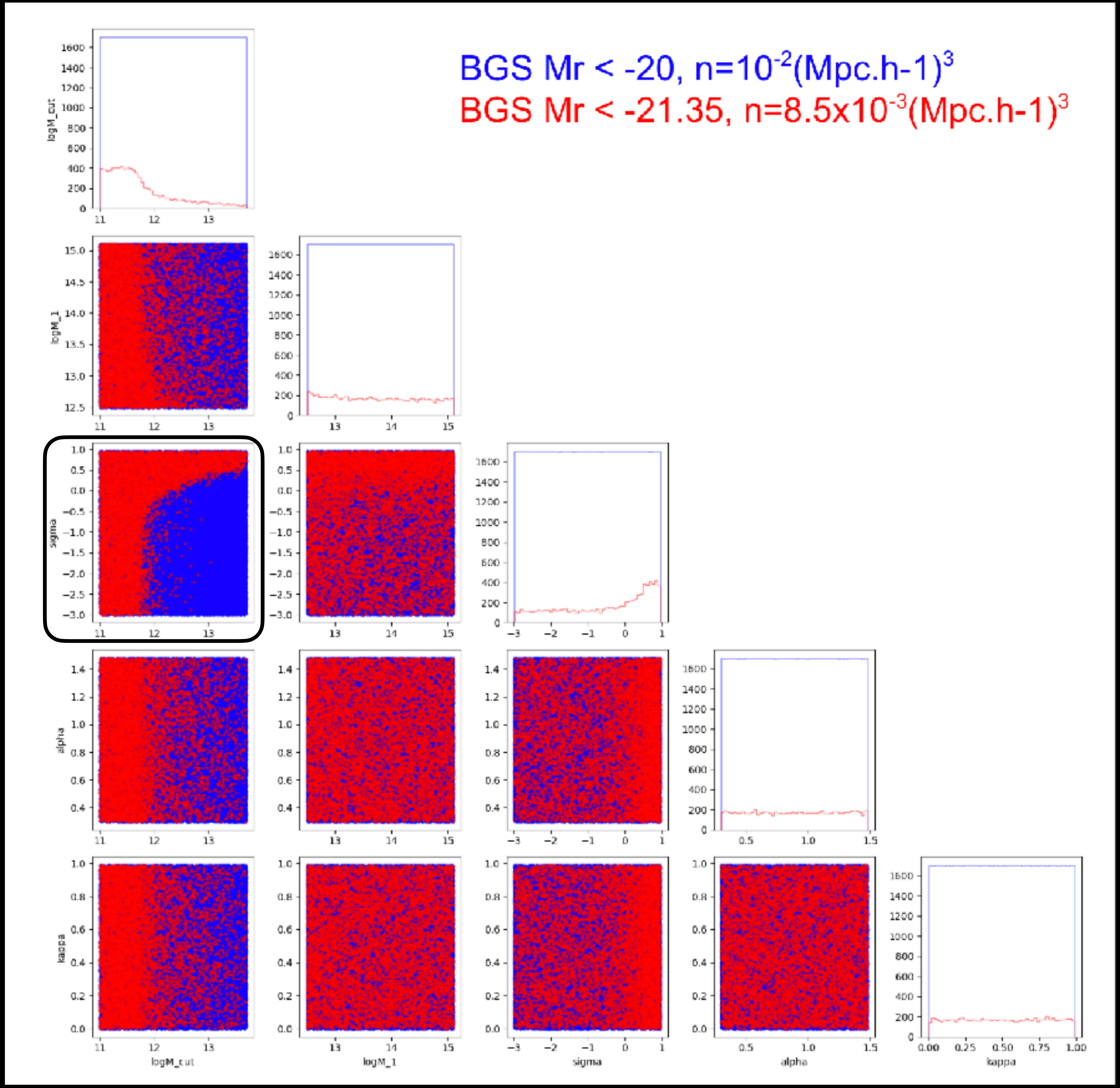
[1] Maksimova et al. 2021





Simon Bouchard (2024-2027)
Optimal summary statistics for BGS

Alternative Clustering Methods with the DESI BGS



Bouchard, Zarrouk et al. *in prep*

Dark matter halo catalogues
(Abacus N-body simulations[1])

[1] Maksimova et al. 2021

Galaxy mock catalogues
(AbacusHOD[2])

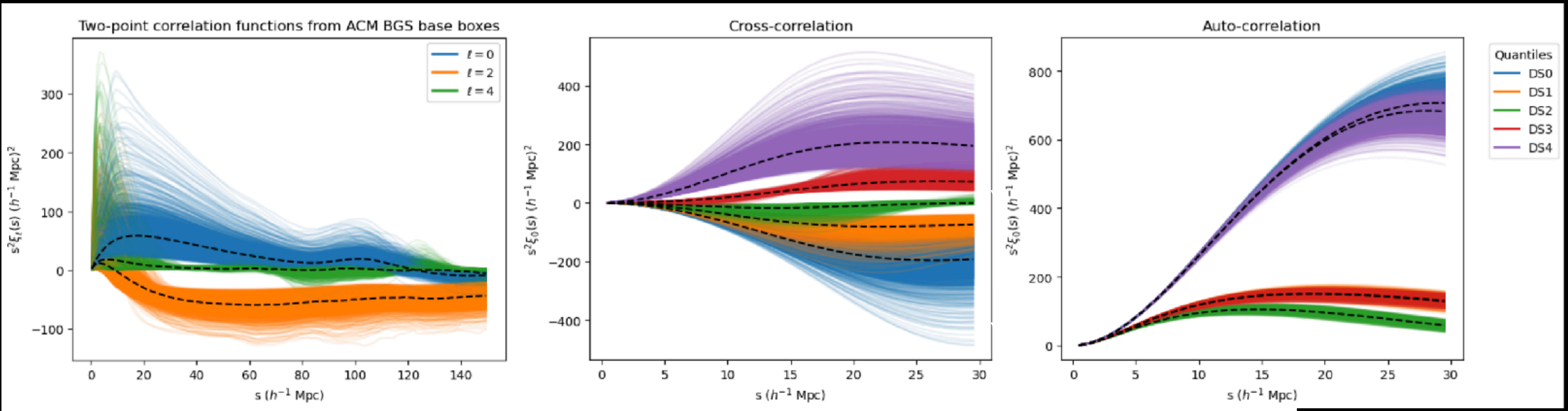
[2] Yuan et al. 2022

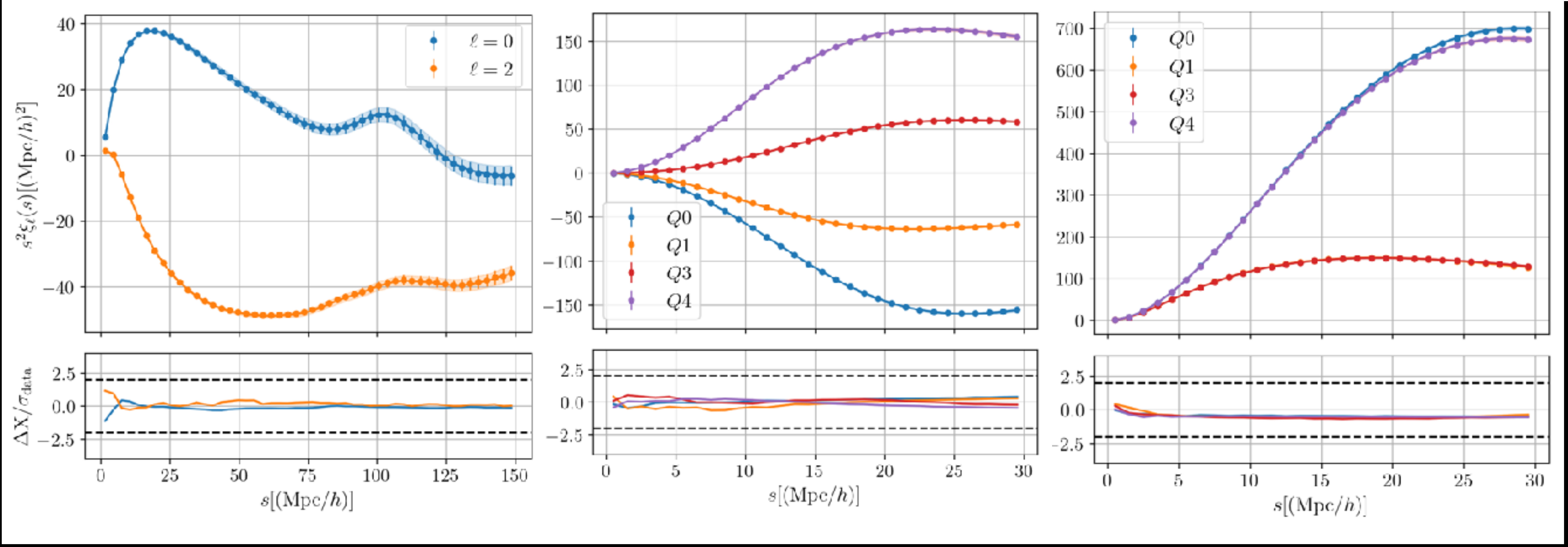
BGS storage
- dark matter catalogues: 8.5 Tb
- galaxy catalogues: ~50 Tb —> not saved, analysis on-fly

