The Astrophysical Journal, 990:99 (43pp), 2025 September 10 https://doi.org/10.3847/1538-4357/ad781d

A Parameter-masked Mock Data Challenge for Beyond-two-point Galaxy Clustering Statistics*

The Beyond-2pt Collaboration,
Elisabeth Krause[1], Yosuke Kobayashi[1,2], Andrés N. Salcedo[1], Mikhail M. Ivanov[3], Tom Abel[4,5,6], Kazuyuki Akitsu[7], Raul E. Angulo[8,9], Giovanni Cabass[10], Sofia Contarini[11,12,13], Carolina Cuesta-Lazaro[14,15,16], ChangHoon Hahn[17], Nico Hamaus[18,19], Donghui Jeong[20,21], Chirag Modi[22,23], Nhat-Minh Nguyen[24,25], Takahiro Nishimichi[2,26,27], Enrique Paillas[28,29], Marcos Pellejero Ibañez[30], Oliver H. E. Philcox[31,32], Alice Pisani[17,22,33,34], Fabian Schmidt[35], Satoshi Tanaka[26], Giovanni Verza[22,36], Sihan Yuan[4,6], and Matteo Zennaro[37]
[1] Department of Astronomy/Steward Observatory, The University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721, USA; krausee@arizona.edu, yosukekobayashi@arizona.edu, ansalcedo@arizona.edu
[2] Department of Astrophysics and Atmospheric Sciences, Faculty of Science, Kyoto Sangyo University, Motoyama, Kamigamo, Kita-ku, Kyoto 603-8555, Japan
[3] Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4] Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, 452 Lomita Mall, Stanford, CA 94305, USA
[5] Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305, USA
[6] SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA
[7] Theory Center, Institute of Particle and Nuclear Studies, High Energy Accelerator Research Organization(KEK), Tsukuba, Ibaraki 305-0801, Japan
[8] Donostia International Physics Center (DIPC), Paseo Manuel de Lardizabal 4, 20018 Donostia-San Sebastian, Spain
[9] IKERBASQUE, Basque Foundation for Science, E-48013, Bilbao, Spain
[10] Division of Theoretical Physics, Ruđer Bošković Institute, Zagreb HR-10000, Croatia
[11] Dipartimento di Fisica e Astronomia "Augusto Righi"—Alma Mater Studiorum Università di Bologna, via Piero Gobetti 93/2, I-40129 Bologna, Italy
[12] INFN-Sezione di Bologna, Viale Berti Pichat 6/2, I-40127 Bologna, Italy
[13] INAF-Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Piero Gobetti 93/3, I-40129 Bologna, Italy
[14] The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[15] Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[16] Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
[17] Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA
[18] Universitäts-Sternwarte München, Fakultät für Physik, Ludwig-Maximilians-Universität, Scheinerstr. 1, 81679 München, Germany
[19] Excellence Cluster ORIGINS, Bolzmannstr. 2, 85748 Garching, Germany
[20] Department of Astronomy and Astrophysics and Institute for Gravitation and the Cosmos, The Pennsylvania State University, University Park, PA 16802, USA
[21] School of Physics, Korea Institute for Advanced Study, Seoul, Republic of Korea
[22] Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA
[23] Center for Computational Mathematics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA
[24] Leinweber Center for Theoretical Physics, University of Michigan, 450 Church Street, Ann Arbor, MI 48109-1040, USA
[25] Department of Physics, College of Literature, Science and the Arts, University of Michigan, 450 Church Street, Ann Arbor, MI 48109-1040, USA
[26] Center for Gravitational Physics and Quantum Information, Yukawa Institute for Theoretical Physics, Kyoto University, Kyoto 606-8502, Japan
[27] Kavli Institute for the Physics and Mathematics of the Universe (WPI), The University of Tokyo Institutes for Advanced Study (UTIAS), The University of Tokyo, Kashiwa, Chiba 277-8583, Japan
[28] Waterloo Centre for Astrophysics, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[29] Department of Physics and Astronomy, University of Waterloo, ON N2L 3G1, Canada
[30] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh, EH9 3HJ, UK
[31] Department of Physics, Columbia University, New York, NY 10027, USA
[32] Simons Society of Fellows, Simons Foundation, New York, NY 10010, USA
[33] Aix-Marseille University, CNRS/IN2P3, CPPM, 163 Av. de Luminy, 13009, Marseille, France
[34] The Cooper Union for the Advancement of Science and Art, 41 Cooper Square, New York, NY 10003, USA
[35] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Straße 1, 85748 Garching, Germany
[36] Center for Cosmology and Particle Physics, Department of Physics, New York University, 726 Broadway, New York, NY 10003, USA
[37] University of Oxford, Astrophysics, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK
Received 2024 May 15; revised 2024 August 29; accepted 2024 September 4; published 2025 September 1

Abstract

The past few years have seen the emergence of a wide array of novel techniques for analyzing high-precision data from upcoming galaxy surveys, which aim to extend the statistical analysis of galaxy clustering data beyond the linear regime and the canonical two-point (2pt) statistics. We test and benchmark some of these new techniques in a community data challenge named "Beyond-2pt," initiated during the Aspen 2022 Summer Program "Large-Scale Structure Cosmology beyond 2-Point Statistics," whose first round of results we present here. The challenge data set consists of high-precision mock galaxy catalogs for clustering in real space, in redshift space, and on a light cone. Participants in the challenge have developed end-to-end pipelines to analyze mock catalogs and extract

* We use *to mask* (and derived forms) in place of the formerly common *to blind* throughout this manuscript.

unknown ("masked") cosmological parameters of the underlying ΛCDM models with their methods. The methods represented are density-split clustering, nearest neighbor statistics, `BACCO` power spectrum emulator, void statistics, `LEFTfield` field-level inference using effective field theory (EFT), and joint power spectrum and bispectrum analyses using both EFT and simulation-based inference. In this work, we review the results of the challenge, focusing on problems solved, lessons learned, and future research needed to perfect the emerging beyond-2pt approaches. The unbiased parameter recovery demonstrated in this challenge by multiple statistics and the associated modeling and inference frameworks supports the credibility of cosmology constraints from these methods. The challenge data set is publicly available, and we welcome future submissions from methods that are not yet represented.

Unified Astronomy Thesaurus concepts: Large-scale structure of the universe (902); Astrostatistics techniques (1886)

1. Introduction

Cosmic large-scale structure—the large-scale distribution of galaxies in the late Universe—is shaped by matter density fluctuations in the early Universe, the growth rate of the fluctuations, and the expansion rate of the smooth cosmic background. Ongoing and upcoming large-scale structure surveys will collectively map the distribution of galaxies over an extensive fraction of the sky at unprecedented depth (R. Laureijs et al. 2011; M. Takada et al. 2014; A. Aghamousa et al. 2016). These observations will enable precision tests of fundamental physics, placing tight constraints on inflation, neutrino masses, and novel properties of dark matter, gravity, and dark energy.

On large scales, matter and galaxy distributions preserve the almost perfectly Gaussian statistics from the initial conditions according to the standard inflationary models. The two-point (2pt) correlation function or power spectrum captures all information in a Gaussian field. They therefore have been the linchpin of cosmological inference and interpretation.

The situation, however, is beginning to change. Larger and higher-resolution simulations (e.g., K. Heitmann et al. 2019; T. Nishimichi et al. 2019; T. Ishiyama et al. 2021; N. A. Maksimova et al. 2021; J. DeRose et al. 2023; J. Schaye et al. 2023) and more realistic galaxy mocks (e.g., P. Behroozi et al. 2019; J. DeRose et al. 2019; A. Balaguera-Antolínez et al. 2023; S. Yuan et al. 2018; M. M. Abidi & T. Baldauf 2018; A. R. H. Stevens et al. 2024; S. Yuan et al. 2023a; C. H. To et al. 2024) have been developed to extend the inferences to smaller scales (e.g., R. M. Reddick et al. 2014; B. A. Reid et al. 2014; J. U. Lange et al. 2022, 2023; S. Yuan et al. 2022a; K. Storey-Fisher et al. 2024). Furthermore, the gain from small-scale information will become less limited by signal-to-noise ratios in galaxy samples from upcoming galaxy surveys.

On smaller scales, 2pt statistics cannot capture all the information available in the data: even with Gaussian initial perturbations, non-Gaussian features naturally emerge in the galaxy distribution as a result of nonlinear gravitational evolution and galaxy formation. Beyond-2pt statistics aim to extract information from non-Gaussian features arising from nonlinear clustering. Examples of the first cosmological constraints from galaxy clustering using beyond-2pt statistics include bispectrum analyses of PSCz (H. A. Feldman et al. 2001), 2dF (L. Verde et al. 2002), VVDS (C. Marinoni et al. 2008), Sloan Digital Sky Survey (SDSS) DR6−DR7 (E. Gaztanaga et al. 2009), and SDSS-III BOSS galaxies (H. Gil-Marín et al. 2017). More recent examples include analyses of the galaxy bispectrum (G. D'Amico et al. 2020, 2024a; O. H. E. Philcox & M. M. Ivanov 2022; M. M. Ivanov et al. 2023; C. Hahn et al. 2024), three-point function (N. S. Sugiyama et al. 2023), skew spectrum (J. Hou et al. 2024), density-split clustering (DSC; E. Paillas et al. 2023), cosmic voids (N. Hamaus et al. 2020; S. Contarini et al. 2023), and wavelet scattering transform of the galaxy density field (B. Régaldo-Saint Blancard et al. 2024; G. Valogiannis et al. 2024). While the discussion here focuses on galaxy clustering, beyond-2pt statistics are similarly gaining traction in weak-lensing analyses (e.g., A. Petri et al. 2015; J. Harnois-Déraps et al. 2021; M. Gatti et al. 2022; S. Heydenreich et al. 2022; S. Cheng et al. 2025).

In most cases, these analyses provided competitive or complementary constraints to those derived from standard 2pt analyses applied to the same data. Beyond-2pt statistics are poised to mature into a prominent role in cosmological inference from forthcoming galaxy clustering data as nonlinear clustering becomes more accessible and statistical power increases, due to improvements in the inference models and the clustering signal-to-noise ratio, respectively.

With great statistical power comes great systematic responsibility. How (in)sensitive are the beyond-2pt statistics to their modeling choices? Do they respond in known ways to observational systematics? Modeling Beyond-2pt statistics and thus physics on the associated nonlinear scales adds considerable complexities in models of both matter density perturbation and matter–galaxy connection, i.e., galaxy bias. Current Beyond-2pt statistics adopt a wide array of approaches for each modeling step: the matter field is modeled with a variety of perturbative approaches or *N*-body simulations, and the matter–galaxy connection is usually encoded via bias expansions, Halo Occupation Distribution (HOD) models, or subhalo abundance matching (SHAM) schemes.

In addition to modeling systematics, how well do we understand the likelihoods and covariances of these novel statistics? Can we evaluate and estimate them at the precision required by current and future surveys? Specifically, intractable non-Gaussian likelihoods[38] must be approximated with simulations. Furthermore, analytic covariances are not always available; hence, simulations and mocks are again required to estimate the covariances. The difficulties inherent in the generation of numerous high-resolution simulations and mock data add another layer of complexity to such analyses, which could potentially compromise their robustness. Consequently, the first aim of this data challenge is to validate a variety of beyond-2pt statistics and their modeling approaches.

Our second aim is to address the following question: How much information can we (robustly) extract from galaxy

[38] See, e.g., C. Hahn et al. (2019) or N.-M. Nguyen et al. (2021) for recent examples of non-Gaussian likelihoods and their impacts on cosmological inference from galaxy clustering.

clustering? Although individual comparisons of information content between standard and beyond-2pt statistics have been made elsewhere, no direct comparison between the beyond-2pt methods presented here was attempted with the same survey volume and with the same galaxy density. This challenge therefore aims to facilitate a community exercise to study the information content in Beyond-2pt statistics by providing a standard set of simulated mock data. A critical quantity that controls the amount of beyond-linear information is the maximum cutoff scale in each analysis, $k_{\rm max}$. As each beyond-2pt analysis might have a different sensitivity to $k_{\rm max}$, we let each group determine their own $k_{\rm max}$ up to which they can trust their model against model misspecifications. However, to avoid fine-tuning while simultaneously encouraging better uncertainty quantification and systematic control in each analysis, we mask the parameters of the mock catalogs, which include both the cosmological parameters and the galaxy HOD parameterization(s) plus their parameter values.

For galaxy clustering, only a single public, parameter-masked mock challenge exists prior to this work—the PTchallenge, introduced in T. Nishimichi et al. (2020)—that specifically targets 2pt statistics. PTchallenge was a cornerstone for the establishment of EFT as a mature tool to analyze the full shape of the galaxy power spectrum. In particular, it demonstrates the ability of one-loop EFT theories to recover *masked* cosmological parameters from both amplitude and shape of the (nonlinear) power spectrum at a subpercent precision.

Important technical lessons from PTchallenge are (1) learning parameter degeneracies and optimizing analysis choices with respect to them, (2) developing a methodology to select scale cuts, and (3) understanding the role of EFT parameters and their priors. Since the original publication, the true cosmology of the PTchallenge has been kept masked, and the PTchallenge continues to serve as a testing ground for new galaxy power spectrum models, e.g., Lagrangian EFT (S.-F. Chen et al. 2020, 2021), a simulation-based plus HOD-based emulator (Y. Kobayashi et al. 2020, 2022), and a perturbation-based shape-fitting approach (S. Brieden et al. 2021, 2022). However, the PTchallenge only made available data in the form of redshift-space power spectrum measurements.

Our "Beyond-2pt" mock challenge extends the PTchallenge (T. Nishimichi et al. 2020) in several ways. First, we present the data directly at the level of mock galaxy catalogs, rather than summary statistics thereof. Estimator validation is therefore part of the challenge for each clustering statistic. Second, we present mock catalogs at three different complexity levels: real-space or redshift-space snapshots and a light cone. These mocks cover a redshift range around $z \sim 1$ with different masked flat ΛCDM cosmologies and different mock galaxy populations with masked HOD parameterization(s). The range of complexity levels enables participation by statistics and methods at different maturity levels, while also facilitating robustness tests of different modeling prescriptions. Third, prior to this, many prescriptions were only tested on mock galaxies resembling those found in the BOSS CMASS sample within a Planck-like cosmology. This challenge—extending both galaxy and cosmology models—therefore marks a step forward in this regard as well.

Finally, we extend the challenge to beyond-2pt statistics, showing their constraints on cosmological parameters, [$\Omega_{\rm m}$, σ_8], side by side with the corresponding constraints from BACCO P, a representative of 2pt statistics, for all setups and dissecting the information content they capture. The present challenge presents a first such benchmark and sets the stage for more detailed comparisons in the future. With eight independent analyses submitted from seven international teams, this challenge represents the first community effort toward developing optimal strategies to extract cosmological information from upcoming galaxy surveys.

The mocks and analyses presented in this paper assume a flat ΛCDM cosmology, which can be described by five parameters: (1) the dimensionless Hubble parameter h; (2) the matter fluctuation or primordial power spectrum amplitude, parameterized by either σ_8 or $A_{\rm s}$; (3) the power spectrum spectral index $n_{\rm s}$; (4) the density parameter of cold dark matter $\Omega_{\rm cdm}$; and (5) the density parameter of baryons $\Omega_{\rm b}$. Different analyses adopt priors on either the density parameters Ω_x or the physical density parameters $\omega_x = \Omega_x h^2$. Though no analysis varies neutrino mass, some assume massless ($\omega_\nu = 0$) while others adopt minimal-mass neutrinos ($\omega_\nu = 0.0006442$). The main results of this challenge are summarized and presented in marginalized constraints on σ_8 and the total matter density, $\Omega_{\rm m} = \Omega_{\rm cdm} + \Omega_{\rm b} + \Omega_\nu$.

To set the stage for the summary of results in Section 2, we briefly review the information content in galaxy clustering, in the context of the ΛCDM model (with Gaussian initial conditions), as considered throughout this paper. In linear theory, all cosmological information encoded in galaxy clustering is captured by the linear power spectrum, particularly its scale dependence and amplitude. The scale dependence is sensitive to both the initial density perturbations, captured by the spectral index $n_{\rm s}$, and the stress-energy components that determine the background evolution and growth of perturbations. The physical density parameters $\omega_{\rm m,b}$ determine two characteristic scales: the Hubble horizon at matter–radiation equality and the sound horizon at photon–baryon decoupling, which are imprinted on the broadband shape of the matter power spectrum and baryonic acoustic oscillation (BAO) wiggles, respectively. All the challenge mocks include this information.

In real space, i.e., without accounting for observational effects of galaxy peculiar motion, the amplitude of the (linear) galaxy power spectrum at redshift z is proportional to the product of the amplitude of primordial fluctuations, the linear growth factor $D(z)$, and the linear galaxy bias b_1. That is, the cosmology parameter σ_8 is completely degenerate with the unknown linear galaxy bias parameter. Observational effects encode additional information: the redshift-space distortion (RSD) effect contains information on the growth rate f, which breaks the degeneracy between σ_8 and b_1, and the Alcock–Paczynski (AP) effect contains geometry information, which constrains $\Omega_{\rm m}$ (within ΛCDM). Growth information from RSD is included in redshift-space and light-cone mocks, while geometry information from AP is only available on the light cone.