BGS computational resources
- 800 CPU hours for producing the dark matter halos catalogues
- from few hundreds to few thousands GPU hours for generating the training set and computing the statistics

Simon Bouchard (2024-2027)
Optimal summary statistics for BGS

Alternative Clustering Methods with the DESI BGS



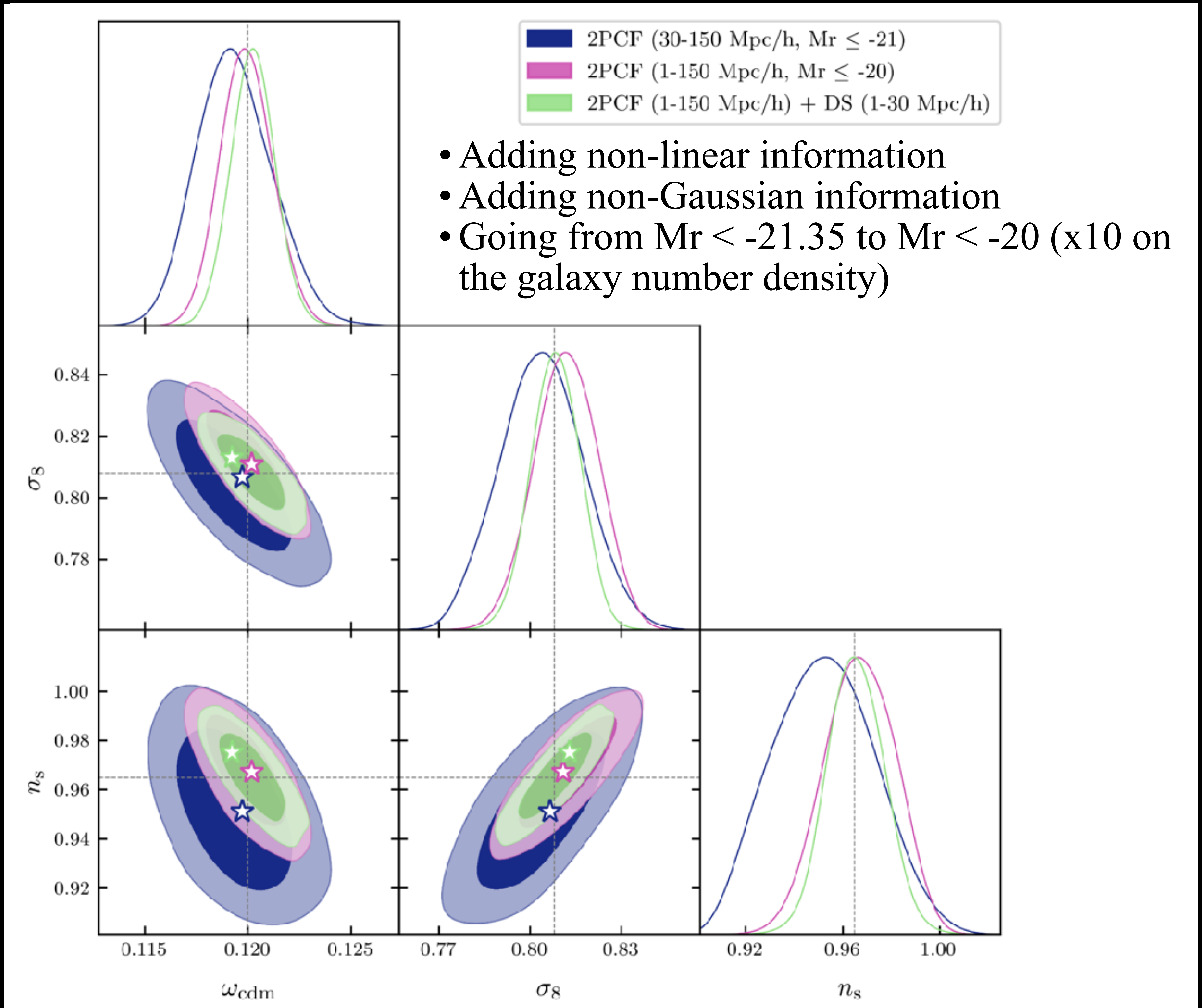


Our training set (from the Abacus suite of simulations) —> 85 cosmologies, 100 HODs per cosmology

Simon Bouchard (2024-2027)
Optimal summary statistics for BGS

Bouchard, Zarrouk et al. *in prep*

Alternative Clustering Methods with the DESI BGS



- Adding non-linear information
- Adding non-Gaussian information
- Going from Mr < -21.35 to Mr < -20 (x10 on the galaxy number density)

Bouchard, Zarrouk et al. *in prep*

Dark matter halo catalogues (Abacus N-body simulations[1])

[1] Maksimova et al. 2021

↓

Galaxy mock catalogues AbacusHOD[2]

[2] Yuan et al. 2022

↓

Summary statistics

↓

Simulation-based model (SUNBIRD[3])

[3] Cuesta-Lazaro et al. 2023

↓

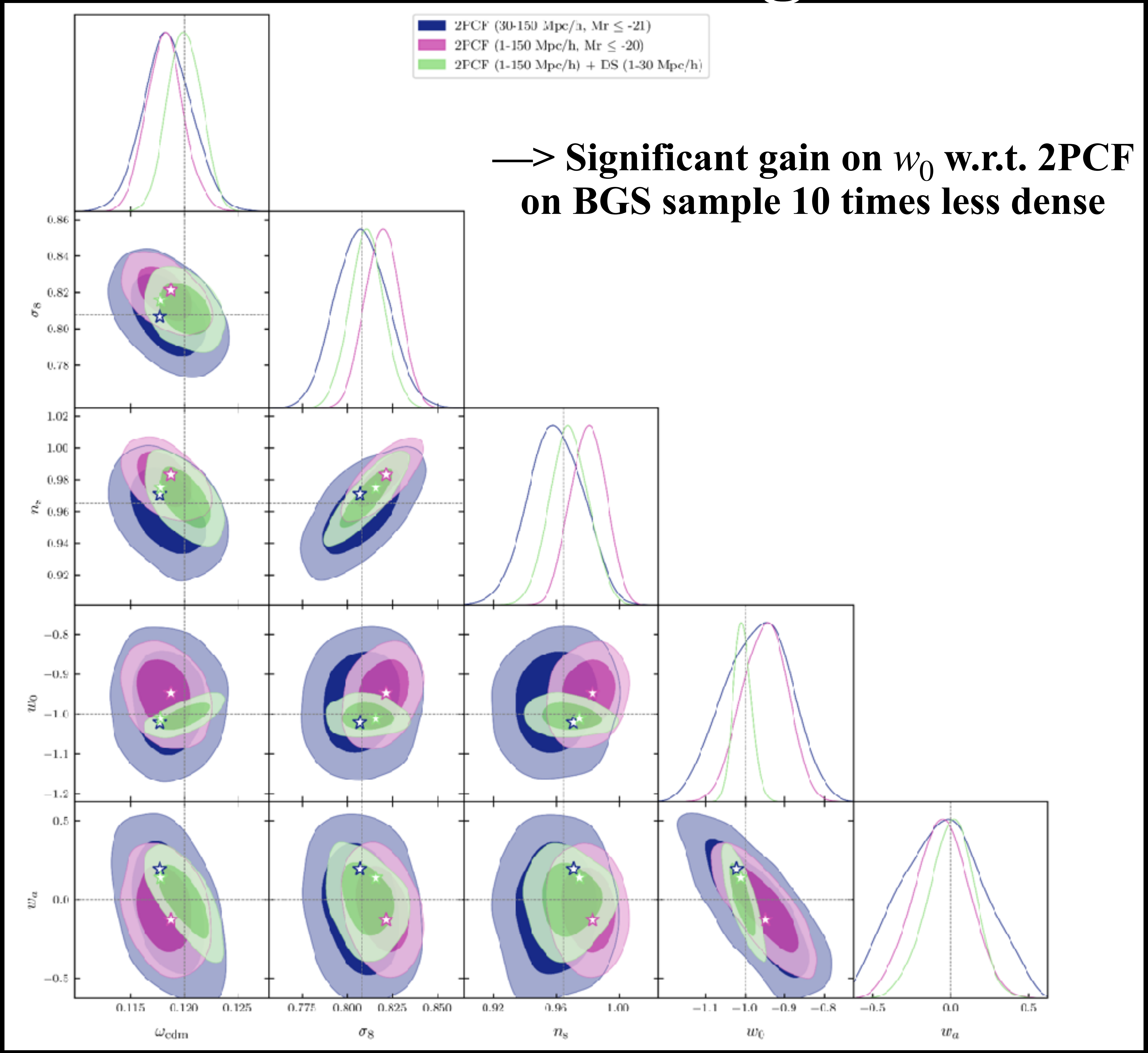
Parameter inference

↓

Cosmological constraints ΛCDM

Simon Bouchard (2024-2027)
Optimal summary statistics for BGS

Alternative Clustering Methods with the DESI BGS



—> Significant gain on w_0 w.r.t. 2PCF on BGS sample 10 times less dense

Bouchard, Zarrouk et al. *in prep*

Dark matter halo catalogues (Abacus N-body simulations[1])

[1] Maksimova et al. 2021

Galaxy mock catalogues AbacusHOD[2]

[2] Yuan et al. 2022

Summary statistics

Simulation-based model (SUNBIRD[3])

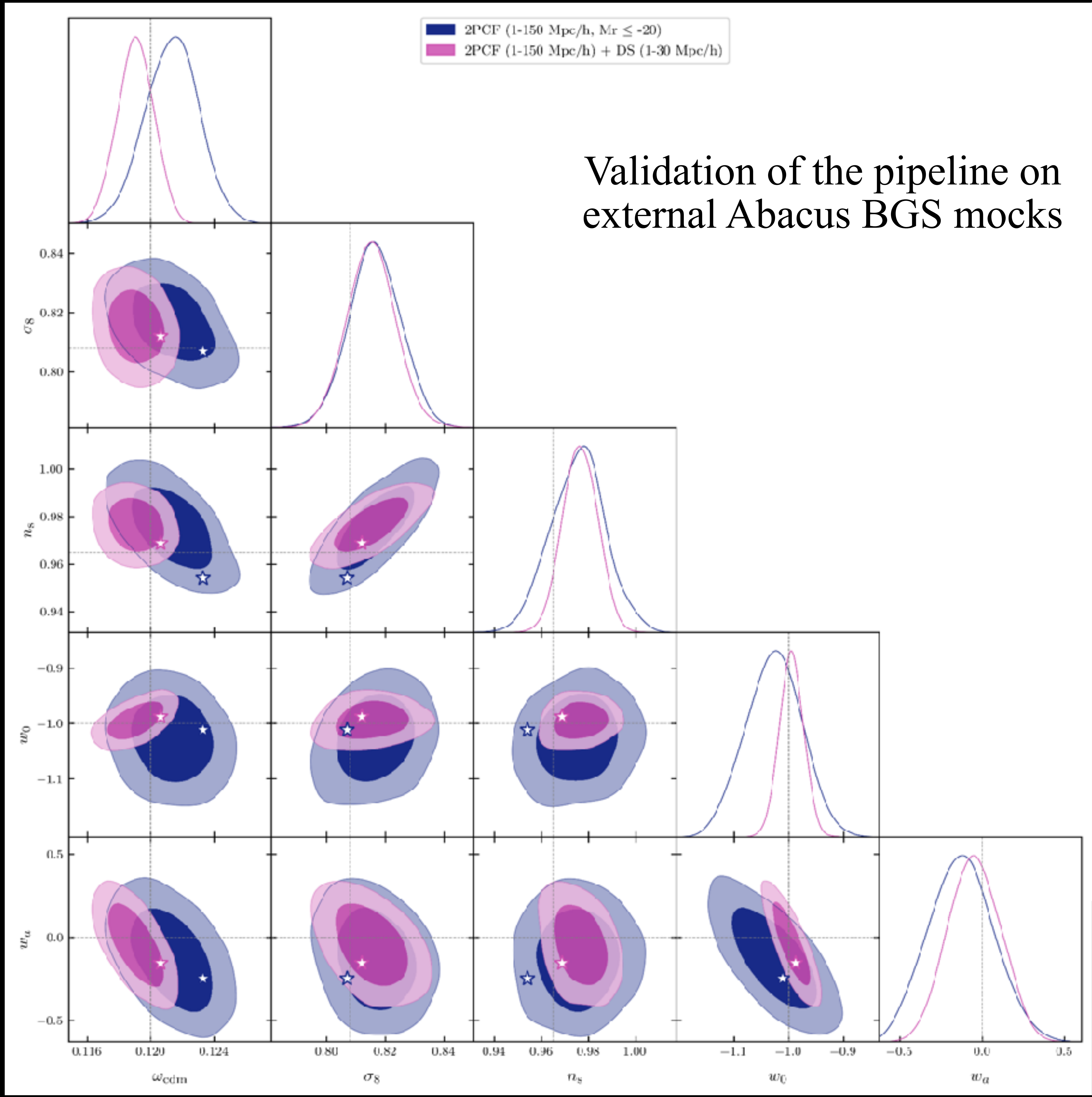
[3] Cuesta-Lazaro et al. 2023

Parameter inference

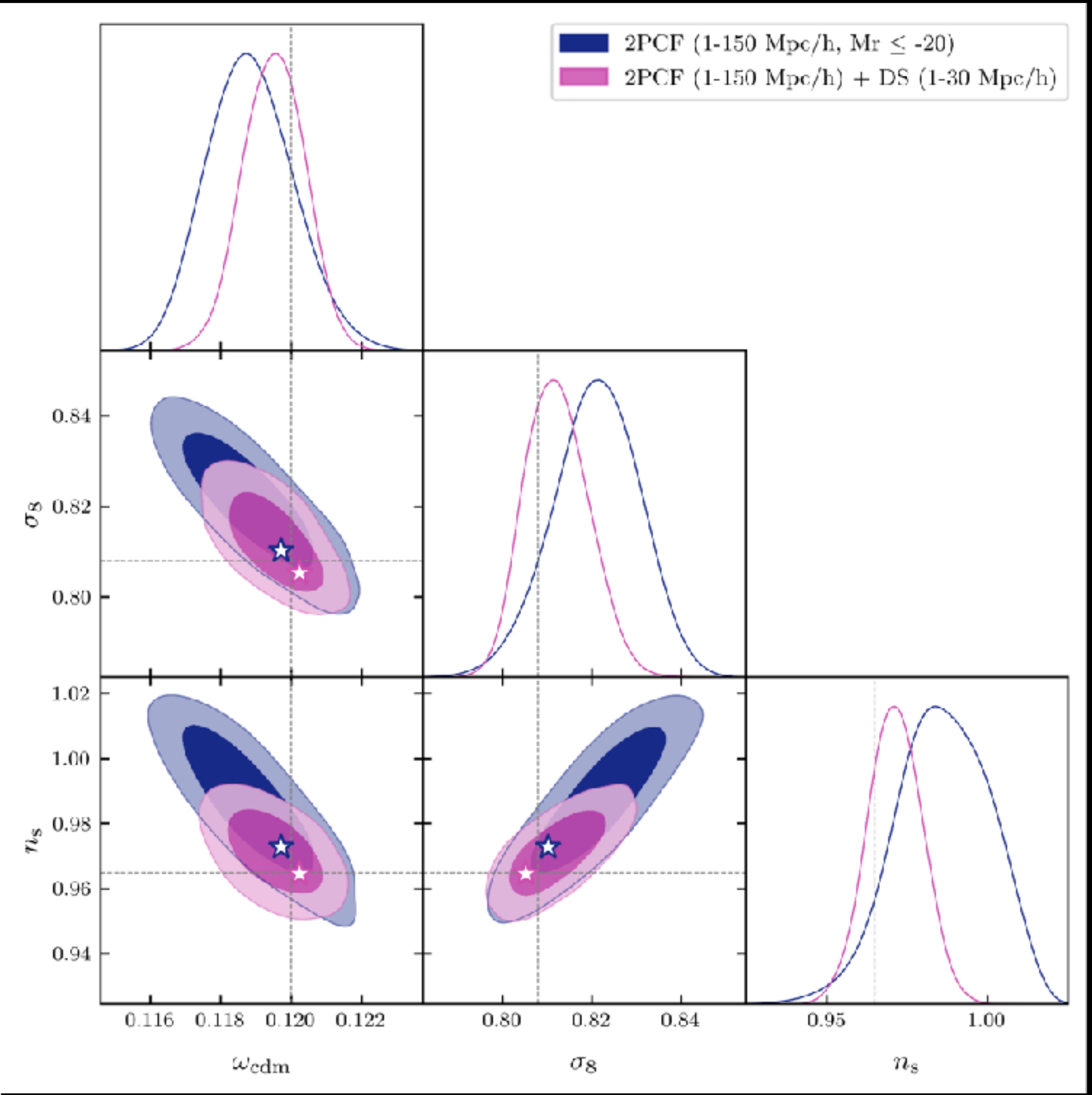
Cosmological constraints $w_o w_a$CDM

Simon Bouchard (2024-2027)
Optimal summary statistics for BGS

Alternative Clustering Methods with the DESI BGS



Validation of the pipeline on external Abacus BGS mocks



Bouchard, Zarrouk et al. *in prep*

Simon Bouchard (2024-2027)
Optimal summary statistics for BGS

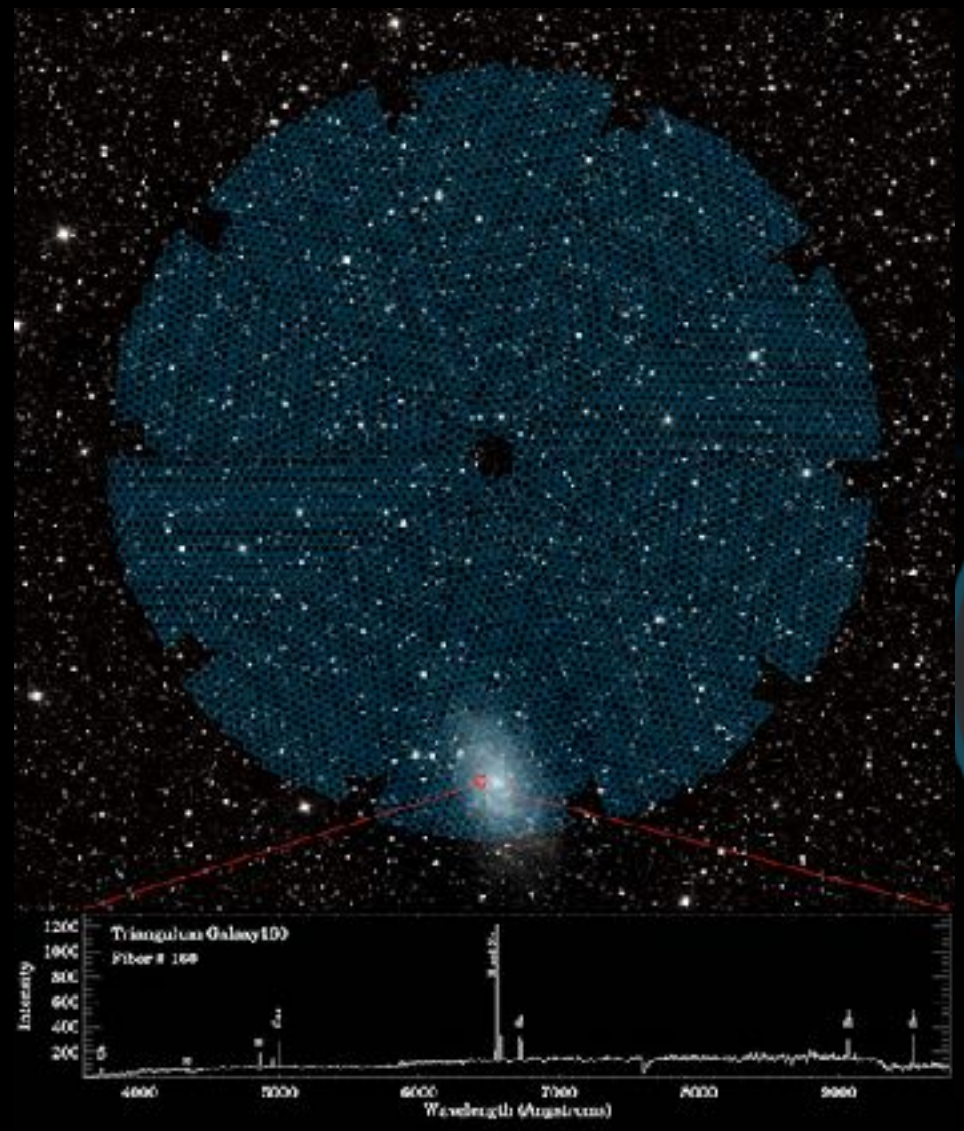
Alternative Clustering Methods within DESI