Beyond the linear regime, quasi- and nonlinear clustering and collapsed structures encode additional information with different parameter degeneracies. The results from this challenge underscore the robust extraction of such nonlinear information either through the nonlinear matter power spectrum (BACCO P) or through statistics that have access to higher-order n-point functions (all other teams). At the power spectrum level, nonlinear evolution introduces additional smearing of the

Table 1
Overview of Participating Analyses and Their Analysis Ingredients

Method	Sections	Mock(s) Analyzed	Gravity Model	Tracer Model	Model Evaluation	Covariance	Cosmology Prior	Team
BACCOP	5.1	all	*N*-body	hybrid-EFT	emulator	analytic + emulator cov.	Equation (7)[a]	M. Pellejero, R. Angulo and M. Zennaro
Density Split	5.6	redshift space	*N*-body	HOD	emulator	1500 mocks + emulator cov.	Equation (32)	E. Paillas and C. Cuesta-Lazaro
EFT FBI	5.3	real space	perturbative	EFT	analytic	analytic	Equation (20)	N. M. Nguyen and F. Schmidt
EFT P+B	5.2	all	perturbative	EFT	analytic	analytic	Equation (10)[a]	M. Ivanov, O. Philcox, G. Cabass and K. Akitsu
*k*NN	5.5	redshift space	*N*-body	HOD	emulator	jackknife + emulator cov.	Equation (27)	S. Yuan and T. Abel
SBI P+B	5.4	redshift space	*N*-body	HOD	galaxy mocks	N/A	Equation (26)	C. Modi and CH. Hahn
VGCF	5.7	light cone	perturbative	linear bias	analytic	jackknife	Table 4	N. Hamaus, S. Contarini, G. Verza and A. Pisani
VSF	5.7	light cone	perturbative	linear bias	analytic	analytic	Table 4	S. Contarini, G. Verza, N. Hamaus and A. Pisani

Note.
[a] Cosmology priors are listed for the baseline validation and original unmasking submission. Teams BACCO P and EFT P+B kindly reran their analyses with different cosmology priors (but otherwise identical analysis choices) to facilitate comparisons.

BAO feature and an enhancement of small-scale power (C. D. Rimes & A. J. S. Hamilton 2005; M. Crocce & R. Scoccimarro 2008; R. E. Angulo et al. 2021). The distinct dependencies of beyond-2pt statistics on cosmology and astrophysics break parameter degeneracies and improve constraints on cosmological parameters. We refer the reader to individual analysis sections, as well as references therein, for detailed discussions on the sources of nonlinear information.

The remainder of the paper is organized as follows. To help readers navigate the following sections, we provide tables summarizing the different analyses (Table 1) and the common notation (Table 2). In Section 2, we summarize and discuss the results of eight parameter-masked analyses. Figures 1–4 present the key results, as constraints on the parameter combination $[\Omega_m, \sigma_8]$ from the redshift-space mocks in Figures 1–2 or from the light-cone mock in Figure 3, and as constraints on the parameter σ_8 from the real-space mocks in Figure 4. We introduce the suite of mock catalogs—the main data product of the "Beyond-2pt" challenge—in Section 3 and the parameter (un)masking procedure in Section 4. In Section 5, each team describes their analysis method. We describe post-unmasking reanalyses and resulting lessons for method refinement and optimization of constraining power in Section 6. We summarize the results of this challenge and discuss implications for future mock challenges and beyond-2pt analyses in Section 7.

Table 2
Notation for Parameter Inference Variables

Variable	Symbol
Data vector	$\widehat{\boldsymbol{d}}$ (or $\widehat{d}_i$)
Cosmological parameters	$\boldsymbol{\Theta} \equiv \{\sigma_8, \Omega_m, h, n_s, \cdots\}$
Nuisance parameters	$\boldsymbol{\Phi} \equiv \{b_1, b_2, \cdots\}$
Parameters	$\boldsymbol{\Omega} = \boldsymbol{\Theta} \cup \boldsymbol{\Phi}$
Model data vector	$\boldsymbol{m}(\boldsymbol{\Omega})$
Covariance	(or $\mathbf{C}_{ij}$)
Covariance estimate	$\widehat{\mathsf{C}}$
Likelihood	$\mathcal{L}[\widehat{\boldsymbol{d}}\|\boldsymbol{\Omega}]$
Posterior	$\mathcal{P}[\boldsymbol{\Omega}\|\widehat{\boldsymbol{d}}]$
Gaussian distribution	$\mathcal{N}$, e.g., $b_1 \sim \mathcal{N}(\mu, \sigma^2)$
Uniform distribution (e.g., prior)	$\mathcal{U}$, e.g., $h \sim \mathcal{U}[0, 1]$

2. Summary of Results

In this section, we summarize the main results of the paper, shown in Figures 1–4, and provide an overview of the eight participating analyses in Table 1.

This challenge is based on a series of mock galaxy catalogs created from *N*-body simulations with a flat ΛCDM cosmology and HOD galaxy–halo connection models. It contains three levels of increasing realism:

1. real-space galaxy distribution from a simulation snapshot with periodic boundary conditions (10 realizations with the same cosmology + HOD parameters at $z = 1$);
2. redshift-space galaxy distribution from a simulation snapshot with periodic boundary conditions (10 realizations with the same cosmology + HOD parameters at $z = 1$); and
3. a light-cone galaxy mock emulating the observational data most closely (one realization, uniform coverage across $0.8 < z < 1.3$ and with a simple footprint bounded by lines of constant R.A. and decl.).

Initially, the organizers communicated to the analysis teams only that the challenge catalogs are HOD-based galaxy mocks in flat ΛCDM cosmologies without observational systematics, hiding the true cosmological parameters and the HOD model used to generate the simulations from the analysis teams. Each analysis team then analyzed these parameter-masked mocks while documenting their model, inference, and analysis choices including scale cuts and parameter priors, before submitting their parameter posteriors to the organizers, who subsequently unmasked the fractional distances between posterior means and the ground truth.

The challenge results presented here indicate whether a given analysis returns parameters consistent with the ground truth given their reported error bars. For real-space and

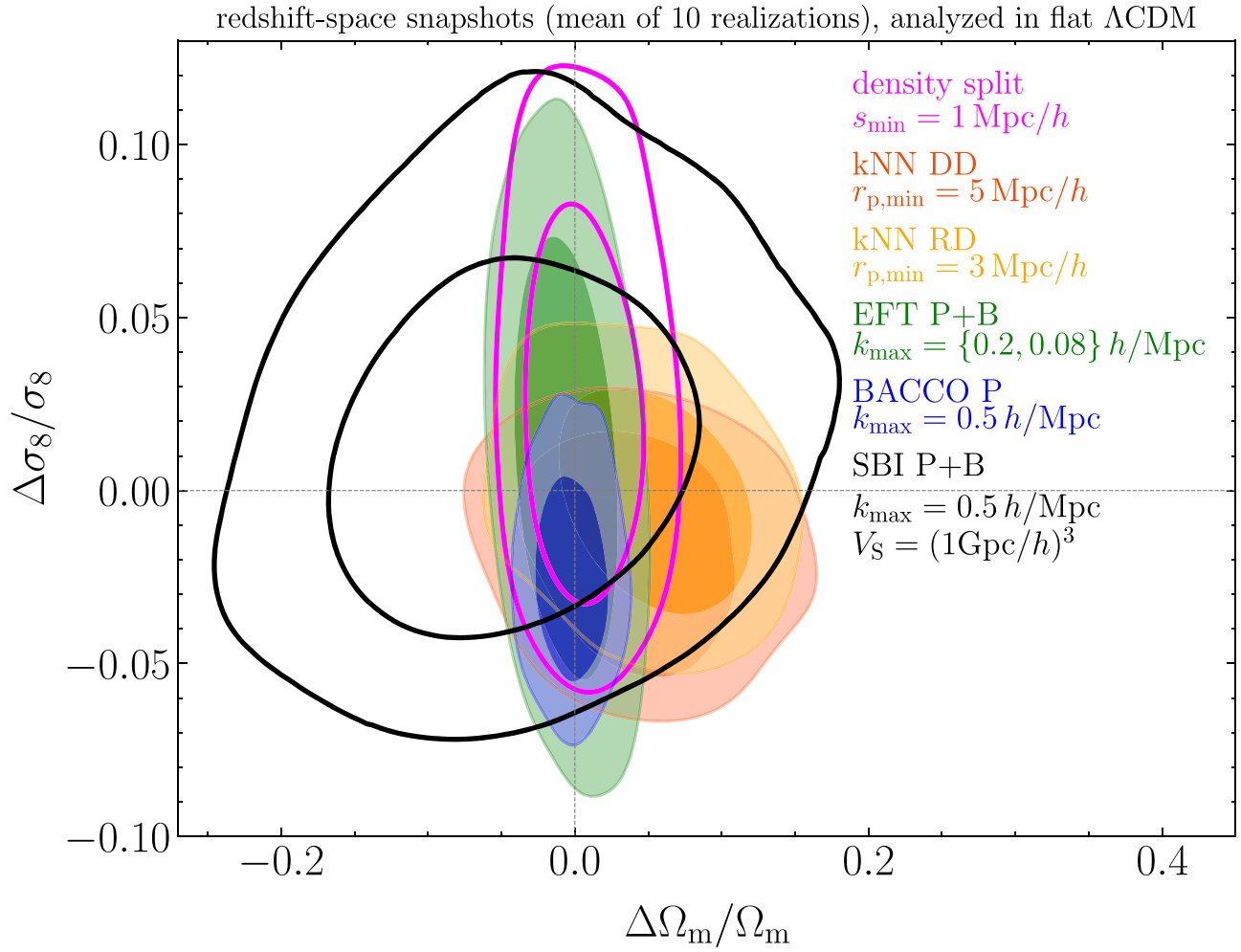

Figure 1. 2D marginalized constraints on $\Omega_{\rm m}$ and σ_8 for parameter-masked analyses of redshift-space mocks (mean of 10 realizations, errors of 1 box), marginalized over the remaining cosmological parameters of flat ΛCDM and nuisance parameters specific to each method. We list the scale cuts due to nonlinear modeling for different analyses, which is specified in terms of redshift-space separation (s), projected radius (r_p), or Fourier mode (k), depending on the analysis method.

redshift-space analyses, analysis teams report results corresponding to the mean of 10 realizations analyzed with the covariance of a single-realization volume to minimize parameter biases due to cosmic variance. Assuming a Gaussian data likelihood, in this setup the probability of a $\geqslant 1\sigma$ fluctuation due to cosmic variance is 0.16%. Hence, biases in the inferred parameters likely indicate model misspecification or incomplete uncertainty modeling. The challenge results do not directly establish whether the reported parameter uncertainties fully reflect the true uncertainties (measurement and model). This is particularly relevant for small-scale analyses and simulation-based methods, based on empirical galaxy–halo connection parameterizations and priors. Hence, the results should not be regarded as a direct quantitative comparison between different analysis methods or summary statistics.

We report results ordered by number of participating analyses for the different mocks. All results on the redshift-space and light-cone mock catalogs are from analyses in flat ΛCDM with priors that are much broader than current-generation observational constraints (Planck Collaboration et al. 2020; S. Alam et al. 2021; DESI Collaboration et al. 2025). However, the cosmology priors differ between all analyses, as illustrated in Figure 5. We note that the void size function (VSF) analysis of the light-cone mock adopts informative priors on three parameters that are weakly constrained by the VSF, with width of three times the current Planck2018 (Planck Collaboration et al. 2020) uncertainties. The analyses on the real-space mock reported here fix all cosmological parameters except σ_8 to their true values.

2.1. Results on Redshift-space Mocks

The redshift-space snapshots were analyzed by most teams, comprising the following analyses: DSC (Section 5.6), two different flavors of nearest neighbor statistics (kNN; Section 5.5), a joint power spectrum plus bispectrum analysis using EFT (EFT P+B; Section 5.2), a hybrid-EFT power spectrum analysis with the BACCO emulator (BACCO P; Section 5.1), and a simulation-based inference analysis of the power spectrum plus bispectrum (SBI P+B; Section 5.4). The SBI P+B analysis was trained on simulations with volume $(1\,h^{-1}\,{\rm Gpc})^3$, while all other analyses assume a covariance matrix corresponding to the sample variance of a $(2\,h^{-1}\,{\rm Gpc})^3$ box, resulting in degradation in constraining power. While it would be convenient to assume a simple survey volume rescaling to compare SBI P+B results with other analyses, we refrain from making such a rescaling since constraining power is a function of both survey volume and scale cuts ($k_{\rm max}$),

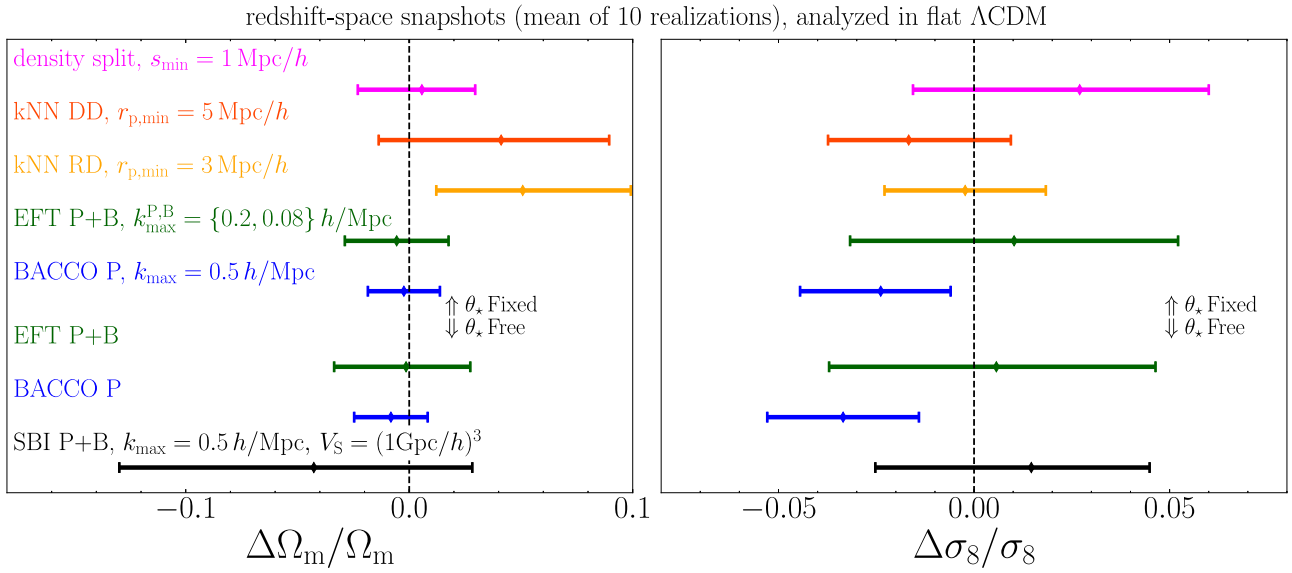

Figure 2. 1D marginalized constraints on $\Omega_{\rm m}$ and σ_8 for parameter-masked analyses of redshift-space mocks (mean of 10 realizations, errors of 1 box), marginalized over the remaining cosmological parameters of flat ΛCDM and nuisance parameters specific to each method.

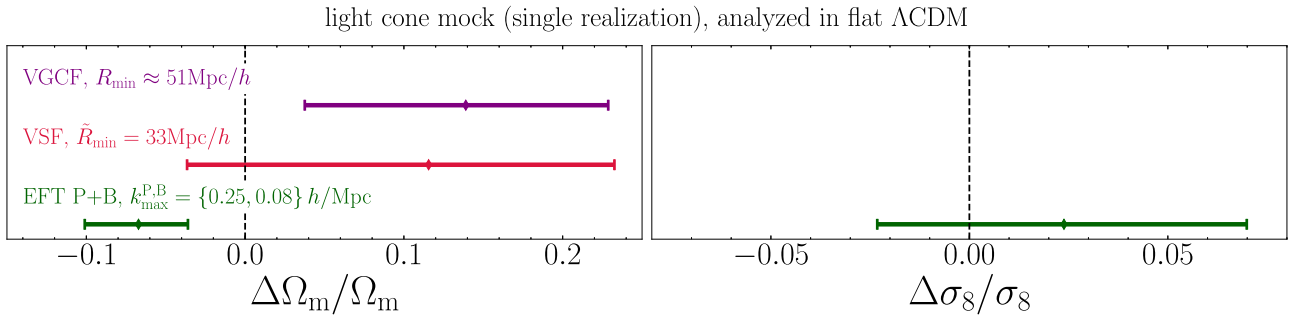

Figure 3. 1D marginalized constraints on $\Omega_{\rm m}$ and σ_8 for parameter-masked analyses of the light-cone mock (single realization), marginalized over the remaining cosmological parameters of flat ΛCDM and nuisance parameters specific to each method. Void size cuts are specified in terms of the effective radius R and the cleaned (rescaled) radius $\tilde{R}$.

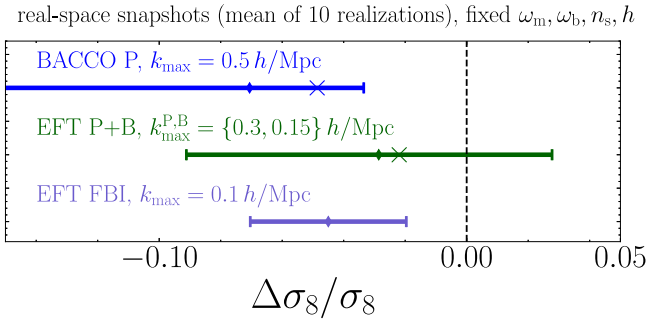

Figure 4. 1D marginalized constraints on σ_8 for analyses of real-space mocks (mean of 10 realizations, errors of 1 box), in a reduced parameter space with all cosmological parameters except σ_8 fixed at the fiducial values. Crosses indicate the maximum a posteriori estimates, to illustrate projection effects.

which need to be calibrated anew when changing the survey volume.

Figure 1 shows the marginalized constraints in the $\Omega_{\rm m}$–σ_8 plane inferred from the different summary statistics. Remarkably, all analysis teams successfully recover the input cosmology within their 1σ confidence region. This result from a parameter-masked challenge further demonstrates the maturity of these reportedly "novel" statistics and their potential for analyses of near-term data.