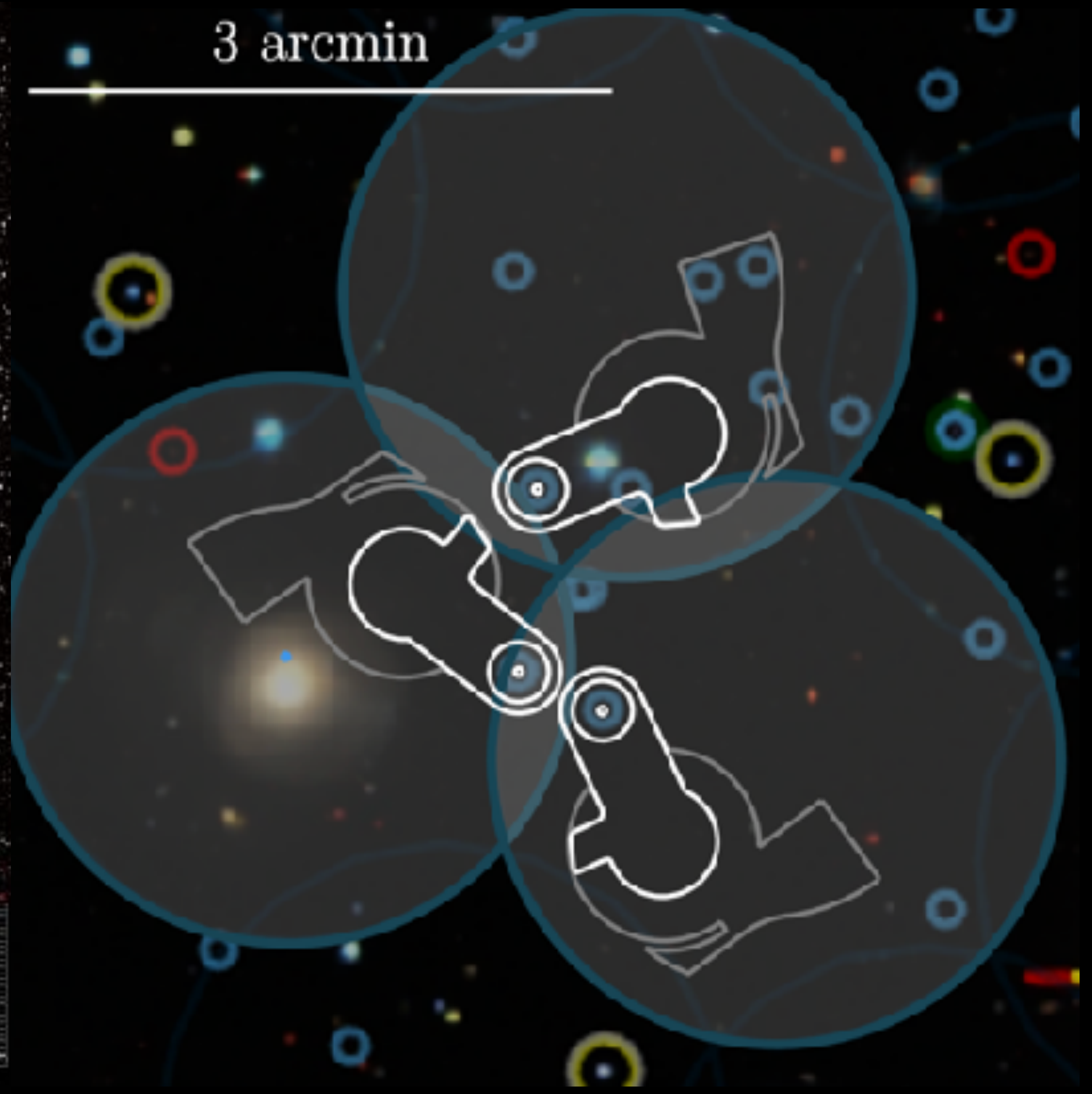
Where are we today?

- ACM working (cubic-box) pipeline for 12 summary statistics
  —> two associated papers: ACM pipeline, emulator mock challenge
- BGS methodology paper in preparation
- Standardisation and robustification of the entire ACM pipeline
  (unit tests, functional tests, automated integration)

Towards DESI DR2 application

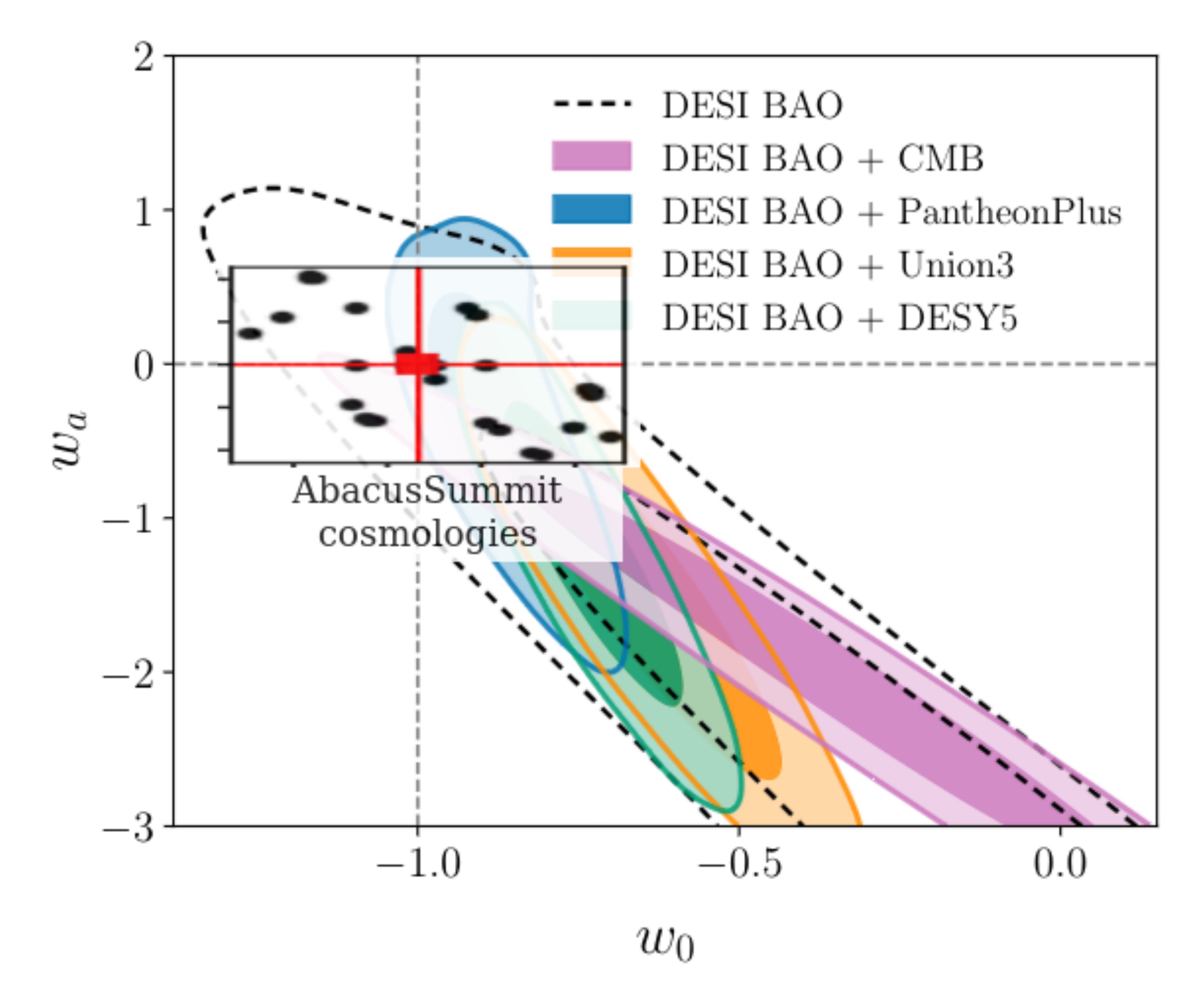
- Include DESI survey geometry
- Include DESI fibre-assignment





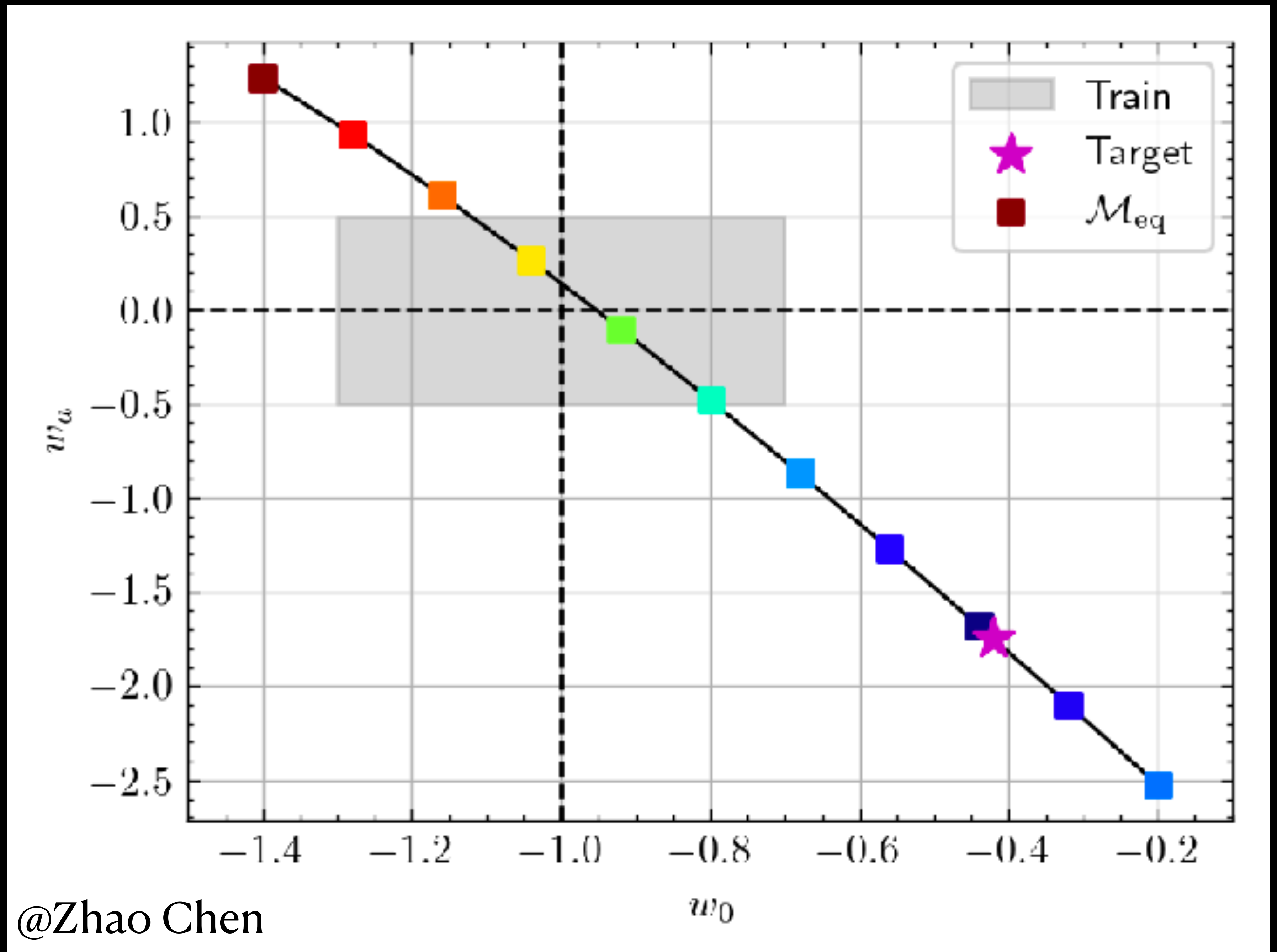
Alternative Clustering Methods within DESI

Constraining $w_0 w_a$CDM



—> Limited $w_o - w_a$ parameter space due to the input suite of simulations (AbacusSummit range)

- One avenue to extend SUNBIRD $w_o - w_a$ parameter space: Spectral equivalence method (Zhao Chen & Yu Yu 2025)



- Another long-term avenue (for DESI DR3): new suite of Abacus simulations should be available next year

Part 2: Field-level inference with BORG/Manticore

Manticore-DESI with Stuart McAlpine, Guilhem Lavaux, Jens Jasche, Pontus Holma
BORG-Peculiar Velocity with Deaglan Bartlett, Solène Degardin, Guilhem Lavaux, Mahmoud Osman, Richard Stiskalek

BORG: Bayesian Origin Reconstruction from Galaxies
Jasche & Wandelt (2013), Jasche et al. (2015), Jasche & Lavaux (2019)

Manticore
McAlpine et al. (2025), McAlpine et al. (2026)

How to extract cosmological information?

Sensitivity to cosmology

Inflation and primordial features — Dark matter and neutrinos — Dark energy

redshift

Initial Gaussian and linear density field — *Today **non-Gaussian** and **non-linear** density field*

← more robust — more optimal →

Two-point statistics (2pt)
- Most mature and robust technique
- Analytical model available up to quasi-linear scales (k_{max}=0.2 h/Mpc)
- Computationally cheap
- Only Gaussian and linear information

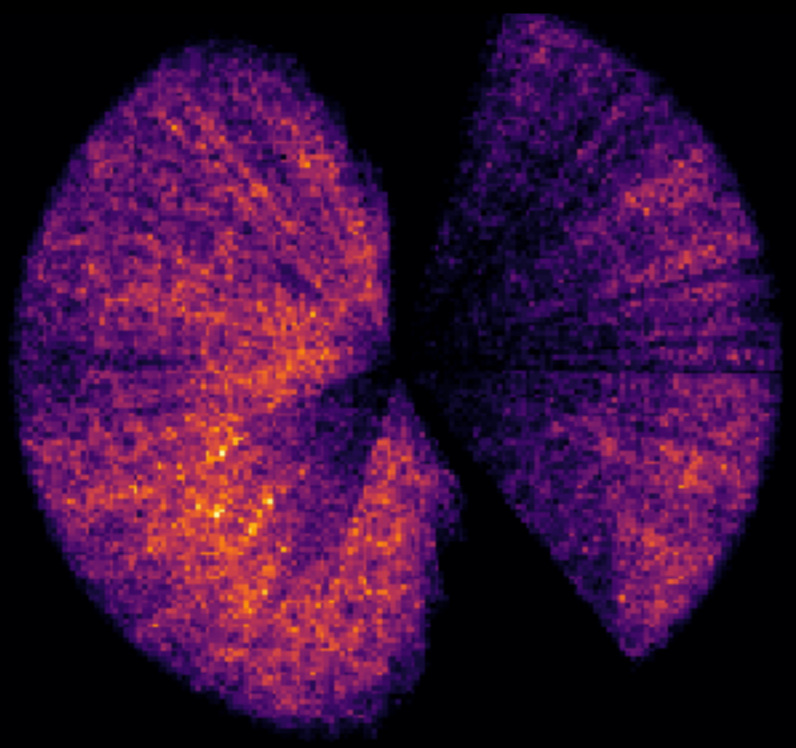
Higher-order statistics (HOS)
- Capture some non-Gaussian information
- Mostly simulation-based model up to k_{max}=0.8 h/Mpc
- Computationally cheaper than FLI but still requires thousands of cosmological simulations
- Factor 2 improvement w.r.t. 2pt*