Figure 2 shows the marginalized posteriors in 1D for $\Omega_{\rm m}$ and σ_8. The nominal distance between the mean and the ground truth (i.e., bias) in $\Omega_{\rm m}$ is the smallest for the EFT P+B, BACCO, and density-split analyses, while for σ_8 the unmasking showed the lowest parameter bias for the EFT P+B, kNN RD, and SBI P+B analyses. After unmasking, further analysis by the BACCO team identified a large emulation uncertainty of the hexadecapole as a potential source of this parameter bias (see Section 5.1.3), which had been unnoticed in previous tests on smaller simulation volumes. We further discuss this post-unmasking reanalysis in Section 6. Focusing on the analyses with fixed θ_*—the angular size of the sound horizon—we find similar error bars on $\Omega_{\rm m}$ from the density-split, EFT P+B, and BACCO methods, implying that, for this particular parameter and mock data, the bulk of the information comes from quasi-linear (BAO) scales. The tightest error bars on σ_8 are obtained with the kNN and BACCO methods.

Both DSC and kNN statistics are modeled with emulators built on the AbacusSummit simulations, which fix θ_*, while the SBI P+B analysis is built on the Quijote simulations, which assume no such constraint. To ease the comparison, the EFT P+B and BACCO P teams ran additional analyses imposing the same prior. A comparison of the EFT P+B constraints in both parameter spaces (top and bottom part of Figure 2) indicates that this difference in cosmology priors has

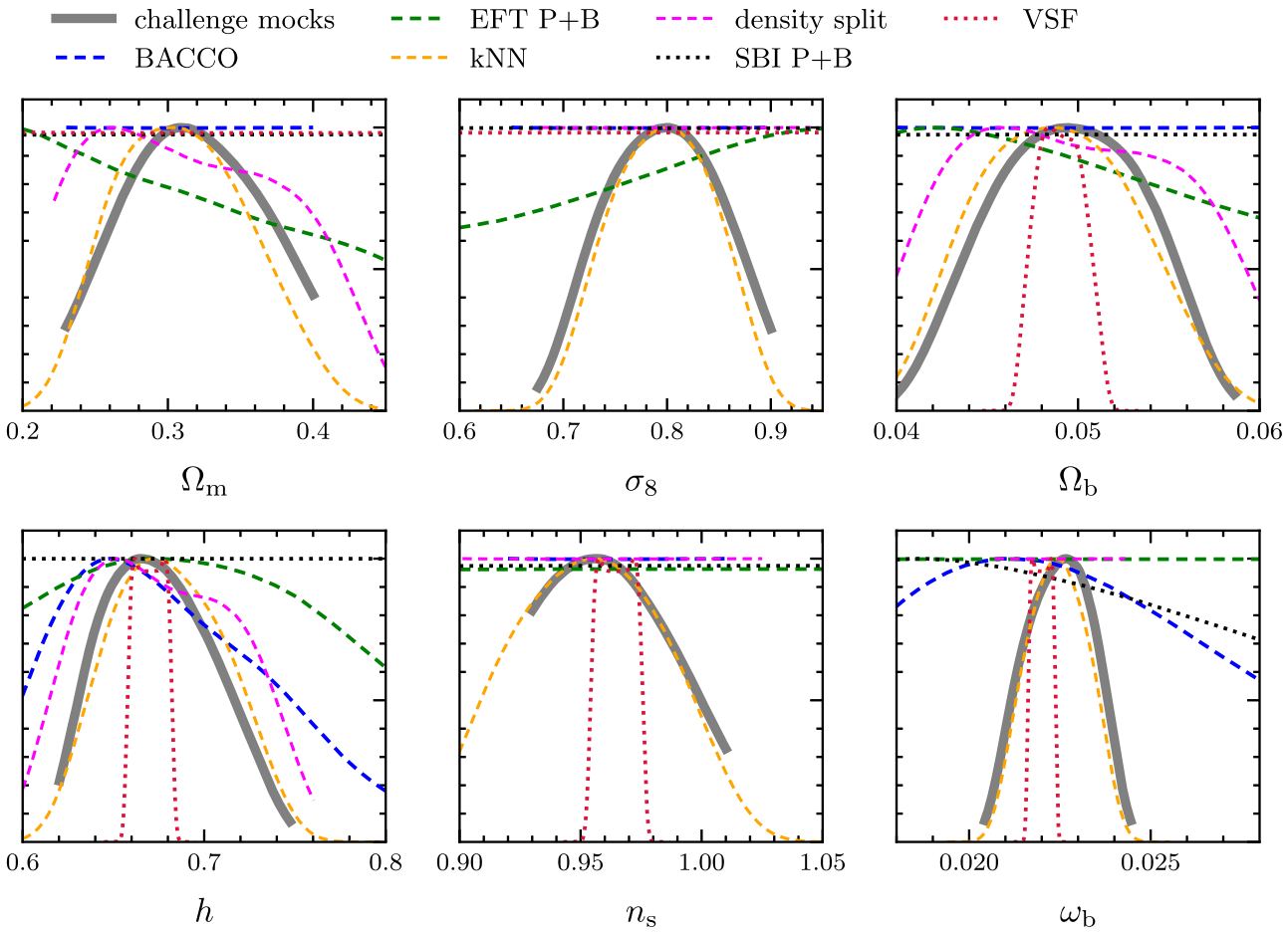

Figure 5. 1D marginalized cosmology priors in flat ΛCDM for the challenge mocks (solid gray line) and those adopted by individual analysis teams. Analyses that fix θ_* are shown with dashed lines, and analyses with free θ_* are shown as dotted lines. While there are only four/five independent parameters in flat ΛCDM with/without θ_* constraint, analysis teams specify their priors in different parameterizations. To facilitate comparison, we project priors in different parameters, noting that not all panels are independent.

limited impact on the constraints shown here. The EFT P+B and SBI P+B results are quite similar in terms of σ_8 constraining power, despite the difference in survey volume. For $\Omega_{\rm m}$, however, the SBI P+B constraint is significantly wider than the EFT P+B result. We believe that the wide SBI P+B posterior on $\Omega_{\rm m}$ is a result of trimming the largest scales in the SBI P+B analysis, which down-weights large-scale modes that are important in the $\Omega_{\rm m}$ recovery from the turnover of the galaxy power spectrum (see M. M. Ivanov et al. 2020a; O. H. E. Philcox et al. 2021b).

2.2. Results on Light-cone Mock

Unlike the redshift-space data discussed above, there exists only one realization for the light-cone mock. Hence, scatter in recovered parameter values is expected owing to cosmic variance of the summary statistics measurements. Since different analyses include (substantially) different scales, they are subject to different cosmic variance modes and may be scattered in different directions.

The effective volume of this mock catalog is similar to that of the BOSS DR12 large red galaxy sample. For this mock, we have submissions from the EFT P+B, the VSF (Section 5.7), and the void-galaxy cross-correlation function (VGCF; Section 5.7) teams; the parameter constraints are shown in Figure 3. Both void methods and EFT P+B successfully recover the input values of $\Omega_{\rm m}$ within 95% confidence level (CL), and the EFT pipeline has additionally recovered the true input value of σ_8. We note that the true cosmological parameters are outside the informative prior (on h, $\omega_{\rm b}$, and $n_{\rm s}$) adopted by the VSF analysis. However, post-unmasking analyses indicate at most minor bias in $\Omega_{\rm m}$ due to the miscentered prior. The impact of prior width on the VSF constraining power is discussed in Section 5.7.4. The void team has not submitted the σ_8 results, as its accurate inference from the VSF is not possible without an independent calibration of the tracer bias inside voids, e.g., taken from realistic mocks of the sample in question (see S. Contarini et al. 2023 and Section 5.7.2). For analyzing the VGCF, in this paper only the AP test is exploited for cosmological inference, which exclusively constrains parameters describing the background evolution.

2.3. Results on Real-space Mock

The real-space challenge boxes were analyzed with the EFT-based field-level Bayesian inference (EFT FBI) approach using the `LEFTfield` code (Section 5.3), EFT P+B, and the `BACCO` power spectrum emulator. As for the redshift-space analyses, the summary statistics measurements were averaged over the 10 realizations of the ΛCDM real-space galaxy catalog from a periodic box, and all `BACCO` and EFT P+B inferences

were performed assuming a covariance of a $(2\ h^{-1}\ \mathrm{Gpc})^3$ box. For EFT FBI, we show the average of the 10 independent posteriors.

All three analysis teams fixed all cosmological parameters to the simulation true values in this challenge, except σ_8. EFT P+B and BACCO had originally validated and unmasked their analyses of the real-space mocks in the full flat-ΛCDM parameter space. After unmasking, both teams kindly reran their analyses with all cosmological parameters except σ_8 set to their truth values to enable comparison with EFT FBI. This allows a direct comparison between σ_8 constraints obtained by the three statistics, as the EFT FBI team and their `LEFTfield` pipeline currently cannot perform a full inference over the ΛCDM cosmological parameter space. This simplified setup, however, still offers insights into the information content in (nonlinear) clustering beyond the 2pt function. The results are presented in Figure 4. The constraints of the `BACCO` analysis are obtained with only the real-space power spectrum and are thus heavily impacted by the b_1–σ_8 degeneracy present at leading order in the power spectrum. The degeneracy is broken only by nonlinear contributions to the power spectrum. The EFT P+B and FBI teams use extended data vectors to mitigate this b_1–σ_8 degeneracy.

All analysis teams recover the true value of σ_8 within 95% CL of one single simulation box. At face value, the EFT FBI analysis achieves the tightest constraints on σ_8. Compared with the EFT P+B analysis, the FBI analysis assumed a more restricted model for galaxy bias and galaxy stochasticity. The two EFT teams subsequently attempted to quantify the impact of their modeling differences in post-unmasking studies; we refer readers to Sections 5.3 and 6 for the results and discussions.

3. Mock Data

The challenge organizers present three different types of mock catalogs based on *N*-body simulations—real-space snapshot, redshift-space snapshot, and light cone—all using the HOD prescription. The mock catalogs are publicly available at this repository .

3.1. N-body Simulations and Halo Catalogs

3.1.1. Flat ΛCDM Snapshot Mocks

After agreement on the range for the cosmological parameters described in Section 4, the organizers created the flat ΛCDM mock catalogs in $z = 1$. Ten realizations have been created for each of the real- and redshift-space mocks, each of which is a cubic simulation box with a comoving side length of $2\ h^{-1}$ Gpc. The *N*-body simulations have been run by using the cosmological *N*-body simulation code `GINKAKU` (T. Nishimichi et al. 2025, in preparation). This code employs the tree particle-mesh (TreePM) method to compute the gravitational force in an expanding periodic box in comoving coordinates. The short-range tree force is implemented based on the Framework for Developing Particle Simulators (`FDPS`; M. Iwasawa et al. 2016; D. Namekata et al. 2018), a public library for general particle simulations. FDPS facilitates scalable computations on modern supercomputer systems, ensuring optimized workload balance with efficient domain decomposition. The tree force is further accelerated by SIMD instructions implemented in the `Phantom-GRAPE` library (K. Nitadori et al. 2006; A. Tanikawa et al. 2012, 2013). Details of the long-range PM force can be found in K. Yoshikawa & T. Fukushige (2005; see also T. Ishiyama et al. 2009, 2012, for recent implementations).

The initial conditions are generated using second-order Lagrangian perturbation theory (2LPT; R. Scoccimarro 1998; M. Crocce et al. 2006) at $z = 49$. Similarly to the existing *N*-body simulation ensembles like `AbacusSummit` or `Dark Quest`, the organizers have run the simulations without the effect of massive neutrino dynamics and have incorporated massive neutrinos only in the matter transfer functions with the abundance $\omega_\nu = 0.000644$. The transfer function has been computed at $z = 0$ and rescaled to the starting redshift $z = 49$ using the linear growth factor with matter density $\Omega_\mathrm{m} = (\omega_\mathrm{b} + \omega_\mathrm{cdm} + \omega_\nu)h^{-2}$, but without the scale dependence induced by massive neutrinos.

For each realization, dark matter halos are populated with mock galaxies as described in Section 3.2. We keep the choice of halo finder masked, as differences in halo finder can contribute to model misspecification in the highly nonlinear regime (e.g., C. Hahn et al. 2023b; C. Modi et al. 2025). Neither real-space nor redshift-space mocks include the AP effect (C. Alcock & B. Paczynski 1979). This effect contributes to constraining the cosmological parameters in a more realistic galaxy spectroscopic survey setting and is taken into account in the light-cone mocks described next.

3.1.2. Flat ΛCDM Light-cone Mocks

The single-realization ΛCDM light-cone mock is based on one of the publicly available light-cone halo catalogs from the `AbacusSummit` set of simulations (N. A. Maksimova et al. 2021; B. Hadzhiyska et al. 2022) generated with the `abacus` cosmological *N*-body code (L. H. Garrison et al. 2019). The specific `AbacusSummit` cosmology realization was chosen to be within the parameter range of all emulators participating in this challenge (see Section 4). The organizers shared that the light-cone mock is one of the `AbacusSummit` cosmologies with participants only after unmasking.

Two teams (*EFT P+B* and *Cosmic Voids*) submitted results for the light-cone mock. Both analyses rely on analytic modeling prescriptions, eliminating concerns that the same halo mock catalog might inadvertently be used in training an emulator for this challenge.

3.2. HOD Galaxy Mocks

To generate mock galaxy catalogs for our challenge, we take advantage of the HOD formalism (e.g., A. A. Berlind & D. H. Weinberg 2002; J. S. Bullock et al. 2002; Z. Zheng et al. 2005; V. Gonzalez-Perez et al. 2018; A. N. Salcedo et al. 2022a; S. Yuan et al. 2024b). In practice, we specify a parameterized form for the mean halo-mass-dependent occupation of galaxies and stochastically populate the dark matter halos with galaxies. The galaxy distribution within a host halo broadly traces the host's internal structure but may differ in detail in a halo-mass-independent way. We may also in principle include some level of galaxy assembly bias, which refers to the possibility for the occupation of halos of a given mass to depend on properties other than mass (e.g., A. P. Hearin et al. 2016; J. E. McEwen & D. H. Weinberg 2018; X. Xu et al. 2021; A. N. Salcedo et al. 2022b; K. Wang et al. 2022; G. D. Beltz-Mohrmann et al. 2023; S. Contreras et al. 2023a; Z. Zhai et al. 2023b). In this context we may implement galaxy assembly bias with respect to halo

properties present in our simulation catalogs, or environmental properties of the simulation particle distribution. The light-cone mock may include redshift evolution of some of our HOD parameters. In our redshift-space mocks (including the light cone) we also add galaxy peculiar velocities that may include velocity bias (e.g., F. C. Van Den Bosch et al. 2005; B. A. Reid et al. 2014; H. Guo et al. 2016; S. Yuan et al. 2018; D. Anbajagane et al. 2022; J. U. Lange et al. 2022; G. D. Beltz-Mohrmann et al. 2023; J. Kwan et al. 2023; Z. Zhai et al. 2023a; K. J. Kwon & C. Hahn 2024). The positions of mock galaxies are computed by modulating the real-space coordinates along the z-axis according to their peculiar velocities. We emphasize that in constructing the mock galaxy catalogs in this challenge we have remained broadly within the existing literature on the galaxy–halo connection and have by no means attempted to confound the challenge participants.

4. Parameter Masking Implementation

4.1. Pre-unmasking Information

Initially, the only information communicated to the analysis teams was the cosmology parameter space (flat wCDM) and that the galaxy assignment is based on an HOD technique, without specifying the detailed HOD implementation.

During initial test runs on these wCDM mocks, several teams realized that the (still-masked) cosmologies were likely outside the parameter support of some emulator-based methods.

The analysis teams then agreed on a common range for the cosmological parameters, which correspond to the intersection of the parameter spaces supported by the `AbacusSummit`, `Aemulus`, and `BACCO` emulators, which include a hard constraint on the angular scale of the sound horizon at decoupling θ_* imposed in the `AbacusSummit` simulation suite (N. A. Maksimova et al. 2021).

A script implementing this parameter restriction was shared with all analysis teams,[39] and the challenge organizers chose new sets of cosmological parameters to generate flat ΛCDM mock data within the specified parameter range. The cosmology prior for the mock catalogs and the priors of individual analyses (documented in Section 5) are shown in Figure 5.

Each analysis team documented their analysis choices and unmasking criteria, as well as potential caveats of their analysis (which could be revisited in the event of an unmasking surprise).

We show only parameter offsets to not unmask future participants and enable continued use of the mock data sets as a benchmark for testing novel analysis techniques. The organizers encourage future submissions from other analysis teams and commit to continuous unmasking.

4.2. Unmasking Process

There is only one parameter unmasking stage. Each analysis team chooses when they are ready to unmask and share their parameter chains with the challenge organizers, who produce the plots described in Section 2 and share values of the (marginalized) parameter shifts $\Delta\Omega_{\rm m}$ and $\Delta\sigma_8$ with the analysis team. Analysis teams agreed not to share parameter values with other teams but to only show plots with offset parameter values.

[39] Available at but not incorporated by any of the participating analyses.

4.3. Post-unmasking Analyses

In the event that an analysis team decides to adjust their analysis choice(s) after unmasking, they will document all post-unmasking tests in the paper and commit to showing the original unmasking result along with the post-unmasking update.

In case some novel analyses may require expanded discussions and future work to reach the maturity required for precision cosmology analyses, post-unmasking analyses and discussions are encouraged for this paper.

4.4. Accommodations

In practice, the organizers unmasked submissions from all analysis teams but one in a joint video conference. To encourage submissions from analyses from emerging methods with less extensive validation, the unmasking plot shared in the unmasking video conference showed results from all submissions with anonymized labels (e.g., "Team Blue" and "Team Yellow"). Each analysis team was informed only of their own label in advance, and the organizers did not share with participants the list of participating analyses. Organizers and participants had agreed on the option for analysis teams to withdraw from the challenge after unmasking to provide anonymity to unexpected results. Ultimately, no analysis team made use of the option to withdraw their submission.