*Euclid collaboration V 2023

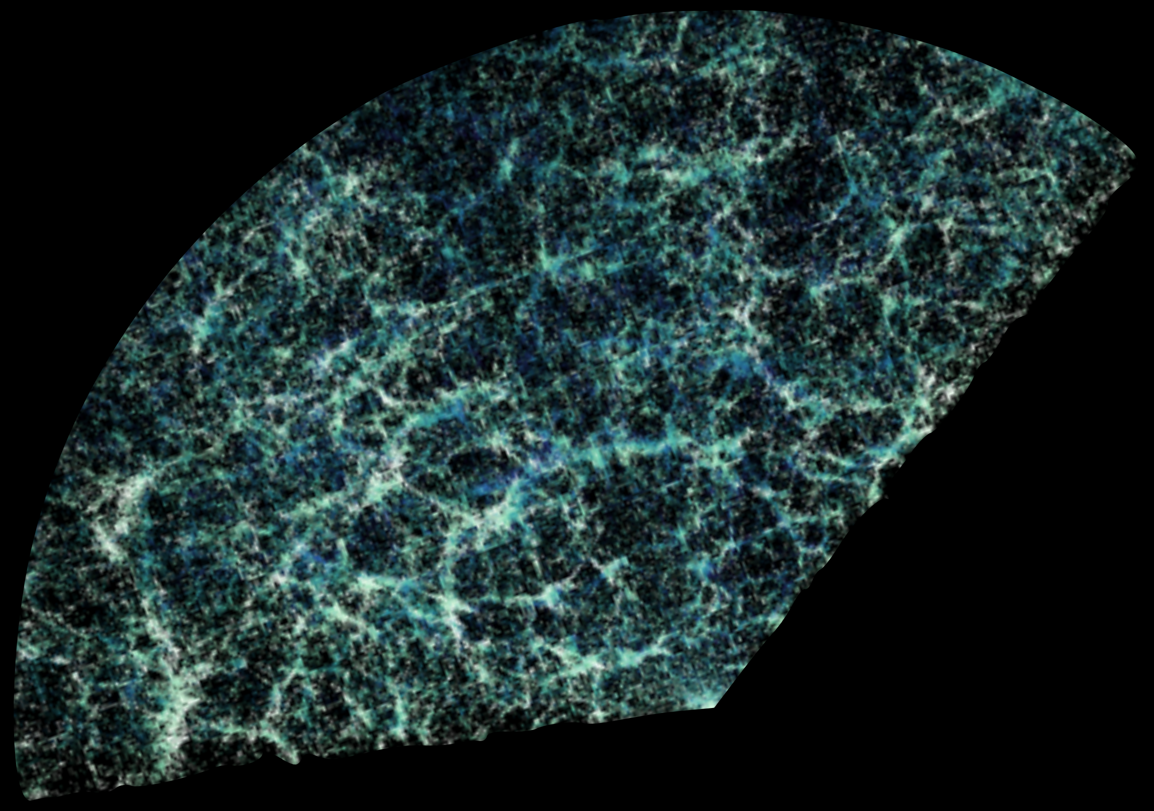
Field-level inference (FLI)
- Retrieve the maximal information & deliver constrained simulations
- Grid resolution up to k_{max}=0.8 h/Mpc
- Computationally very expensive
- Factor 2-5 improvement w.r.t. HOS**?

** Nguyen et al. 2025, Porqueres et al. 2023

Bayesian field-level inference: how does it work?



Voxel-by-voxel comparison



Cosmic variance-aligned simulated DESI DR1 Universe

Real DESI DR1 Universe

Digital twins of the Universe (cosmic variance-aligned / constrained simulations) offer a unique way of confronting theory against data

- Detailed probabilistic maps of the matter distribution and its history
- Access all latent variables (velocity field, tidal field, …)
- Posterior predictive tests and cross-validation checks

Three main ingredients:

1) The forward model (or the simulator)
2) The (explicit) likelihood
3) The sampler

Three main challenges:

- Model-mispecification
- Scalability with larger and denser datasets
- Multi-million dimensional parameter space

Bayesian field-level inference

Initial phases

World model

$$\pi(x, \Omega, M \,|\, d) = \pi(M \,|\, d)\pi(\Omega \,|\, d, M)\pi(x, d, M, \Omega)$$

Model parameters

$\pi(M \,|\, d)$

Model evidence
(here ΛCDM)

$\pi(\Omega \,|\, d, M)$

Parameter estimation
$\Omega_m, \sigma_8, n_s, h, \ldots$ + astro + nuisance

Ensemble view: laws and parameters, statistically consistent

$\pi(x \,|\, d, M, \Omega)$

Specific realisation
Initial phases / modes

Realization view: our Universe

Output of the re-simulations at the inferred initial conditions: digital twins of the Universe

The ultimate goal is to access $\pi(x, \Omega \,|\, M, d)$.
But it requires:

1. Fast gradient-based sampler (HMC) for the very high dimensional parameter space: $\sim 10^{8-9}$ parameters (mesh resolution)
2. Requires an accurate (differentiable) model that links the initial Gaussian random field to a given observable
3. Precise knowledge of the selection functions and mitigation of systematic effects at the field level

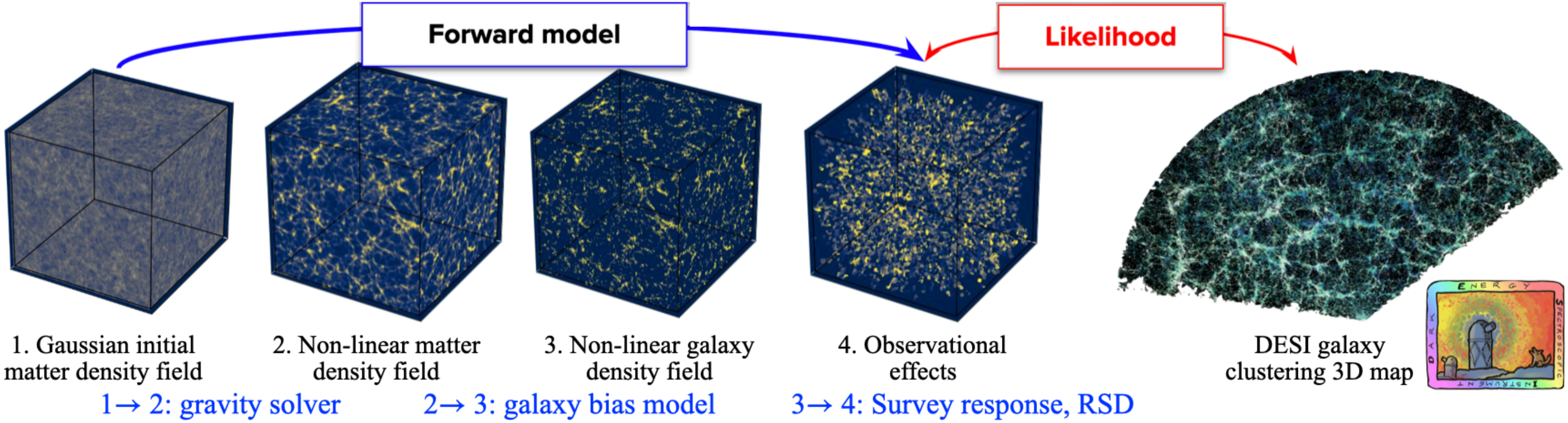
How? One way to do FLI is with BORG

Jasche & Wandelt (2013), Jasche et al. (2015), Jasche & Lavaux (2019)

Field-level inference with galaxy clustering

$\pi(x, \Omega_{nuisance} \,|\, \Omega_{cosmo}, M, d)$

Poisson likelihood for number counts



1—> 2: Gravity solver

Default: COLA, but faster options already exist.

2 —> 3: Galaxy bias model

Default: Sigmoid Truncated Double Power Law (STDPL) but non-local relation and stochasticity

3 —> 4: Selection function

Field-level inference with galaxy clustering

The likelihood
—> Generalised Poisson likelihood (McAlpine et al. 2025, 2026)

$$\Phi(N|\lambda, b) = \frac{\lambda(1-b)}{N!}[\lambda(1-b)+Nb]^{N-1}\, e^{-\lambda(1-b)-Nb}$$

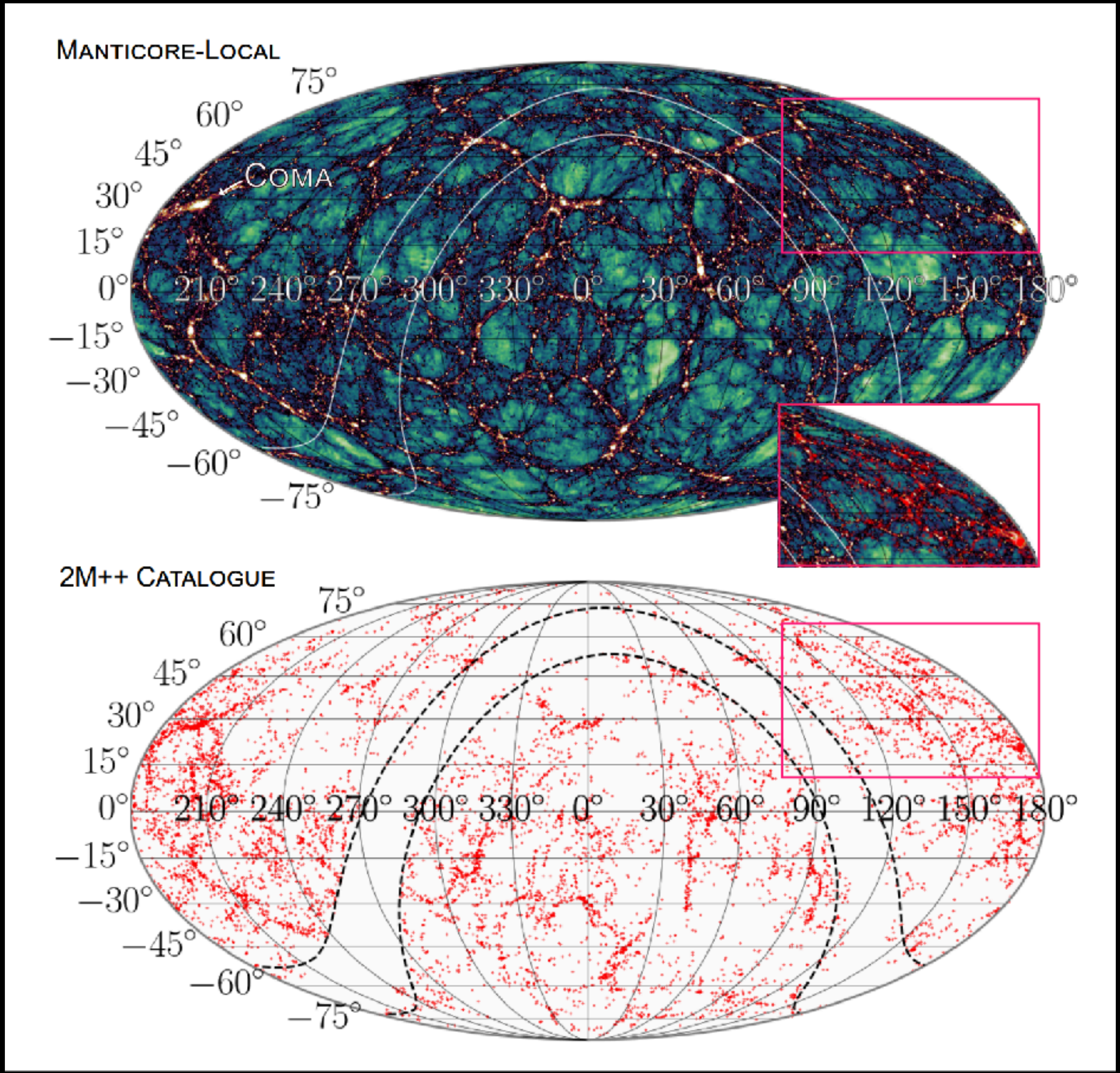
- Overdispersion: $b(\rho) = b_\alpha \rho^{b_\beta}$
- Enables $\sigma^2 > \mu$ (super-Poissonian)

Physics-informed priors (McAlpine et al. 2025, 2026)

$$H_{\text{total}} = E_{\text{var}}(x) + E_{\text{ps}}(\tilde{x}) + E_{\text{mom}}(\delta_{\text{final}})$$

- Gaussianity: $E_{\text{var}} = \frac{1}{2}\sum x_i^2$
- Isotropy: $E_{\text{ps}} \propto \sum(P(k_i)/\bar{P} - 1)^2$
- ΛCDM moments: E_{mom} constrains mean, variance, skewness

Successful proof-of-concept: Manticore-Local



McAlpine et al. 2025

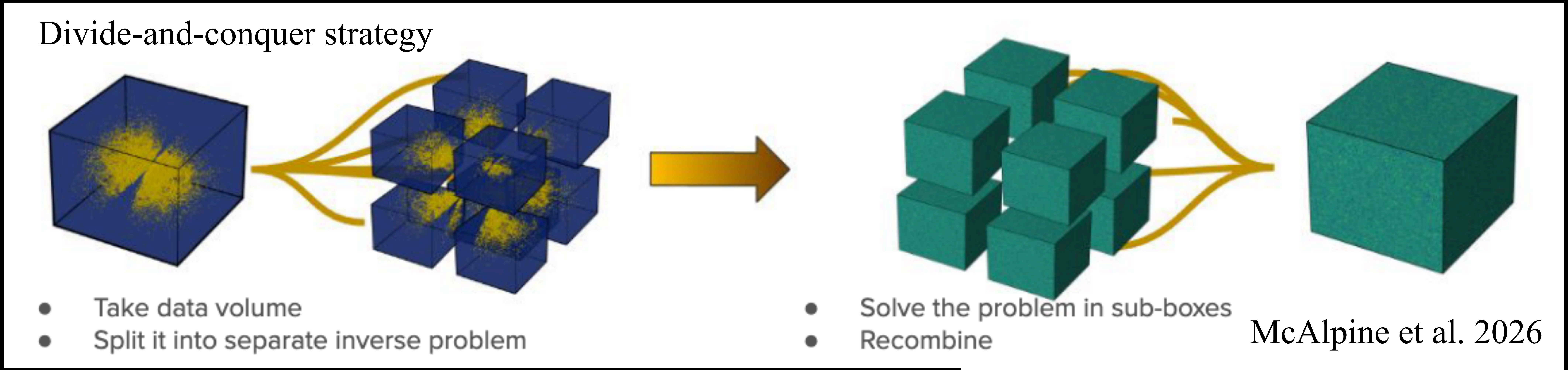
Posterior predictive tests

- ✅ Initial power spectrum (white noise field)
- ✅ Power spectrum and bispectrum of the matter field at z=0
- ✅ Halo mass function
- ✅ Velocity field
- ✅ Galaxy clusters detection and properties

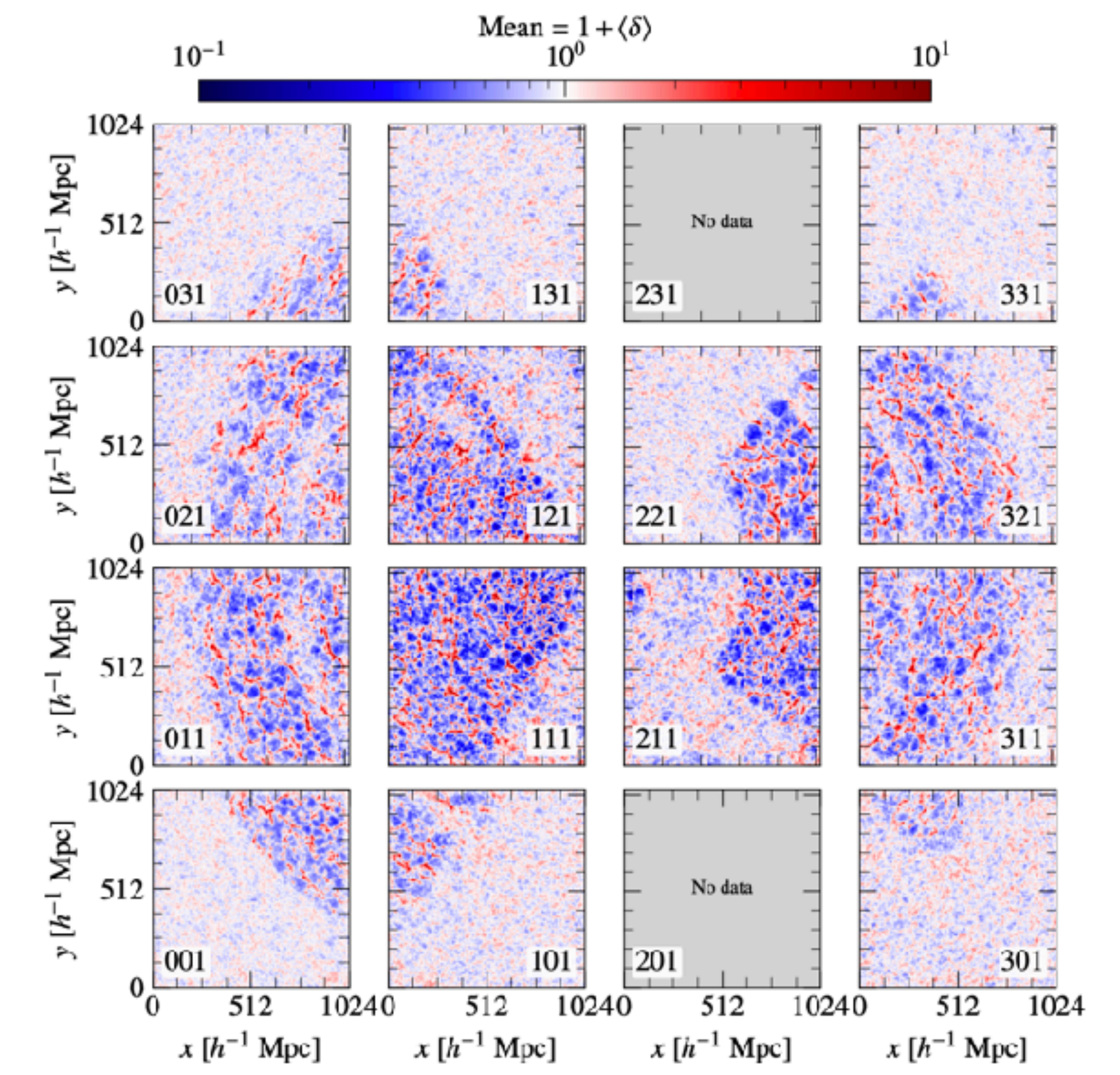
Which additional tests to showcase the limitations of the galaxy bias model?

The limitations of the generalized Poisson likelihood?