The organizers communicated extensively with analysis teams to facilitate direct comparisons between analyses while preserving the parameter masking. Specifically, the organizers coordinated additional submissions from `BACCO` and EFT P+B with different cosmology priors: (i) ΛCDM without hard θ_* constraint (see Section 4.1) to enable comparison with the SBI P+B analyses based on `Quijote` simulations without θ_* constraint (Figure 2), and (ii) fixing all parameters except σ_8 to the true cosmology of the real-space mocks for comparison with EFT FBI results (Figure 4). For the latter comparison, truth values were shared after unmasking of the `BACCO` and EFT P+B submissions for the full parameter space. In addition, the organizers shared the normalized initial conditions of real-space mock `box 1` and truth values of all cosmological parameters except σ_8 with the EFT FBI team.

Feedback from participants suggests that these accommodations enabled broader participation and further improved the participants' experience.

5. Analysis Methods

The following subsections detail the modeling, inference, and parameter unmasking choices of the different analysis teams.

Notation. We adopt the notation in Table 2 for common variables. For example, with this notation, the Gaussian data likelihood adopted in most inferences is given by

$$-2\log\mathcal{L}[\hat{\boldsymbol{d}}|\boldsymbol{\Omega}] = (\boldsymbol{m}(\boldsymbol{\Omega}) - \hat{\boldsymbol{d}})^T\mathsf{C}^{-1}(\boldsymbol{m}(\boldsymbol{\Omega}) - \hat{\boldsymbol{d}}) + \text{constant}. \quad (1)$$

Throughout, we assume a flat ΛCDM cosmology and use $H(z)$ for the Hubble rate at redshift z and $D_A(z)$ for the angular diameter distance.

HOD Models. Analyses of galaxy clustering that extend into the highly nonlinear regime require a model for the galaxy–halo connection that relates galaxies to individual halos, for example, HOD models (see Section 3.2 for extended

references). This is in contrast to the EFT bias expansion relating the smoothed galaxy density field to the matter density on large scales.

Three of the participating analyses (SBI P+B 5.4, kNN statistics 5.5, DSC 5.6) follow the HOD approach, but with somewhat different implementations and model extensions. Specifically, analysis 5.4 uses one parameterization of the HOD model (see C. Hahn et al. 2023b, for details), while analyses 5.5 and 5.6 use a somewhat different parameterization (see S. Yuan et al. 2022b for details). For brevity, we refer to the two models as HOD1 and HOD2, respectively, in this overview.

First, we briefly introduce the vanilla HOD formalism. Statistically, the HOD can be summarized as a probabilistic distribution $P(n_g|\boldsymbol{X}_h)$, where n_g is the number of galaxies of the given halo and $\boldsymbol{X}_h$ is some set of halo properties. Typically the galaxy population is divided into central and satellite populations and assumes that the central galaxy occupation follows a Bernoulli distribution whereas the satellites follow a Poisson distribution.

The vanilla HOD model has been extensively used to describe magnitude-limited galaxy samples (Z. Zheng et al. 2007). It parameterizes the mean galaxy occupation as

$$\bar{n}_{\rm cent}(M) = \frac{f_{\rm ic}}{2}\,{\rm erfc}\left[\frac{\log_{10}(M_{\rm cut}/M)}{\sqrt{2}\,\sigma}\right], \tag{2}$$

$$\bar{n}_{\rm sat}(M) = \left[\frac{M - M_0}{M_1}\right]^{\alpha}, \tag{3}$$

where the five vanilla parameters characterizing the model are $M_{\rm cut}$, M_1, σ, α, and M_0, augmented by a central galaxy incompleteness parameter $f_{\rm ic}$, defined to be $0 < f_{\rm ic} \leqslant 1$. $M_{\rm cut}$ characterizes the minimum halo mass to host a central galaxy, and σ describes the steepness of the transition from 0 to $f_{\rm ic}$ in the number of central galaxies. M_0 gives the minimum halo mass to host a satellite galaxy, which is also commonly parameterized as $\kappa M_{\rm cut} = M_0$; M_1 characterizes the typical halo mass that hosts one satellite galaxy; and α is the power-law index on the satellite galaxy occupation.

Both HOD models slightly deviate from this vanilla form. HOD1 fixes $f_{\rm ic} = 1$, effectively requiring all massive halos to host a central galaxy. HOD2 varies $f_{\rm ic}$ implicitly by rescaling the predicted number density to match the observed number density. HOD2 also adds a modulation term $\bar{n}_{\rm cent}(M)$ to the satellite occupation function to largely remove satellites from halos without centrals.

The two models also differ in their ways of determining galaxy positions and velocities once the number of galaxies per halo is computed. In both models, the position and velocity of the central galaxy are set to be the same as those of the halo center. However, for satellite galaxies HOD1 assigns the satellite galaxy positions within the virial radius of the halo following a Navarro–Frenk–White (NFW) profile (J. F. Navarro et al. 1997), while the velocities are solved from Jeans equations. In HOD2, the satellite galaxies are randomly assigned to halo particles with uniform weights, each satellite inheriting the position and velocity of its host particle.

Both models also augment the vanilla HOD by adding additional flexibilities (see S. Yuan et al. 2022b; C. Hahn et al. 2023b, for implementation details):

1. *Velocity bias*. The galaxy velocities may not perfectly follow the dark matter halo and particle velocities. HOD1 rescales galaxy velocities with parameters $\eta_{\rm cen}$ and $\eta_{\rm sat}$, which set the velocity dispersions of central and satellite galaxies relative to halo velocity dispersion: $\sigma_{\rm cen} = \eta_{\rm cen}\sigma_{\rm cen}$ and $\sigma_{\rm sat} = \eta_{\rm sat}\sigma_{\rm sat}$. In HOD2, we define velocity bias parameters $\alpha_{\rm vel,c}$ and $\alpha_{\rm vel,s}$ to modulate the peculiar velocities of the central and satellite galaxies with respect to the host halo center, respectively. In this definition, $\alpha_{\rm vel,c} = 0$ and $\alpha_{\rm vel,s} = 1$ indicates no velocity bias.
2. *Galaxy assembly bias*. Galaxy occupation can also depend on secondary halo properties beyond halo mass, a phenomenon commonly referred to as galaxy assembly bias or galaxy secondary bias (see R. H. Wechsler & J. L. Tinker 2018, for a review). HOD1 adds the halo concentration as a secondary dependency in galaxy occupation via the Heaviside assembly bias model, which is described in detail in A. P. Hearin et al. (2016). HOD2 offers the option of using either the halo concentration or the halo environment in a 5 h^{-1} Mpc filter as the secondary dependency. Both models adopt two assembly bias parameters to modulate the central and satellite occupations separately.
3. *Baryonic effects*. Both HOD implementations account for baryonic effects by modulating the radial distribution of satellite galaxies relative to the halo density profile. HOD1 includes a parameter $\eta_{\rm conc}$ that sets the ratio between the concentration of satellite and halo profile. HOD2 includes a parameter s that modulates the radial satellite galaxy profile, with $s = 0$ indicating no radial bias, $s > 0$ indicating a more extended (less concentrated) profile of satellites relative to the halo, and vice versa for negative s.

Note that the three HOD-based analyses all employ different priors for the HOD parameters, motivated by sensitivity studies specific to each statistic or based on conservative estimates, as detailed in their respective analysis sections.

5.1. BACCO Hybrid Emulator[40]

In this subsection, we describe the results of the Beyond-2pt challenge obtained by the BACCO-hybrid emulator approach. This emulator was presented in M. Zennaro et al. (2023) and M. Pellejero Ibañez et al. (2023) and has been thoroughly tested on SHAM extended (SHAMe) techniques (S. Contreras et al. 2021a, 2021b) and survey-based HOD techniques (Euclid Collaboration et al. 2024; A. Nicola et al. 2024). The model has been further extended to study intrinsic alignments in F. Maion et al. (2024) and to generate field-level predictions in M. Pellejero Ibañez et al. (2024).

5.1.1. Data and Estimators

This analysis focuses on the power spectrum $P(k)$ for the ΛCDM boxes in real space and the first nonzero multipoles ($\ell = 0, 2, 4$) of the power spectrum $P_\ell(k)$ for the redshift-space ΛCDM boxes, shown in Figure 6. We measure these power spectra by mapping the galaxy positions into a mesh of 1024^3 cells and then perform a fast Fourier transform, as usual,

[40] Authors: M. Pellejero Ibañez, R. E. Angulo, M. Zennaro.

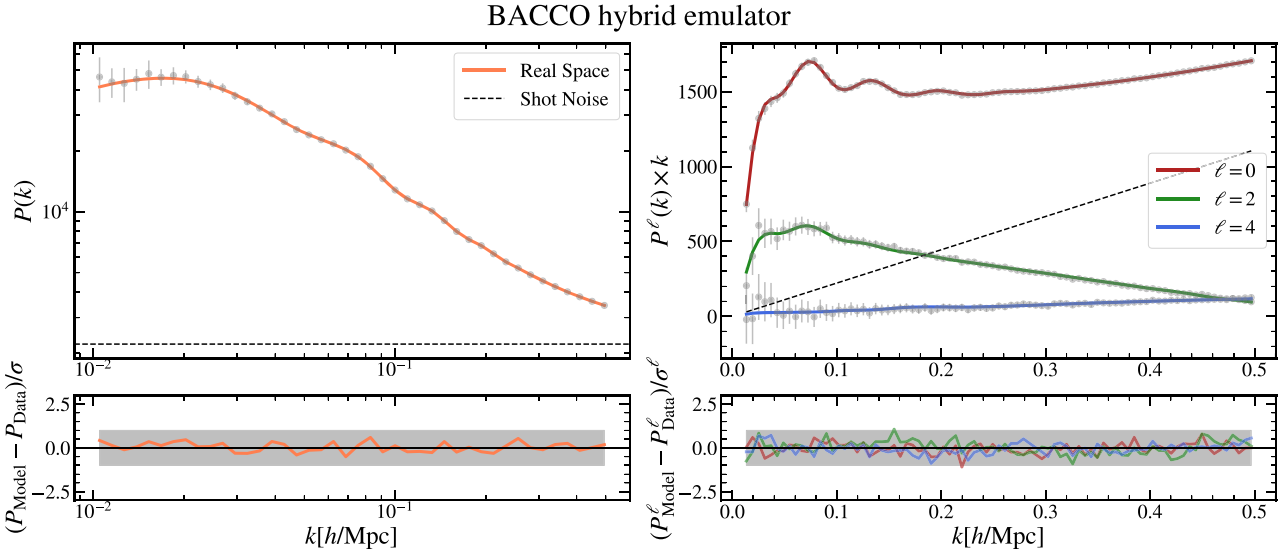

Figure 6. Left panel: power spectrum fit of the BACCO hybrid emulator model to the mean of the 10 ΛCDM mock boxes. The error bars correspond to the Gaussian approximation of the covariance matrix at the volume of one of the boxes. Right panel: same as the left panel, but for the 10 ΛCDM mock boxes in redshift space. We show the three multipoles used in this work.

applying interlacing and "triangular shape cloud" as the deposit method.

5.1.2. Model

The modeling of the clustering statistics in the so-called "hybrid" approaches (C. Modi et al. 2020; M. Pellejero Ibañez et al. 2022) contains two ingredients: (i) a map from Lagrangian, $\boldsymbol{q}$, to Eulerian, $\boldsymbol{x}$, space using the displacement field as measured on N-body simulations, $\boldsymbol{\psi}(\boldsymbol{q})$; and (ii) a functional relation between the matter and galaxy density fields, i.e., a bias model, $F(\delta_{\rm L}(\boldsymbol{q}))$.

N-*body Displacements*. Regarding the first ingredient, we can define completely the Eulerian overdensity field as

$$1 + \delta_{\rm tr}(\boldsymbol{x}) = \int d^3q\; w(\boldsymbol{q})\; \delta_D(\boldsymbol{x} - \boldsymbol{q} - \boldsymbol{\psi}(\boldsymbol{q})). \tag{4}$$

For the BACCO hybrid model $\boldsymbol{\psi}(\boldsymbol{q})$ is measured in N-body simulations (as opposed to PT) by comparing the Lagrangian position of simulation particles with their position in a snapshot at any desired redshift.

Galaxy–Halo Connection. We employ a second-order Lagrangian bias model (T. Matsubara 2008; V. Desjacques et al. 2018):

$$\begin{aligned} \mathrm{w}(\boldsymbol{q}) = F(\delta_{\rm L}(\boldsymbol{q})) = 1 &+ b_1\delta_{\rm L}(\boldsymbol{q}) + b_2(\delta^2_{\rm L}(\boldsymbol{q}) - \langle\delta^2_{\rm L}(\boldsymbol{q})\rangle) \\ &+ b_s(s^2(\boldsymbol{q}) - \langle s^2(\boldsymbol{q})\rangle) + b_\nabla\nabla^2\delta_{\rm L}(\boldsymbol{q}). \end{aligned} \tag{5}$$

Here $\delta_L(\boldsymbol{q})$ stands for the linear field and s^2 is the traceless part of the tidal field, $s^2 = s_{ij}s^{ij} = (\partial_i\partial_j\phi(\boldsymbol{q}) - 1/3\delta^{\rm K}_{ij}\delta_{\rm L}(\boldsymbol{q}))^2$, with $\phi(\boldsymbol{q})$ the linear gravitational potential. The function $\mathrm{w}(\boldsymbol{q})$ weighs the importance of different Lagrangian fields in representing the density of tracers at a given $\boldsymbol{x}$.

Redshift Space. We base our redshift-space modeling on our work in M. Pellejero Ibañez et al. (2022). In order to account for this effect, we substitute $\boldsymbol{\psi}(\boldsymbol{q})$ with $\boldsymbol{\psi}^s(\boldsymbol{q})$ accounting for the shifts in the line-of-sight (LOS) direction due to the velocity field as follows:

$$\boldsymbol{\psi}^s(\boldsymbol{q}) = \boldsymbol{\psi}(\boldsymbol{q}) + \frac{\hat{\boldsymbol{q}}_z \cdot \boldsymbol{v}_{\rm tr}(\boldsymbol{q})}{aH}\,\hat{\boldsymbol{q}}_z, \tag{6}$$

where we define the velocity $\boldsymbol{v}_{\rm tr}$ based on the velocities of the N-body simulation. Concretely, $\boldsymbol{v}_{\rm tr}(\boldsymbol{x}) = \boldsymbol{v}(\boldsymbol{x})$, the velocity of the matter particle, if the tracer is outside of a halo, and $\boldsymbol{v}_{\rm tr}(\boldsymbol{x}) = \boldsymbol{v}_{\rm halo}(\boldsymbol{x})$ if the tracer is inside of a halo. This distortion roughly accounts for the so-called "Kaiser effect" (N. Kaiser 1987) but also incorporates additional contributions resulting from the nonlinearity of the halo velocity. We can model the effect of the intracluster velocities on the redshift-space galaxy field by applying a convolution along the LOS direction $\delta^s_{\rm tr}(\boldsymbol{s}) *_z [(1 - f_{\rm sat})\delta_D(s_z) + f_{\rm sat}\exp(-\lambda_{\rm FoG}s_z)]$, where $*_z$ represents the convolution along the LOS z.

Emulator. The BACCO hybrid emulator employed a suite of high-resolution simulations (first introduced in R. E. Angulo et al. 2021) together with cosmology rescaling (R. E. Angulo & S. D. M. White 2010) to densely sample a target cosmological parameter space. Then, we compute the Eulerian fields weighted by their corresponding Lagrangian bias fields. We estimate their power and cross-power spectra, and finally, we use these data to train a neural network. As a result, this provides accurate and extremely fast predictions of the 2pt statistics, which makes it possible to use our model in cosmological data analyses, as presented in M. Pellejero Ibañez et al. (2023). At this point, we further include two noise terms to account for shot noise to the power spectrum in the form of Noise $= 1/\bar{n}(\epsilon_1 + \epsilon_2 k^2)$. Note that in M. Pellejero Ibañez et al. (2023) we tested the need to include μ-dependent stochastic terms, finding them not required for the range of scales explored in this analysis. This does not imply that the term is absent; rather, it indicates that the velocity from halos in the simulation, combined with the Finger-of-God (FoG) free parameters, accounts for this dependency. We further tested whether the inclusion of a $k^2\mu^2$ dependency in the noise changed our results, finding a negligible impact.

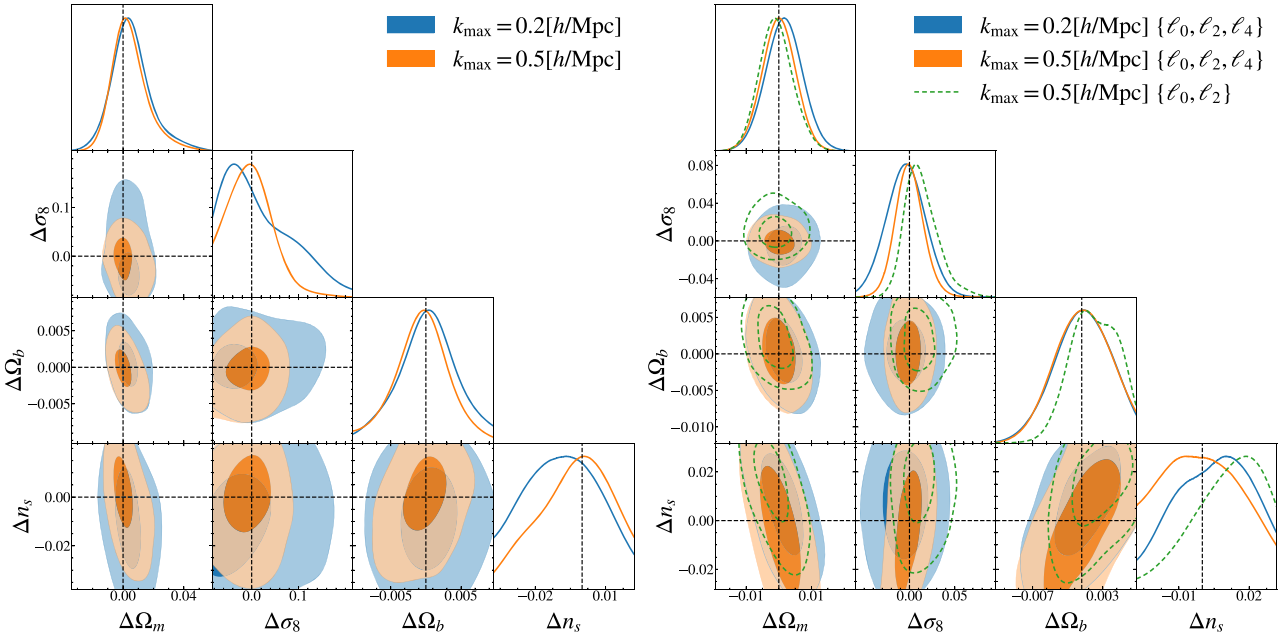

Figure 7. Left panel: posterior distribution inferred by the BACCO hybrid emulator model to the mean of the 10 ΛCDM mock boxes in real space as a function of $k_{\rm max}$. Right panel: same as the left panel, but for the 10 ΛCDM mock boxes in redshift space. We have displaced values by subtracting the mean at the $k_{\rm max} = 0.5[h\ {\rm Mpc}^{-1}]$ result to keep the inferred cosmology masked.

Another important aspect is how we address the binning discreteness effect caused by a finite fundamental k. To mitigate this effect, we apply a binning correction for each k mode and each multipole. This is done by evaluating the theoretical power spectrum on the same mesh and using the resulting power spectrum divided by the nonbinned one.

5.1.3. Inference

Priors. Our cosmological parameter priors are defined by the limits of the emulator. They are given by

$$\begin{aligned} \Omega_{\rm m} &\sim \mathcal{U}[0.23, 0.4], \quad & \sigma_{8,{\rm c}} &\sim \mathcal{U}[0.65, 0.9], \\ \Omega_{\rm b} &\sim \mathcal{U}[0.04, 0.06], \quad & n_{\rm s} &\sim \mathcal{U}[0.92, 1.01]. \end{aligned} \tag{7}$$

All beyond ΛCDM parameters are set to their default values. Specifically, neutrino mass is set to zero, and evolving dark energy parameterizations are not taken into account. For bias parameters, we use uninformative priors, given by

$$\begin{aligned} b_1 &\sim \mathcal{U}[0, 2], \quad & b_2 &\sim \mathcal{U}[-2, 2], \\ b_s &\sim \mathcal{U}[-3, 3], \quad & b_\nabla &\sim \mathcal{U}[-6, 6], \\ \lambda_{\rm FoG} &\sim \mathcal{U}[0, 1], \quad & f_{\rm sat} &\sim \mathcal{U}[0, 1], \\ \epsilon_1 &\sim \mathcal{U}[0, 2], \quad & \epsilon_2 &\sim \mathcal{U}[-4, 4]. \end{aligned} \tag{8}$$

Covariance. We make use of a Gaussian covariance. In real space, this is diagonal and proportional to the amplitude of $P(k)$ and shot noise. In redshift space, it is block diagonal in the multipoles, with correlations given by the Wigner symbols. We are aware that this covariance is not accurate when pushing to scales as small as $k \approx 0.5\ h\,{\rm Mpc}^{-1}$ since nonlinearities are not taken into account. However, for this challenge, the number density of mocks is such that the shot noise dominates the signal at scales of $k \approx 0.3\ h\,{\rm Mpc}^{-1}$. This shot-noise contribution will dominate at scales where nonlinearities become relevant, making their contribution less important (see, e.g., L. Blot et al. 2019; D. Wadekar & R. Scoccimarro 2020). Nevertheless, an extended study on the impact of nonlinear and non-Gaussian covariance terms will be of interest for dense or highly biased galaxy samples going forward.

Likelihood. Due to the central limit theorem, it is safe to assume a Gaussian likelihood shape for the power spectrum in this case.

Sampling and Validation. We make use of the public code MULTINEST[41] Bayesian inference tool for recovering credibility intervals (see F. Feroz & M. P. Hobson 2008; F. Feroz et al. 2009, 2019, for more details). The best-fit values are also extracted from the maximum likelihood values of these chains. We set the number of live points to 1200 and the evidence tolerance to 0.08.

5.1.4. Analysis Choices

Scale Cuts. We analyze real- and redshift-space power spectra down to nonlinear scales of $k_{\rm max} = 0.5\ h\,{\rm Mpc}^{-1}$, motivated by our findings in previous works (M. Zennaro et al. 2022; M. Pellejero Ibañez et al. 2022, 2023). Specifically, in M. Zennaro et al. (2022) we created thousands of SHAMe mocks with different physical parameters to validate the model and to compute physical priors on the bias parameters, including a mock that closely resembles the hydrodynamical simulation IllustrisTNG galaxy clustering (S. Contreras et al. 2023b). These scale cuts pass all the parameter drift tests described in the *Unmasking Criteria* subsection and shown in Figure 7.

Unmasking Criteria. We follow the following main criteria before unmasking to assert the robustness of our measurements:

1. We verify that the reduced χ^2 value of the best-fit value is ≈0.1 for the ΛCDM mock sample mean. The value $\chi^2_{\rm red} = 0.1$ is determined for the real-space mocks, while the value $\chi^2_{\rm red} = 0.12$ was determined for the redshift-space mocks.

[41] Available at https://github.com/farhanferoz/MultiNest.

2. We determine whether any of the parameter posteriors intersect the priors' limits. This is remedied by extending the prior range for the nuisance parameters. In the case of cosmological parameters, however, our emulator provides the priors. For example, we discovered that the σ_8–b_1 degeneracy broadens the constraints and partially affects the real-space emulator priors in the $2\sigma-3\sigma$ region. However, the $1\sigma-2\sigma$ region is well within our emulator's priors, so we believe these results to be reliable.
3. We investigate the consistency of the inferred parameters with respect to various $k_{\rm max}$ values. In particular, we compute the evolution of the posterior distribution at $k_{\rm max} = \{0.1, 0.2, 0.3, 0.4, 0.5\}\, h\,{\rm Mpc}^{-1}$ and find no discernible change, as illustrated in Figure 7.
4. We test for redshift-space analysis whether adding or removing multipoles gives consistent cosmological parameter constraints. We find no tensions between results containing only monopole, monopole plus quadrupole, or monopole plus quadrupole and hexadecapole.
5. We include our theory error budget based on the emulator uncertainties. This is discussed in Appendix A of M. Pellejero Ibañez et al. (2023). For the test, we add these error estimates in quadrature to the covariance matrix diagonal elements. We find no strong dependence of the recovered cosmological marginalized values on this theory error budget. We note that the theory error should always be included; however, if this error is small compared to the data error, its effect becomes negligible.

Caveats. We caution that several assumptions could potentially bias our constraints: The first is the choice of a Gaussian covariance with no off-diagonal elements. The second is the possible underestimation of the emulator errors at the cosmology of interest. The third is the priors of the emulator. Even though we expect the true cosmologies to lie within these priors, strong degeneracies, such as the σ_8–b_1 direction in real space, might make our priors too informative. The fourth is possible model simplifications. To construct the BACCO emulator hybrid model, we make use of several approximations in the galaxy-to-matter and galaxy velocity assignments. Specifically, the current galaxy velocity model lacks implementation of the mass-dependent FoG effect and strong velocity bias. These model misspecifications might bias our cosmological parameter estimations.

Post-unmasking Studies. After unmasking, we found that our results for σ_8 exhibited a 1σ shift compared to the true values. This is not entirely unexpected, as cosmic variance can introduce shifts in the contours, and locating them within a $1\sigma-3\sigma$ range does not imply a failure of the model. Additionally, our assumption of a diagonal Gaussian covariance matrix exacerbates this issue by ignoring nondiagonal correlation terms that arise on small scales.

However, to mitigate the impact of cosmic variance, parameter estimation was performed using the mean of 10 independent mocks. Naively, this should reduce the contour size by a factor of approximately $\sqrt{10} \sim 3$. Taking this into account, our contours appear to be at $\sim 3\sigma$ from the true values. To confirm this, it would be necessary to reanalyze the data with reduced error bars corresponding to a volume 10 times larger than each individual mock. Unfortunately, all tests conducted on the BACCO hybrid model in redshift space thus far have been based on the assumption of a BOSS-like sample (M. Pellejero Ibañez et al. 2023) with an effective volume of $V_{\rm eff} = 2.8\,{\rm Gpc}^3 \sim 0.9\,{\rm Gpc}^3\ h^{-3}$, and our emulator's noise level remains too high for the galaxy sample and survey volume of this challenge.

Nevertheless, we can investigate whether incorporating or removing theory errors impacts our predictions. As previously mentioned, we already conducted this test by including errors in quadrature. Based on M. Pellejero Ibañez et al. (2023), we are aware that typical uncertainties in the emulator account for $\sim$0.5% in the monopole, $\sim$1% in the quadrupole, and $\sim$10% in the hexadecapole amplitudes. During this examination, we did not realize the low values of the hexadecapole, as depicted in Figure 7. Consequently, the theory error we included in the noisiest of our estimations was underestimated.

Therefore, we decided to present the measurement of cosmological parameters without the hexadecapole, as shown in dashed lines in Figure 7. The shift found is of around half a sigma, and the slight decrease in constraining power deems our results unbiased with respect to the true values. It is important to reiterate that this is a post-unmasking finding, and further tests on the model are required to rule out potential failures when studying the HOD models employed in this work.

5.2. EFT P+B: Analysis Method[42]

In this subsection, we describe the analysis of the Beyond-2pt mocks with the EFT of large-scale structure, focusing on the large-scale power spectrum and bispectrum. The key theoretical underpinnings of this are described, e.g., in D. Baumann et al. (2012), J. J. M. Carrasco et al. (2012), and M. M. Ivanov (2023),[43] with recent applications to galaxy surveys and other large-scale structure data shown in, e.g., M. M. Ivanov et al. (2020a, 2020b), G. D'Amico et al. (2020), O. H. E. Philcox & M. M. Ivanov (2022), and S.-F. Chen et al. (2022). We apply this methodology to all challenge catalogs.

Information Sources. We briefly discuss the various sources of information in the full shape of the galaxy power spectrum, which set the parameter degeneracy directions. More details on the sources of information can be found in M. M. Ivanov et al. (2020a). From the shape of the galaxy power spectrum, one can determine $\omega_{\rm b}$, $\omega_{\rm m} \equiv \Omega_{\rm m} h^2$, and $n_{\rm s}$ regardless of the distance to the sample. These parameters then predict absolute scales such as that of the matter−radiation equality and the sound horizon ($k_{\rm eq}$ and $k_{\rm BAO}$, respectively).

In contrast, distance information is encoded in the angular scales $\theta_{\parallel}$ and $\theta_{\perp}$, for both BAO and the matter−radiation equality. This constrains the Hubble parameter h and, if multiple redshifts are available, $\Omega_{\rm m}$ through the growth rate evolution. For periodic boxes, there is no distance information, since all quantities are measured in h^{-1} Mpc units; thus, all distances are equivalent to H_0^{-1}.

From growth, we constrain $b_1\sigma_8(z_{\rm data})$ from the real-space power spectrum or $b_1^3\sigma_8^4(z_{\rm data})$ from the bispectrum. Loop corrections additionally yield $b_1^2\sigma_8^4(z_{\rm data})$, and infrared resummation directly measures $\sigma_8(z_{\rm data})$ (see T. Baldauf et al. 2015; L. Senatore & M. Zaldarriaga 2015; D. Blas et al. 2016b, 2016a; M. M. Ivanov & S. Sibiryakov 2018; A. Vasudevan et al. 2019), though these effects are comparatively small at the high redshifts of this challenge. In redshift space we instead measure $b_1\sigma_8(z_{\rm data})$ from the monopole and $f\sigma_8(z_{\rm data})$ from the quadrupole and

[42] Authors: M. M. Ivanov, O. Philcox, K. Akitsu, G. Cabass.
[43] See footnote 6 of R. C. Nunes et al. (2022) for an extended set of references.

hexadecapole power spectrum moments. Loops and the bispectrum monopole help break degeneracies and give directly $f(z_{\rm data})$ and $\sigma_8(z_{\rm data})$.

5.2.1. Data and Estimators

In all cases, we analyze the large-scale power spectrum and bispectrum extracted from the mock catalogs. We use $P(k)$ and $B(k_1, k_2, k_3)$ in real space; in redshift space we consider the power spectrum monopole, quadrupole, and hexadecapole (P_0, P_2, and P_4) but restrict to the bispectrum monopole (since higher moments do not add significant signal; see M. M. Ivanov et al. 2023). For the redshift-space mocks, we supplement the data vector with the real-space power spectrum proxy, Q_0, which gives additional information on smaller scales without bias from FoGs (M. M. Ivanov et al. 2022b; see also R. Scoccimarro 2004; G. D'Amico et al. 2024c).

For the periodic box mocks, correlators are estimated using fast Fourier transforms, as usual (R. Scoccimarro 2015). For the light-cone mock catalog, we use window-free estimators (O. H. E. Philcox 2021a, 2021b), as implemented in the Spectra-Without-Windows code,[44] making use of the mask and random file. In this case, we split the sample into two redshift bins of equal density, with effective redshifts $z_1 = 0.92$ and $z_2 = 1.17$. Knowledge of redshift evolution is necessary to break the geometric degeneracy.

5.2.2. Model

Perturbation Theory. We model the power spectrum and bispectrum with the EFT of large-scale structure, as implemented in `CLASS-PT` (A. Chudaykin et al. 2020).[45] The power spectrum multipoles are modeled using the (infrared-resummed) one-loop theory (M. M. Ivanov et al. 2020a), and we use tree-level theory for the bispectrum (M. M. Ivanov et al. 2022a), noting that higher loops give limited gains (O. H. E. Philcox et al. 2022; G. D'Amico et al. 2024b).

Galaxy–Matter Connection. We assume the following bias expansion up to third-order (renormalized) operators (M. M. Ivanov et al. 2020a):

$$\delta_g = b_1\delta + \frac{1}{2}b_2\delta^2 + b_{\mathcal{G}_2}\mathcal{G}_2 + b_{\Gamma_3}\Gamma_3, \tag{9}$$

where δ_g and δ are the galaxy and matter overdensities, respectively, and $\mathcal{G}_2$ and Γ_3 are Galileon tidal operators. The model additionally includes stochastic contributions from shot noise $P_{\rm shot}$, $A_{\rm shot}$, $B_{\rm shot}$, a_0, and a_2 (including k^2 scale dependence) and counterterms $\{c_0, c_2, c_4, \tilde{c}\}$, encapsulating small-scale physics such as halo formation and velocity effects. This makes no assumptions on the form of the galaxy–halo connection, except that it is statistically isotropic and homogeneous on large scales and obeys Einstein's equivalence principle (V. Desjacques et al. 2018).

5.2.3. Inference

Priors. We assume the following priors on cosmological parameters for all analyses:

$$\begin{aligned} &\omega_{\rm cdm} \sim \mathcal{U}[-\infty, \infty], \quad 10^9 A_{\rm s} \sim \mathcal{U}[0.5, 5], \\ &n_{\rm s} \sim \mathcal{U}[0.87, 1.07], \quad \omega_{\rm b} \sim \mathcal{U}[0.01, 0.035], \end{aligned} \tag{10}$$

with the angular size of the sound horizon θ_* fixed to the value known to all participants. We fix the neutrino mass to zero and assume a flat Universe in all cases. For bias parameters, we use weakly informative priors, given by O. H. E. Philcox & M. M. Ivanov (2022):

$$\begin{aligned} &b_1 \in \text{flat}[0, 4], \quad b_2 \sim \mathcal{N}(0, 1^2), \\ &b_{\mathcal{G}_2} \sim \mathcal{N}(0, 1^2), \quad b_{\Gamma_3} \sim \mathcal{N}\left(\frac{23}{42}(b_1 - 1), 1^2\right), \end{aligned} \tag{11}$$

where $\mathcal{N}(\mu, \sigma^2)$ indicates a Gaussian distribution with mean μ and variance σ^2. Note that these priors are consistent with the recent measurements (M. M. Ivanov et al. 2024), as well as earlier results (M. M. Abidi & T. Baldauf 2018; T. Lazeyras & F. Schmidt 2018). Similarly, we use Gaussian priors for the counterterms c_0, c_2, c_4, c_1, $\tilde{c}$ and stochasticity parameters a_0, a_2, $P_{\rm shot}$, $A_{\rm shot}$, $B_{\rm shot}$:

$$\begin{aligned} &\frac{c_0}{[\text{Mpc}/h]^2} \sim \mathcal{N}(0, 30^2), \quad \frac{c_2}{[\text{Mpc}/h]^2} \sim \mathcal{N}(30, 30^2), \\ &\frac{c_4}{[\text{Mpc}/h]^2} \sim \mathcal{N}(0, 30^2), \quad \frac{\tilde{c}}{[\text{Mpc}/h]^4} \sim \mathcal{N}(500, 500^2), \\ &\frac{c_1}{[\text{Mpc}/h]^2} \sim \mathcal{N}(0, 51^2), \quad P_{\rm shot} \sim \mathcal{N}(0, 1^2), \\ &a_0 \sim \mathcal{N}(0, 1^2), \quad a_2 \sim \mathcal{N}(0, 1^2), \\ &B_{\rm shot} \sim \mathcal{N}(1, 1^2), \quad A_{\rm shot} \sim \mathcal{N}(0, 1^2), \end{aligned} \tag{12}$$

where we use the same convention[46] as M. M. Ivanov et al. (2022a; see also O. H. E. Philcox & M. M. Ivanov 2022; O. H. E. Philcox et al. 2022), but with $\bar{n} \approx 5 \times 10^{-4}\, h^3\,\text{Mpc}^{-3}$ of the challenge boxes.