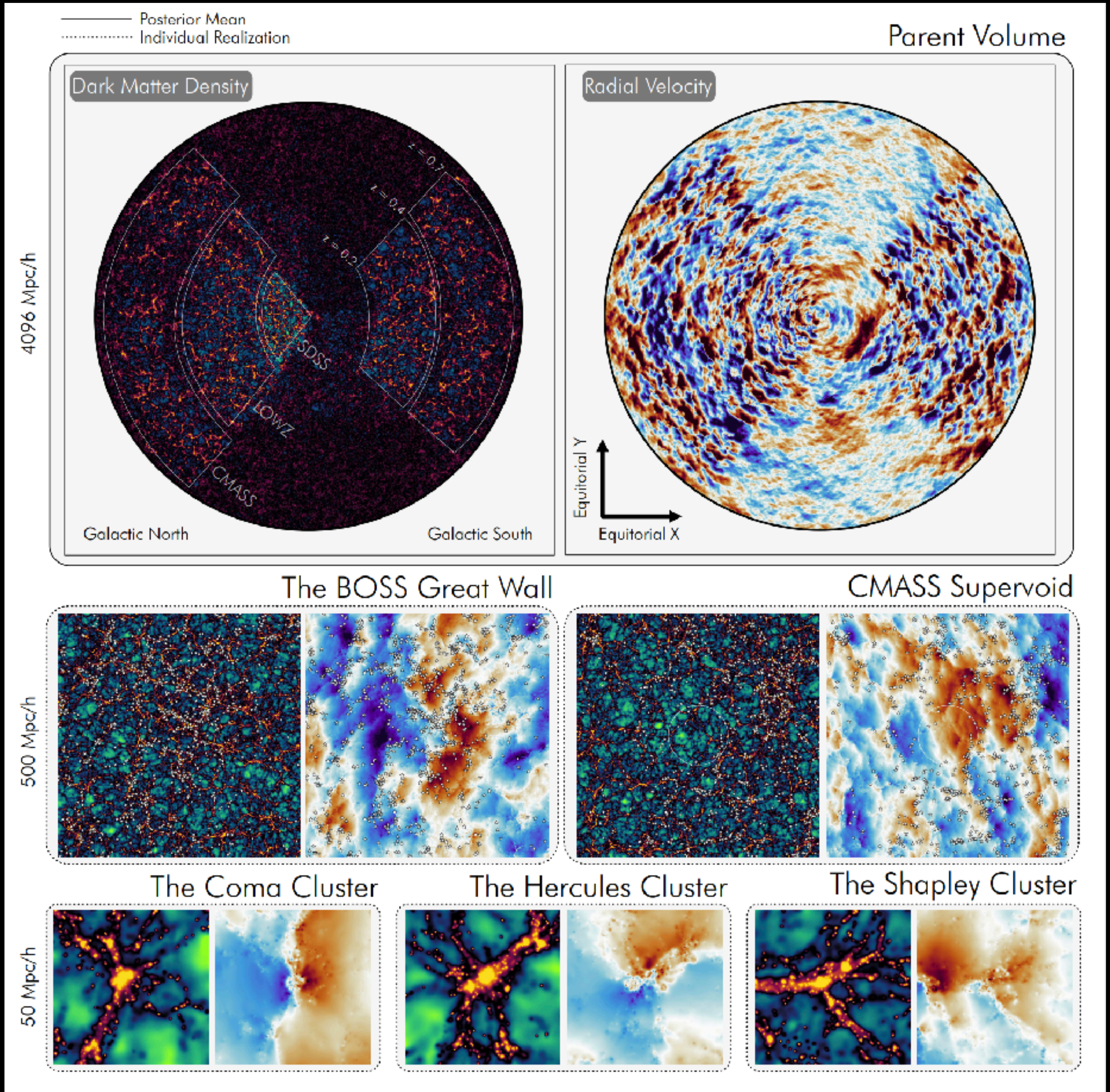
Another major challenge: scalibility



Manticore-Local	Manticore-Deep
L = 700 Mpc/h	L = 4096 Mpc/h
N = 256^3	N = 1024^3
—> 2.7 Mpc/h spatial resolution	—> 4 Mpc/h spatial resolution
	L_{subvol} = 1024 Mpc/h
	N = 256^3
	—> 4 Mpc/h spatial resolution



Successful proof-of-concept: Manticore-Deep



McAlpine et al. 2026

Posterior predictive tests

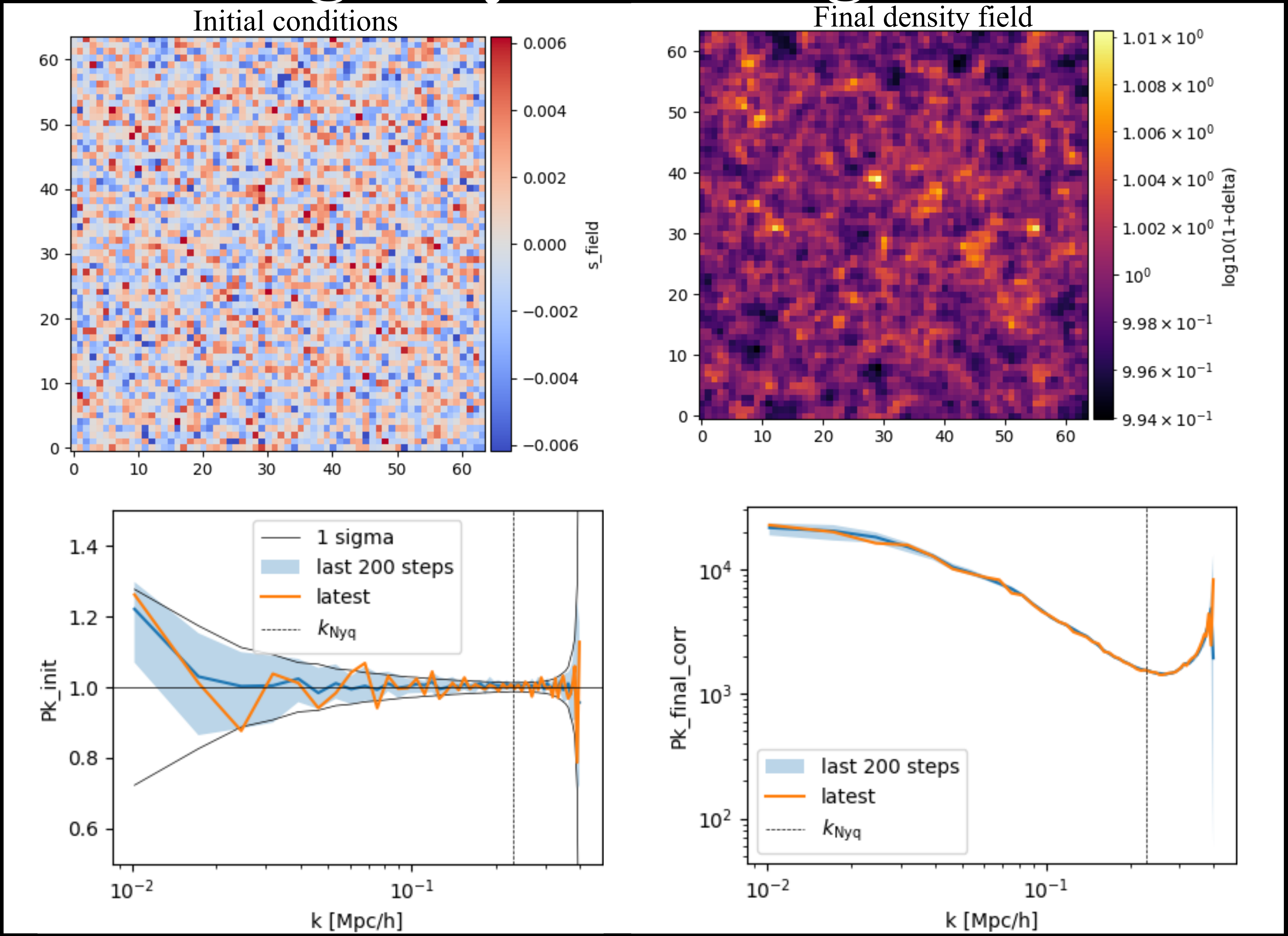
- ✓ Initial power spectrum (white noise field)
- ✓ Power spectrum and bispectrum of the matter field at z=0
- ✓ Halo mass function
- ✓ Convergence field and cross-correlation with Planck PR3 CMB map at 7.4σ
- ✓ Velocity field and detection of tSZ effect at 3.5σ

Which additional tests to showcase the limitations of the galaxy bias model?

The limitations of the generalized Poisson likelihood?

Field-level inference with galaxy clustering

Preliminary tests of Manticore against DESI BGS DR1 mock data using low resolution grid



Next steps:
- Benchmarch for target resolution $N=1024^3$
- Tests/refinement of the galaxy bias model
- Account for observational systematic effects

Summary / open questions

Alternative Clustering Methods within DESI

Working pipeline for 12 summary statistics validated against non-HOD simulations

Is it enough to validate the galaxy-halo connection assumptions?

Including observational systematics: fibre-assignment

Is the method and gain robust against systematics?

Extend the emulator $w_o - w_a$ parameter space

Do we have the required suite of simulations for simulation-based model?

Field-level inference within DESI with BORG

Validated framework for inferring initial conditions
at fixed cosmology with previous generation datasets

Digital twins are powerful tools for testing the cosmological model
What about cosmological parameter constraints?

Galaxy bias model
Poisson likelihood

Analytical vs AI-driven galaxy bias model?
Is Poisson likelihood really enough for inferring initial conditions?

Including observational systematics: fibre-assignment

Is the method robust against systematics?