Covariance. We assume a Gaussian covariance matrix for the power spectrum multipoles and bispectrum monopole, defined explicitly in A. Chudaykin & M. M. Ivanov (2019) and M. M. Ivanov et al. (2022a, 2023). This is exact in the linear regime (where modes are uncorrelated) and found to be highly accurate on our scales of interest, due to the high shot noise and limited cosmological information available at high k (D. Wadekar et al. 2020). The covariance is diagonal in k but includes correlations between different power spectrum multipoles, where necessary. We do not include cross-covariance between the power spectrum and bispectrum, since this is formally of higher order in EFT and found to be unnecessary in M. M. Ivanov et al. (2022a) for a similar choice of scale cuts.

The power spectrum covariance is computed using the specific realizations of the power spectrum multipoles (since these are measured at high significance). For the bispectrum covariance, we adopt an iterative procedure, first computing the covariance using a fiducial cosmology and then updating with

[44] Available at https://github.com/oliverphilcox/Spectra-Without-Windows.
[45] Available at https://github.com/michalychforever/CLASS-PT.
[46] Note that O. H. E. Philcox & M. M. Ivanov (2022) and O. H. E. Philcox et al. (2022) used 2 standard deviations for stochastic parameters of the high-z samples because their physical number density $\bar{n}$ was twice as low as the fiducial one; see the public likelihoods for more information.

the best-fit parameters from a likelihood analysis. This procedure is found to converge quickly.

Likelihood. On large scales, the data are well described by a Gaussian likelihood. This holds by the central limit theorem and also follows from perturbation theory. The covariance can be computed precisely on large scales, and the number of bins is reasonable.

Sampling and Validation. We sample the various parameters with `MontePython` (T. Brinckmann & J. Lesgourgues 2019), analytically marginalizing over nuisance parameters that enter the model linearly (O. H. E. Philcox et al. 2021a). This model has been validated on several suites of simulations, including Multi-Dark-Patchy, Las Damas, the Perturbation Theory Challenge, Nseries, and Outer Rim (M. M. Ivanov et al. 2020a, 2022b, 2023; T. Nishimichi et al. 2020; M. M. Ivanov 2021; O. H. E. Philcox & M. M. Ivanov 2022; O. H. E. Philcox et al. 2022; A. Chudaykin & M. M. Ivanov 2023). We do not include a theoretical error covariance (due to our choice of scale cuts), but we note that this can be, in principle, incorporated (T. Baldauf et al. 2016a; A. Chudaykin & M. M. Ivanov 2019; A. Chudaykin et al. 2021b). We consider our Markov Chain Monte Carlo (MCMC) chains converged when they satisfy the Gelman–Rubin (G-R) criterion $R-1<0.03$.

5.2.4. Analysis Choices

Scale Cuts. The EFT P+B analyses adopt the following scale cuts (in $h\,\mathrm{Mpc}^{-1}$ units):

1. real-space mock (periodic box, $z=1$, $V=8\ (\mathrm{Gpc}/h)^3\ \times 10$): $k^P\leqslant 0.3$, $k^B\leqslant 0.15$;
2. redshift-space mock (periodic box, $z=1$, $V=8\ (\mathrm{Gpc}/h)^3\ \times 10$): $k^{P_\ell}\leqslant 0.20$, $0.20<k^{Q_0}\leqslant 0.4$, $k^{B_0}\leqslant 0.08$;
3. light-cone mock ($z=[0.8,\ 1.3]$, $V=5.3\ (\mathrm{Gpc}/h)^3$ (for our fiducial cosmology $\Omega_\mathrm{m}=0.31$, $h=0.676$)): $k^{P_\ell}\leqslant 0.25$, $0.25<k^{Q_0}\leqslant 0.4$, $k^{B_0}\leqslant 0.08$.

These are motivated by previous works and tests on large-volume simulations (M. M. Ivanov et al. 2020a, 2022a, 2022b; T. Nishimichi et al. 2020; O. H. E. Philcox & M. M. Ivanov 2022) and validated for this challenge through parameter drift plots for all challenge mocks. As an example, we show parameter posteriors for the redshift-space mock for four different choices of $k^{P_\ell}_\mathrm{max}=0.15,\ 0.2,\ 0.25,\ 0.3\ h\,\mathrm{Mpc}^{-1}$ in Figure 8. The parameter drifts in this plot suggest that results at $k^{P_\ell}_\mathrm{max}=0.3\ h\,\mathrm{Mpc}^{-1}$ are biased, because one can clearly see an upward shift of σ_8 by 1σ, consistent with the effects of two-loop corrections previously observed (e.g., T. Nishimichi et al. 2020; A. Chudaykin et al. 2021b). The posteriors for the smaller-scale cuts appear consistent with each other, and we select $k^{P_\ell}_\mathrm{max}=0.2\ h\,\mathrm{Mpc}^{-1}$ as a baseline choice in order to be conservative. The same methodology is applied for the light-cone and real-space analyses, i.e., studying parameter drifts plus using estimates for the theoretical error based on the two-loop estimates (see also A. Chudaykin et al. 2021b; M. M. Ivanov et al. 2022a, for studies of real-space clustering in EFT). Using this methodology, all challenge mock catalogs were systematically analyzed and tested (see consistency checks below). The particular analysis choices applied here were also cross-validated on other mock catalogs such as Nseries, PTchallenge, and Las Damas. We choose a more aggressive scale cut for the light cones, as their volume is small compared to the periodic box simulations, and hence the theory systematic error is smaller than cosmic variance down to smaller scales. In this case, the parameter drift plot is also less conclusive because it is hard to disentangle parameter shifts owing to bias from those resulting from a statistical fluctuation due to the addition of new data at a higher k_max.

We also note that the bispectrum is restricted to larger scales owing to the tree-level modeling. The lower k_max for redshift-space snapshots (compared to real-space ones) is also supported by expectations of the impact of velocity effects that become nonperturbative at larger scales (noting that the nonlinear scale is $k_\mathrm{NL}\sim\sigma^{-1}_\mathrm{FoG}$ in redshift space, as opposed to R^{-1}_halo in real space, if halo formation effects dominate; see M. M. Ivanov 2021; M. M. Ivanov et al. 2022b for detailed discussions).

Consistency Checks. We have run a number of consistency checks that include variation of the bispectrum covariance matrix; stability w.r.t. scale cut changes; exclusion of Q_0, B_0, P_4 from the data vector; and the change of fiducial cosmology in the case of the light-cone mock. We have not found any significant effect when perturbing our baseline choices within a reasonable range.

As far as the choice of fiducial cosmology is concerned, we have explicitly tested that using $\Omega_\mathrm{m}=0.31$ (our baseline choice) or Ω_m from the best fit to the data has negligible difference on the extracted cosmological parameters, even when Ω_m is actually significantly different (by up to $\sim$30%) from its fiducial value 0.31. In this particular example of a 30% difference (which constitutes $\simeq 10\sigma_{\Omega_m}$), the shift in the mean of Ω_m was found to be $0.1\sigma_{\Omega_m}$, while the shift in σ_8 was found to be $0.3\sigma_{\sigma_8}$. This implies that the standard AP parameterization is quite accurate even for large deviations from the fiducial cosmological parameter values.

Unmasking Criteria. We allow our results to be unmasked after visual confirmation of the relevant posteriors and their degeneracy directions, as well as simple tests based on the goodness of fit and analysis of variation with k_max.

For the real-space and redshift-space mocks we additionally analyzed the mean data vectors (from the 10 realizations) with the covariances that correspond to the full volume of 10 boxes. We found some nonnegligible bias for $k^{P_\ell}_\mathrm{max}=0.25\ h\,\mathrm{Mpc}^{-1}$ for the redshift-space mock in this analysis, which was another reason to stick to our baseline choice $k^{P_\ell}_\mathrm{max}=0.2\ h\,\mathrm{Mpc}^{-1}$. For the light-cone mock, however, we only have one realization for the volume that is approximately 40% smaller than that of one snapshot. The two-loop systematic error is smaller than the cosmic variance in this case even at $k^{P_\ell}_\mathrm{max}=0.25\ h\,\mathrm{Mpc}^{-1}$, and therefore we could adopt this more optimistic scale cut in our analysis.

Caveats. Finally, we caution that failure modes of our analysis could include overly optimistic scale cuts (e.g., due to extremely large FoG effects), cosmological or nuisance parameters lying outside the parameter ranges, or galaxy–halo connections violating our symmetry assumptions. The posteriors should be interpreted with caution, as for many cases (e.g., real-space analyses) their shape is highly non-Gaussian. This may lead to parameter projection effects, i.e., 1D marginalized contours may be off owing to marginalization over non-Gaussian posteriors, as was discussed in detail, e.g., in M. M. Ivanov et al. (2020a, 2023), A. Chudaykin et al. (2021a), and O. H. E. Philcox & M. M. Ivanov (2022). Note that our choice of $k_\mathrm{max}=0.3\ h\,\mathrm{Mpc}^{-1}$ for the real-space analyses is somewhat aggressive (done in part in order to facilitate the sampling), so we expect (based on our tests on mock data) that the theory systematic error may be as large as $\sim 1\sigma$.

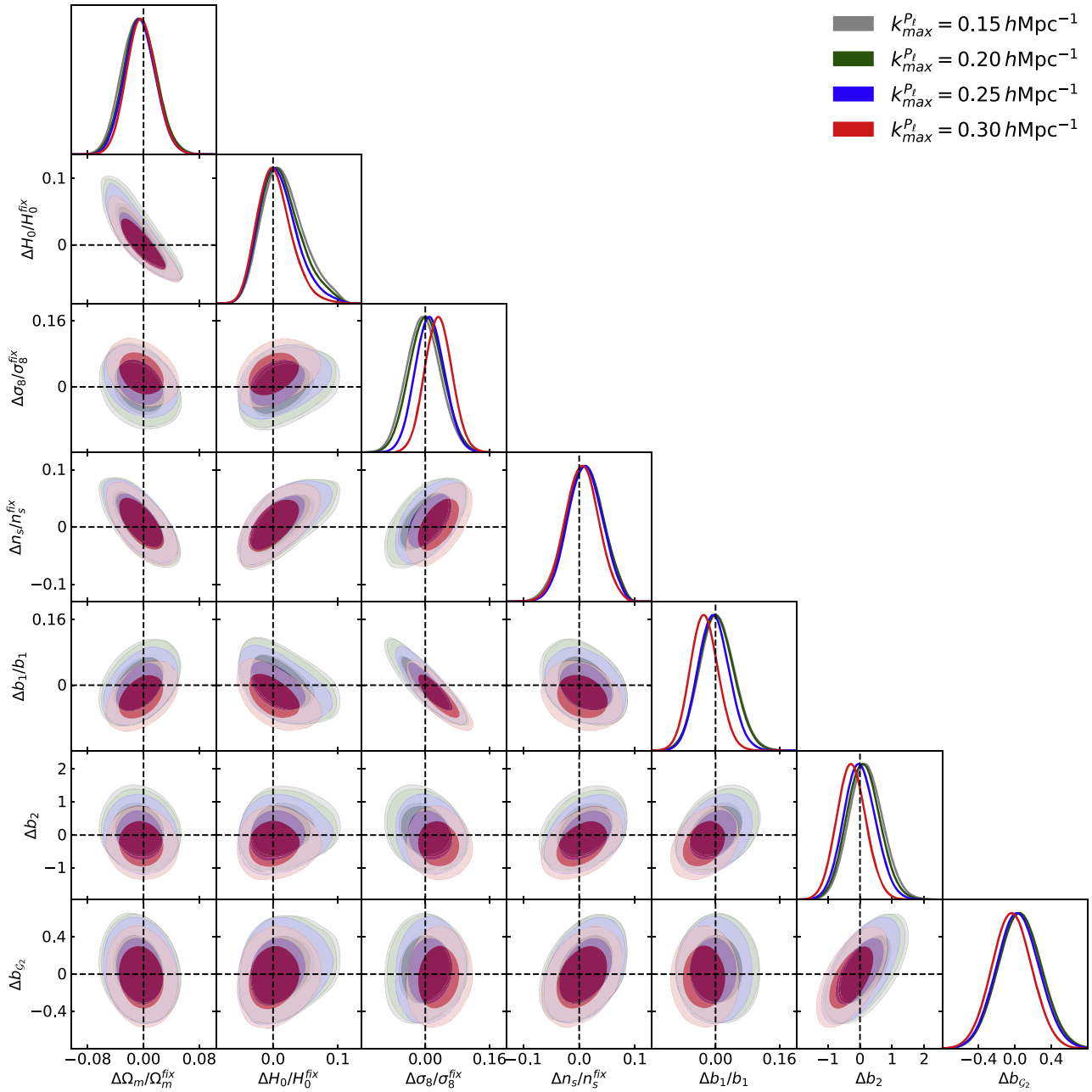

Figure 8. ΛCDM redshift-space mock analysis: parameter constraints from EFT P+B for four choices of the power spectrum scale cut $k^{P_\ell}_{\rm max}$. All parameters are normalized to fixed values taken from the best-fit at $k^{P_\ell}_{\rm max} = 0.2\ h\ {\rm Mpc}^{-1}$ (our fiducial scale cut). Physical baryon density $\omega_{\rm b}$ and some nuisance parameters are not displayed.

5.3. EFT FBI: Analysis Method[47]

In this section, we review the EFT FBI. The EFT FBI directly models and optimally extracts cosmological information from the three-dimensional galaxy density field within the EFT framework. This requires an explicit sampling of all Fourier modes of the initial conditions—in our case, corresponding to $90^3 \simeq 730{,}000$ parameters—in addition to bias, stochastic, and cosmological parameters. We refer readers to F. Schmidt et al. (2019) for the first EFT FBI groundwork and N.-M. Nguyen et al. (2024) for the most recent advance. We further note parallel efforts on cosmological field-level inference from galaxy clustering,[48] within (T. Baldauf et al. 2016b; M. Schmittfull et al. 2019; G. Cabass 2021; M. Schmittfull et al. 2021; G. Cabass et al. 2024) and outside of the EFT framework (U. Seljak et al. 2017; D. K. Ramanah et al. 2019; A. Andrews et al. 2023; A. E. Bayer et al. 2023b; Z. Wang et al. 2023), as well as the most recent advances within the community of machine learning for cosmology (e.g., C. Cuesta-Lazaro & S. Mishra-Sharma 2024; L. Doeser et al. 2024; M. Ho et al. 2024; D. Saadeh et al. 2025).

[47] Authors: N.-M. Nguyen, F. Schmidt.

[48] We define the term "field level" in relation to the (uncompressed) three-dimensional density field of biased tracers or matter, thereby restricting the scope of the discussion and literature.

Our EFT FBI analysis uses `LEFTfield`, a Lagrangian, EFT-based forward modeling of cosmological density fields. The code was first introduced in F. Schmidt (2021a), with extensions and validations later shown in A. Kostić et al. (2023), B. Tucci & F. Schmidt (2024), J. Stadler et al. (2023), I. Babić et al. (2024), and N.-M. Nguyen et al. (2024).

In this challenge, we analyze the real-space mocks. For the clustering of galaxies in their comoving rest frame, i.e., real space, both the matter δ and galaxy density field δ_g contain advection contributions that involve the Lagrangian displacement field. Since matter and galaxies comove on large scales following Einstein's equivalence principle, the displacement is the same for galaxies and matter. This direct consequence of the equivalence principle therefore ensures that the advection contribution to δ_g can be uniquely predicted (V. Desjacques et al. 2018), hence the bias–σ_8 degeneracy breaking.

5.3.1. Data and Estimators

Our data vector is the entire three-dimensional galaxy density field, namely $\hat{\boldsymbol{d}} \equiv \delta_g$, at $z = 1.0$. Similarly to N.-M. Nguyen et al. (2024), we construct a filtered galaxy density field $\delta_{g,\Lambda}$ from the galaxies in each catalog by a nonuniform-to-uniform fast Fourier transform. Here Λ is the cutoff scale of the sharp-k filter, such that all Fourier modes above Λ are set to zero.

5.3.2. Model

The `LEFTfield` forward model we employ in this challenge involves a *gravity model* evolving the initial conditions $\hat{s}$ to the Eulerian matter density field and a *galaxy model* connecting the Eulerian matter density field to the model prediction for the data vector $\delta_{g,\Lambda}$.

We note that our pre-unmasking analyses in the challenge had concluded before the analysis in N.-M. Nguyen et al. (2024) started—where the latter analyzed different data sets than this challenge, consisting of mass-selected main halos. Hence, while both works use the same *gravity model*, they differ in the *galaxy model*. Specifically, our baseline analyses here use a more restricted model compared to N.-M. Nguyen et al. (2024). We will highlight the differences below and revisit some of the main differences in our post-unmasking study.

Gravity Model. We predict gravitational evolution using the nth-order Lagrangian perturbation theory (LPT) framework (T. Matsubara 2015), as described in Section 2 and Appendices A–B of F. Schmidt (2021b). The convergence of this n-LPT scheme was demonstrated by F. Schmidt (2021b) in their Section 6, up to $n = 6$. In this challenge, we adopt $n = 2$, i.e., the 2LPT model (T. Buchert 1992; F. R. Bouchet et al. 1995).

Galaxy Model. We describe the matter–galaxy connection with the EFT bias expansion:

$$\delta_{g,\Lambda} = \sum_O b_O O[\delta_\Lambda], \tag{13}$$

where O are operators constructed out of the filtered Eulerian matter density field δ_Λ, together with their associated galaxy bias coefficients b_O. Similar to N.-M. Nguyen et al. (2024), we expand Equation (13) in the Eulerian basis. However, unlike N.-M. Nguyen et al. (2024), here we do so only up to second-order operators:

$$O \in [\delta,\ \delta^2 - \langle\delta^2\rangle,\ K^2 - \langle K^2\rangle,\ \nabla^2\delta], \tag{14}$$

where

$$K^2 \equiv (K^2_{ij}) = \left(\left[\frac{\partial_i\partial_j}{\nabla^2} - \frac{1}{3}\delta_{ij}\right]\delta\right)^2 \tag{15}$$

is the gravitational tidal field squared and $\nabla^2\delta$ is the leading higher-derivative operator. Note that these coefficients are equivalent to the subset of coefficients b_1, b_2, $b_{\mathcal{G}_2}$, c_0 in the EFT P+B model (Section 5.2). Compared to Equation (14), N.-M. Nguyen et al. (2024) add the full set of four third-order bias operators, while the EFT P+B team adds a single third-order bias term (Γ_3) that is relevant for the one-loop power spectrum. In our EFT-convergence test (see Figure 10) and post-unmasking study (see Figure 23), we additionally consider the full set of third-order bias terms and quantify the impact on our σ_8 constraints (Figure 4).

5.3.3. Inference

We sample and estimate the marginalized posterior of the amplitude rescaling parameter $\alpha = \sigma_8/\sigma_8^{\rm fiducial}$ given by

$$\begin{aligned}\mathcal{P}(\alpha|\widehat{\boldsymbol{d}}_\Lambda) &= \int\mathcal{D}\hat{s}\int d\sigma_\epsilon\int d\{b_O\}\,\mathcal{P}(\alpha,\{b_O\},\sigma_\epsilon,\hat{s}|\widehat{\boldsymbol{d}}_\Lambda)\\ &\propto \int\mathcal{D}\hat{s}\int d\sigma_\epsilon\int d\{b_O\}\,\mathcal{L}(\widehat{\boldsymbol{d}}_\Lambda|\alpha,\{b_O\},\sigma_\epsilon,\hat{s})\\ &\quad\times\mathcal{P}(\alpha)\mathcal{P}(\{b_O\})\mathcal{P}(\sigma_\epsilon)\mathcal{P}(\hat{s}),\end{aligned} \tag{16}$$

where the parameters to be explicitly sampled and numerically marginalized over are $\boldsymbol{\Phi} \equiv \{b_O, \sigma_\epsilon, \hat{s}\}$. Here σ_ϵ denotes the galaxy stochasticity, further described in the *Covariance* section.

The normalized initial conditions $\hat{s}$ (sometimes referred to as "phases"), which are drawn from a unit Gaussian prior, are related to the linear density field via

$$\delta^{(1)}(\boldsymbol{k}, z) = \alpha\left[\frac{N^3_{\rm grid}}{L^3}P_{\rm L}(k, z)\right]^{1/2}\hat{s}(\boldsymbol{k}). \tag{17}$$

In Equation (17), $P_{\rm L}(k, z)$ is the linear matter power spectrum in the fiducial cosmology, while L and $N_{\rm grid}$ are the length and grid size of the $\hat{s}(\boldsymbol{k})$ grid, respectively. Note that the finite volume L^3 and finite cutoff Λ imply that a finite grid is sufficient to explicitly represent every relevant Fourier mode of the initial conditions.

Likelihood. Following F. Schmidt et al. (2019, 2020), G. Cabass & F. Schmidt (2020a), F. Schmidt (2021a), A. Kostić et al. (2023), and N.-M. Nguyen et al. (2024), we adopt a Gaussian likelihood in Fourier space,

$$\begin{aligned}\ln\mathcal{L}(\widehat{\boldsymbol{d}}_\Lambda|\alpha,\hat{s},\{b_O\},\sigma_\epsilon) = -\frac{1}{2}\sum_{\boldsymbol{k}\neq 0}^{k_{\rm max}}\Bigg[&\ln 2\pi\sigma^2_\epsilon + \frac{1}{\sigma^2_\epsilon}|\widehat{\boldsymbol{d}}_\Lambda(\boldsymbol{k})\\ &- \delta_{g,\Lambda}[\alpha,\hat{s},\{b_O\}](\boldsymbol{k})|^2\Bigg]\\ &+ \text{const.},\end{aligned} \tag{18}$$

where we have used the fact that the galaxy stochasticity is homogeneous and isotropic in the galaxy comoving rest frame, i.e., in real space, and hence has a diagonal covariance matrix in Fourier space $\mathsf{C} = \sigma^2_\epsilon$. Note that Equation (18) introduces

the analysis cutoff scale $k_{\rm max}$ where $k_{\rm max} \leqslant \Lambda$. Unlike N.-M. Nguyen et al. (2024), here we choose a cubic sharp-k filter to implement the analysis cutoff $k_{\rm max}$ in Equation (18) and further set $k_{\rm max} = \Lambda/1.4$. A value of $k_{\rm max} < \Lambda$ is chosen to reduce the magnitude of higher-derivative terms that are controlled by the length scale Λ^{-1}. We have performed a detailed study of the optimal choice of $k_{\rm max}/\Lambda$ neither here nor in N.-M. Nguyen et al. (2024).

Equation (18) does not capture a coupling of galaxy stochasticity to large-scale perturbations ("density-dependent noise"), or non-Gaussianity of the noise. While the former can be incorporated via a real-space formulation of the likelihood (G. Cabass & F. Schmidt 2020b), the latter is technically more difficult to incorporate at the field level. Apart from the bias terms (see *Galaxy Model* above), this constitutes the second main model difference to the EFT P+B analysis.

Covariance. Even when limiting to a diagonal covariance in Fourier space, galaxy stochasticity is not necessarily white noise on all scales. In fact,

$$\sigma_\epsilon = \sigma_\epsilon(k) = \sigma_{\epsilon,0}[1 + \sigma_{\epsilon,k^2}k^2 + \sigma_{\epsilon,k^4}k^4 + \cdots], \tag{19}$$

in full generality (V. Desjacques et al. 2018). Though physically motivated, σ_{ϵ,k^2} and σ_{ϵ,k^4} add significant parameter degeneracies and correlations. Therefore, unlike N.-M. Nguyen et al. (2024), herein we consider only the scale-independent leading-order contribution[49] in our baseline analyses to minimize sampling cost.[50] Equation (19) then reduces to $\sigma_\epsilon(k) = \sigma_{\epsilon,0}$. This still leaves open two possibilities: (1) fixing $\sigma_{\epsilon,0} = \sigma_{\epsilon_{\rm Poisson}}$, where the latter follows the expectation for Poisson shot noise; and (2) inferring $\sigma_{\epsilon,0}$ from the data itself. We choose option 1 in our baseline analyses and consider the subleading correction $\propto k^2$, as employed in N.-M. Nguyen et al. (2024), in a post-unmasking analysis described below.

Priors on Cosmology and Bias. We adopt wide uniform priors for the amplitude rescaling parameter α and the bias parameters b_O. To summarize,

$$\mathcal{P}(\alpha) = \mathcal{U}(0.5,\, 1.5), \tag{20}$$

$$\begin{aligned}\mathcal{P}(b_\delta) = \mathcal{U}(0.0,\, 4.0),\ \ \mathcal{P}(b_{\delta^2}) = \mathcal{P}(b_{K^2}) = \mathcal{U}(-4.0,\, 4.0),\\ \mathcal{P}(b_{\nabla^2\delta}) = \mathcal{U}(-20.0,\, 20.0),\ \ \mathcal{P}(\sigma_\epsilon) = \delta_D(\sigma_{\epsilon_{\rm Poisson}}).\end{aligned} \tag{21}$$

Priors on Initial Conditions. As mentioned above, the cosmological prior on the normalized initial conditions $\hat{s}$, in the absence of primordial non-Gaussianity, is a unit Gaussian prior (free-IC). However, we will also consider a special analysis, on `box 1`, where we were given the ground-truth initial conditions used to initialize one of the N-body simulations underlying the challenge data set (fixed-IC). To summarize, we assume the following priors on $\hat{s}$:

$$\mathcal{P}(\hat{s}) = \begin{cases} \delta_{\rm D}(\hat{s} - \hat{s}_{\rm true}) & ({\rm fixed{-}IC},\ \texttt{box 1}), \\ \mathcal{N}(\hat{s};\, 0,\, {\rm I}) & ({\rm free{-}IC}), \end{cases} \tag{22}$$

where, in a fixed-IC analysis, $\delta_{\rm D}$ indicates the Dirac delta distribution fixing the initial conditions to the true initial conditions $\hat{s}_{\rm true}$ of the simulation data while, in a free-IC analysis, $\mathcal{N}$ indicates the multivariate normal distribution on normalized initial conditions with zero mean and identity covariance.

Sampling. We employ Hamiltonian Monte Carlo sampling to explore the initial conditions $\hat{s}$ while adopting slice sampling to sample other parameters. We refer readers to Section 3.2 of A. Kostić et al. (2023) and references therein for the motivation behind this choice.

MCMC Convergence and Autocorrelation. Unconverged and correlated MCMC chains lead to biases and uncertainties in the estimation of moments of the parameter posteriors. Below, we focus on MCMC convergence diagnostics and autocorrelation estimates for stage 2 analyses (see *Unmasking Criteria* below), as their results can be directly compared to those obtained by other teams, namely EFT P+B and `BACCO`P.

1. *MCMC convergence*. We run two MCMC chains with distinct initial parameter values for each analysis. We first identify the warmup and equilibrium phases in each chain through the parameter drifts within individual chains and the G-R statistics estimated from each and both chains.[51] After removing the warmup phases, we find posteriors estimated from the two chains to be completely consistent.
2. *MCMC autocorrelation*. The univariate autocorrelation function (ACF) provides an estimate for the number of effective, i.e., independent, samples $n^{\rm effective}$ in the MCMC chain for the parameter of interest. In the upper right corner of Figure 9, we show the ACFs of the cosmological and bias parameters as a function of the lag between two samples in the MCMC chain.[52] A null ACF indicates that two MCMC samples are independent draws from the posteriors.
3. *MCMC effective sample size (ESS)*. We target a G-R value of $\hat{R} \sim 0.01$ and an average number of $\sim$255 independent samples of α.[53] The MCMC sampling error on the means of α and σ_8 can be estimated from $\sigma_\alpha/n_\alpha^{\rm effective}$, where $n_\alpha^{\rm effective}$ is the ESS of α. The ESSs of 10 stage 2 analyses range from 193 to 311, with a mean and median of 254.4 and 259.5, respectively. Their G-R statistics range from 0.003 (best) to 0.035 (worst), with a mean of 0.0115 and median of 0.0085.

Posterior Consistency. A unique feature of the EFT FBI (stage 2) analyses is that we individually analyze all 10 data realizations in the real-space suite of mocks. Figure 9 shows the posteriors obtained in all 10 analyses, with their labels hidden to avoid revealing the cosmic variance in each realization. The main takeaway here is that all 10 posteriors are consistent with each other within the 68% confidence limit.

[49] On sufficiently large scales, these k-dependent corrections should be minimal (M. Schmittfull et al. 2019, 2021; F. Elsner et al. 2020; F. Schmidt et al. 2020; F. Schmidt 2021a).

[50] See our post-unmasking study and N.-M. Nguyen et al. (2024) for a more complete treatment of stochasticity.

[51] We observe neither divergence nor slow mixing in MCMC chains across all pre-unmasking and post-unmasking analyses.

[52] We estimate the ACFs using the FFT method, as implemented in `GetDist`. See, e.g., W. H. Press et al. (1986).

[53] We use `GetDist` (A. Lewis 2019) for these diagnostics and estimates. The package estimates the G-R statistics following the classical estimator in S. P. Brooks & A. Gelman (1998) and the effective sample number following their Equation (22).

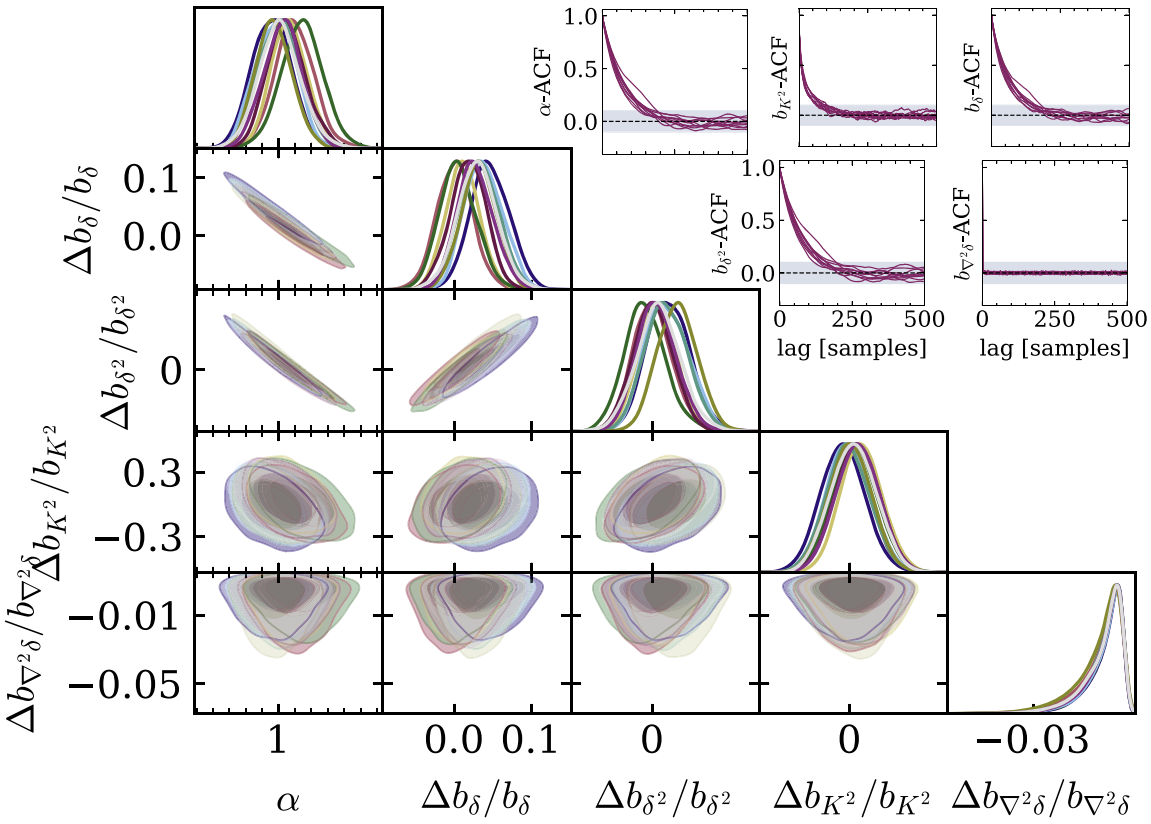

Figure 9. Consistency between posteriors obtained in 10 EFT FBI free-IC analyses on 10 data realizations during stage 2. To avoid unmasking the values of bias parameters, we subtract and divide their values in each chain by the corresponding posterior means obtained from the analysis of `box 1`.

5.3.4. Analysis Choices

Scale Cuts. We determine the fiducial scale cut $k_{\rm max}$ in our baseline analyses by choosing the maximum scale that maintains convergence of the amplitude rescaling parameter α inferred with different choices for the *galaxy model*, e.g., second- versus third-order galaxy bias expansion (Figure 10), and Lagrangian versus Eulerian basis for the bias expansion. Specifically, we define the relative parameter shift

$$\Delta\alpha \Big/ \sigma_\alpha \equiv \frac{\langle\alpha\rangle_A - \langle\alpha\rangle_B}{(\sigma^2_{\alpha_A} + \sigma^2_{\alpha_B})^{1/2}}, \tag{23}$$

where the indices A, B label different analyses with their associated posterior means $\langle\alpha\rangle$ and 68% uncertainties σ_α. We require Equation (23) to be less than 2.0 for final submissions, i.e., the `LEFTfield` result shown in Figure 4. The scale cut we end up with for our stage 1 and stage 2 submissions is $k_{\rm max} = 0.1\,h\,{\rm Mpc}^{-1}$. This then corresponds to the maximum scale up to which the second-order bias expansion is expected to be reliable. We anticipate that a third-order bias expansion will have higher reach in wavenumbers. For more details, see N.-M. Nguyen et al. (2024).

Unmasking Criteria. We unmask the EFT FBI analysis in two stages:

1. In stage 1, we performed the inference on `box 1`, with the initial conditions $\hat{s}$ being fixed to $\hat{s}_{\rm true}$ (fixed-IC) provided by the challenge organizers (only for `box 1`). During stage 1, we performed several EFT- and numerical-convergence tests because analyses are numerically cheap (at fixed initial conditions). After stage 1 submission and before stage 2 analysis started, the organizers provided us a confirmation that the stage 1 posterior mean of σ_8 is within 2σ of the true σ_8 (without specifying which side of the true σ_8 it actually is). For transparency, below in Figure 10 we also highlight the 2σ constraint of the stage 1 submission.
2. In stage 2, *without* any change in our analysis choices, we extended the analysis to all 10 boxes, while also allowing $\hat{s}$ to be explicitly sampled and marginalized over (free-IC). We emphasize that stage 2 analyses were performed *without* the true initial conditions $\hat{s}_{\rm true}$ of the simulation boxes.

In summary, the EFT FBI analyses and results are conducted and submitted in two stages, where σ_8 is fully masked, but $\hat{s}$ is provided for `box 1` in stage 1. The reason behind this choice is that we initially agreed with the challenge organizers to conduct a fixed-IC analysis only. After the submission of our fixed-IC analysis (stage 1), encouraged by the organizers, we decided to pursue a full free-IC analysis (stage 2), in coordination with the `BACCO` P and EFT P+B teams.

Caveats. In early tests or parallel studies (e.g., N.-M. Nguyen et al. 2021, 2024), we have identified an issue of "σ_ϵ collapse" when trying to infer the galaxy stochasticity, parameterized by $\sigma_{\epsilon,0}$, from the data themselves with a wide uniform prior on $\sigma_{\epsilon,0}$: the inferred $\sigma_{\epsilon,0}$ (hence σ_ϵ) tends to an unphysically low value. This is a known issue of inference and sampling using a Gaussian likelihood with unknown covariance, especially on noisy data, i.e., in this case, a galaxy sample with low number density. Ongoing work is devoted to an extended noise model and investigating physically motivated parameterizations and priors for σ_ϵ.

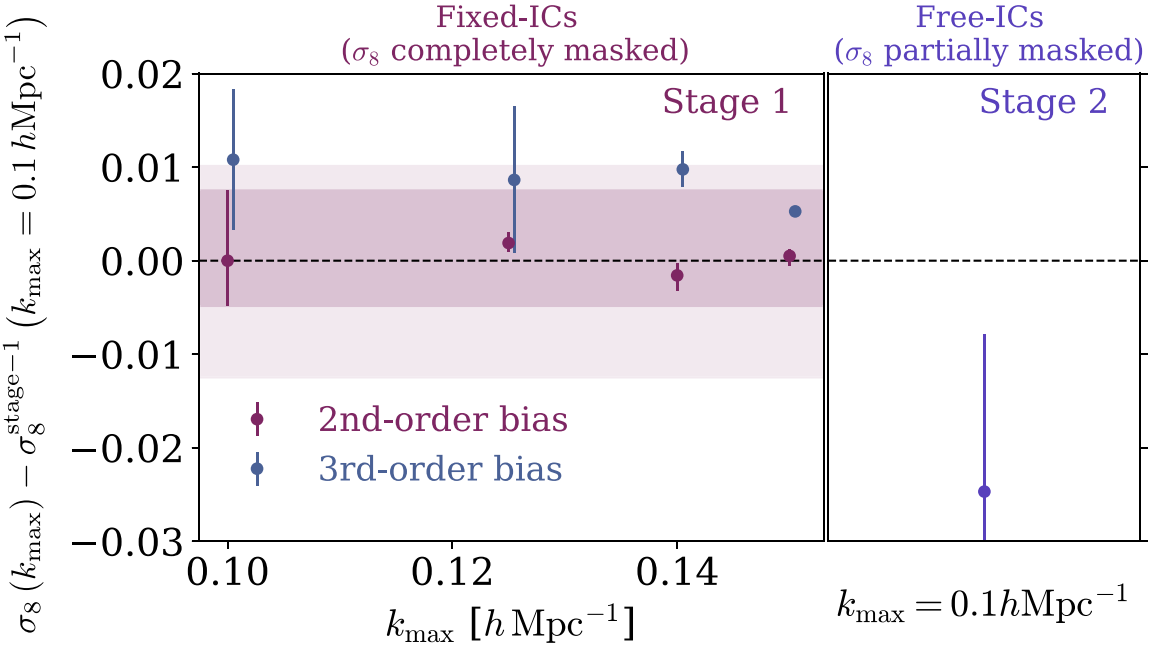

Figure 10. Convergence of $\sigma_8(z=0)$ inferred in our stage 1 and stage 2 analyses, using the same data realization (`box 1`). In both panels, the error bars indicate 68% lower and upper limits on α. All σ_8 values are quoted relative to the stage 1 analysis using the fiducial second-order bias model at $k_{\rm max} = 0.1$. We emphasize that the horizontal dashed line does *not* indicate the ground truth σ_8. The horizontal bands indicate the 68% and 95% confidence intervals of the σ_8 posterior in that scenario. Left (Stage 1): convergence of fixed-IC constraints as a function of the analysis (EFT) cutoff scale $k_{\rm max}$ (Λ). All σ_8 posterior means are within 1σ of the corresponding value at $k_{\rm max} = 0.1\ h\ {\rm Mpc}^{-1}$ for the respective bias model (same color). Magenta and blue points in the left panel indicate the second- and third-order bias model, respectively. The blue points are horizontally displaced to make error bars fully visible. Right (Stage 2): free-IC constraint from `box 1`, again relative to the stage 1 result.

Post-unmasking Studies. Our baseline analyses opted for efficiency in model evaluation and sampling, specifically the ability to scale up our EFT FBI algorithm and individually analyze all 10 realizations of the real-space ΛCDM data suite during our stage 2 analysis. Therefore, we adopted the second-order bias expansion and assumed galaxy stochasticity to be white noise with Poisson amplitude (Equation (21)). Our primary motivation for a post-unmasking study is to facilitate a comparison between our σ_8 constraint and that reported by the EFT P+B team. To this end, we have reanalyzed `box 1` while extending the *galaxy model* in the *model* and the *covariance* in the *inference* as follows:

1. *Galaxy model*. Going from second-order to third-order bias (see Equation (A4) of N.-M. Nguyen et al. 2024).
2. *Covariance*. Allowing for the galaxy stochasticity to be scale dependent with unknown amplitudes, i.e., $\sigma_\epsilon(k) = \sigma_{\epsilon,0}[1 + \sigma_{\epsilon,k^2}k^2]$, where $\sigma_{\epsilon,0}$, σ_{ϵ,k^2} are jointly inferred from the data.

With these new choices, our *model* and *inference* correspond to the same adopted in N.-M. Nguyen et al. (2024). For the cubic bias parameters, we adopt the same uniform prior adopted for the quadractic bias parameters, i.e.,

$$\mathcal{P}(b_{\delta^3}) = \mathcal{P}(b_{K^3}) = \mathcal{P}(b_{\delta K^2}) = \mathcal{P}(b_{O_{\rm td}}) = \mathcal{U}(-4.,4.). \quad (24)$$

For the additional stochastic parameters $\sigma_{\epsilon,0}$ and σ_{ϵ,k^2}, we assume the following flat priors:

$$\mathcal{P}(\sigma_{\epsilon,0}) = \mathcal{U}(0.9,\ 1.1)\sigma_{\epsilon_{\rm Poisson}},\ \ \mathcal{P}(\sigma_{\epsilon,k^2}) = \mathcal{U}(-10.0,\ 100.0). \quad (25)$$

The result of this post-unmasking analysis is shown in Figure 23 under the label "FBI extended." The systematic bias in σ_8 is now completely relieved, while the increased parameter vector leads to a larger error bar compared to the pre-unmasking analysis. Our constraints are further in line with the expectations from the EFT FBI constraints presented in N.-M. Nguyen et al. (2024). One technicality is that here we implement the analysis cutoff $k_{\rm max}$ with a cubic sharp-k filter, i.e., $\prod_{i=1}^{3}\theta_H(|k_i| - k_{\rm max})$, while N.-M. Nguyen et al. (2024) implemented the cutoff with a spherical sharp-k filter, i.e., $\theta_H(|\boldsymbol{k}| - k_{\rm max})$. Therefore, our analyses with $k_{\rm max} = 0.1\ h\ {\rm Mpc}^{-1}$ here involve roughly the same number of modes as (and hence should be compared to) the analyses with $k_{\rm max} = 0.12\ h\ {\rm Mpc}^{-1}$ in N.-M. Nguyen et al. (2024). For more discussion and post-unmasking comparisons, we refer the reader to Section 6.

5.4. SBI: P+B Method[54]

In this section, we outline the analysis of the redshift-space mocks using simulation-based inference (SBI).

SBI uses computationally simulated forward models of the observed data that are evaluated using parameter values spanning the prior (K. Cranmer et al. 2020). It leverages the forward models to learn the likelihood distribution of any measurable data statistics and can, in principle, be applied to analyze many of the summary statistics presented in this paper. Since the likelihood is learned from the forward models, it relaxes the assumptions on the likelihood (e.g., the emulator approach typically assumes a Gaussian likelihood). Furthermore, the forward models can include observational effects to more robustly treat systematics (e.g., S. Yuan et al. 2023a). SBI has been applied to large-scale structure analysis (J. Alsing et al. 2019; N. Jeffrey et al. 2021; C. Hahn et al. 2023a, 2023b;

[54] Authors: C. Modi, C. Hahn.

Table 3
Summary of HOD Parameterizations Used by the SBI (HOD1), *k*NN, and Density-split (HOD2) Analysis Teams

Ingredient	HOD1 (C. Hahn et al. 2023b)	HOD2 (S. Yuan et al. 2022b)
1. Vanilla Z. Zheng et al. (2007) HOD Parameters Equations (2)–(3)		
	satellite occupation cutoff parameter: M_0	satellite occupation cutoff parameter: $\kappa M_{\rm cut}$
	…	$\bar{n}_{\rm sat}(M)$ modulated by $\bar{n}_{\rm cent}(M)$
	satellites assigned via NFW profile	satellites assigned to halo particles
	$f_{\rm ic} \equiv 1$	$f_{\rm ic}$ determined by observed galaxy density
2. Velocity Bias Parameters		
$\alpha_{\rm vel,c}$	…	central velocity bias defined relative to halo center
$\alpha_{\rm vel,s}$	…	satellite velocity bias defined relative to particle
$\eta_{\rm cen}$	central velocity bias defined as additional dispersion	…
$\eta_{\rm sat}$	satellite velocity bias defined as additional dispersion	…
3. Galaxy Assembly Bias Parameters		
$A_{\rm c}$	central occupation conditioned on concentration	…
$A_{\rm s}$	satellite occupation conditioned on concentration	…
$B_{\rm cent}$	…	conditioned on environment or concentration
$B_{\rm sat}$	…	conditioned on environment or concentration
4. Baryonic Effects		
$\eta_{\rm conc}$	concentration ratio of satellite and halo profile	…
s	…	modulation of radial satellite galaxy profile

B. Tucci & F. Schmidt 2024), and our analysis here will most closely follow C. Hahn et al. (2023a).

5.4.1. Data and Estimators

We focus on analyzing the power spectrum multipoles $P_\ell(k)$ for ($\ell = 0, 2, 4$) and bispectrum monopole $B_0(k_1, k_2, k_3)$ using the mean measurement from the 10 ΛCDM redshift-space boxes. Power spectrum multipoles are measured with fast Fourier transforms using Nbodykit (N. Hand et al. 2018) on a 512^3 mesh, while bispectrum is measured using the `pySpectrum` code (C. Hahn 2020) on a 760^3 mesh. In addition, we supplement our data vector with the galaxy number density, $\bar{n}$.[55] We use both the statistics between $k_{\rm min} = 0.009\,h\,{\rm Mpc}^{-1}$ and $k_{\rm max} = 0.5\,h\,{\rm Mpc}^{-1}$. This results in the power spectrum data vector of size 79×3 (for three multipoles) and bispectrum data vector of size 1898.

5.4.2. Model

For SBI, we need to accurately model galaxy clustering statistics over a range of cosmology and galaxy formation parameters. To this end, we use the `Quijote` *N*-body simulation suite (F. Villaescusa-Navarro et al. 2020), specifically the high-resolution ΛCDM Latin hypercube (LH), which consists of 2000 simulations varying over five cosmological parameters: $\Omega_{\rm m}$, $\Omega_{\rm b}$, h, $n_{\rm s}$, and σ_8. Each simulation evolves 1024^3 dark matter particles with the TreePM Gadget-III code in a volume of 1 Gpc/h. For each simulation, there are two sets of dark matter halos identified with ROCKSTAR (P. S. Behroozi et al. 2013) and the friends-of-friends (FoF) halo finder, respectively. Recent work has shown that SBI models trained on one set can lead to biased results on another (C. Modi et al. 2025); hence, we will use both halo catalogs to evaluate robustness of our analysis. However, the final results of this challenge will be derived using halos from ROCKSTAR, which more accurately identifies the position and velocity of halos.

Next, we populate these dark matter halos with galaxies. We use an extended nine-parameter HOD model that includes assembly, concentration, and velocity biases, summarized in Table 3. Five of these parameters are the same as the standard Zheng07 HOD (Z. Zheng et al. 2007) model: $\log M_{\rm min}$, $\sigma_{\log M}$, $\log M_0$, $\log M_1$, and α. These are supplemented with "assembly" bias parameters ($A_{\rm c}$, $A_{\rm s}$), which modify the number of centrals and satellites based on the halo concentration;[56] the "concentration" bias parameter ($\eta_{\rm conc}$), which allows the concentration of satellite galaxies to deviate from the NFW profile; and "velocity" dispersion parameters ($\eta_{\rm c}$, $\eta_{\rm s}$), which rescale the central and satellite velocities over the halo velocity. For further details on implementation of this decorated HOD, we refer the reader to A. P. Hearin et al. (2016) and C. Hahn et al. (2023b).

5.4.3. Inference

Methodology. SBI requires a training data set of ($\boldsymbol{\theta}$, $\boldsymbol{x}$) pairs, where $\boldsymbol{\theta}$ are the model parameters of interest (here cosmology and HOD parameters) and $\boldsymbol{x}$ are the corresponding summary statistic (here power spectrum multipoles, bispectrum, and number density).

Parameter $\boldsymbol{x}$ is generated using the simulations and estimators described above. The data set of $\{(\boldsymbol{\theta}, \boldsymbol{x})\}$ therefore corresponds to samples drawn from the joint distribution $p(\boldsymbol{\theta}, \boldsymbol{x})$. More importantly, $\{(\boldsymbol{\theta}, \boldsymbol{x})\}$ can be used to infer the posterior $p(\boldsymbol{\theta}|\boldsymbol{x})$.

[55] We find that explicitly including number density does not add any significant constraining power but helps make the analysis more stable when the number density varies widely over different cosmologies of the LH for the same HOD parameters.

[56] FoF halos do not measure concentration accurately, and hence we use an analytic relation to estimate this value from halo mass (A. A. Dutton & A. V. Macciò 2014). In this sense it is not strictly assembly bias, but a different parameterization with respect to halo mass.

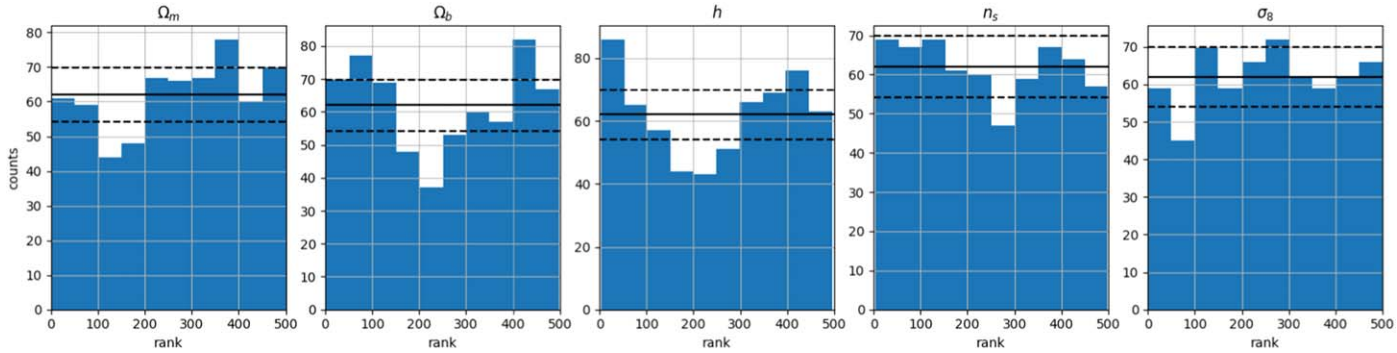

Figure 11. Rank histograms of the five cosmology parameters for the trained ensemble evaluated using the held-out test data set. The solid horizontal black line is the mean number of counts expected for a uniform distribution, and the dashed lines indicate the expected Poisson scatter. Broadly, all the histograms are consistent with uniform distribution.

While different methods can be used to infer $p(\boldsymbol{\theta}|\boldsymbol{x})$ using $\{(\boldsymbol{\theta}, \boldsymbol{x})\}$, we use neural density estimators, which have been used to accurately estimate even high-dimensional distributions with limited training data. More specifically, we train a conditional neural density estimator, q_ϕ, with parameters ϕ to approximate $p(\boldsymbol{\theta}|\boldsymbol{x}) \approx q_\phi(\boldsymbol{\theta}|\boldsymbol{x})$. We do this by minimizing the Kullback–Leibler divergence between $p(\boldsymbol{\theta}|\boldsymbol{x})$ and $q_\phi(\boldsymbol{\theta}|\boldsymbol{x})$, estimated by the log-probability of q_ϕ evaluated over the training data set.

Once $q_\phi(\boldsymbol{\theta}|\boldsymbol{x})$ is successfully trained, we can infer the posterior for observations, $\boldsymbol{x}'$, by simply querying the trained q_{ϕ^*} to generate samples from the posterior: $\boldsymbol{\theta}' \sim q_{\phi^*}(\boldsymbol{\theta}|\boldsymbol{x}')$.

Training Data and Priors. To generate the training data set for SBI, we use the QUIJOTE LH described in the previous section. For each simulation in LH, we sample 10 different HOD parameter values over a prior range and generate 10 galaxy catalogs. Thus, in total, we have 20,000 galaxy catalogs. We used catalogs for 1500 of these simulations (i.e., 15,000 galaxy catalogs) for training, used 200 for validation, and held out 300 for testing. The prior range of cosmological parameters is set by the bounds of QUIJOTE LH,

$$\Omega_{\rm m} \sim \mathcal{U}[0.1, 0.5],\ \sigma_8 \sim \mathcal{U}[0.6, 1.0],$$
$$\Omega_{\rm b} \sim \mathcal{U}[0.03, 0.07],\ n_{\rm s} \sim \mathcal{U}[0.8, 1.2],\ h \sim \mathcal{U}[0.5, 0.9]. \tag{26}$$

For HOD parameters, we shift the central values of prior on $\log M_{\rm min}$, $\log M_0$, and $\log M_1$ for different cosmologies so as to match the given number density on average. This allows us to more efficiently sample the parameter space by focusing only on the regions that will generate simulations loosely consistent with the data. These central values are estimated as follows: we assume a fiducial value of $\alpha = 0.7$ and satellite fraction of 0.2; then, for every cosmology, we set $\log M^\theta_{\rm min} = \log M^\theta_0 = M_h$, the halo mass above which the number of halos is the same as the number of centrals; and then we estimate $\log M^\theta_1$ to match the satellites. The resulting priors are as follows:

$$\log M_{\rm min} \sim \mathcal{U}[\log M^\theta_{\rm min} \pm 0.15],$$
$$\log M_0 \sim \mathcal{U}[\log M^\theta_0 \pm 0.2], \quad \log M_1 \sim \mathcal{U}[\log M^\theta_1 \pm 0.3].$$

For the remaining HOD parameters, we use the following priors for all cosmologies:

$$\alpha \sim \mathcal{U}[0.4, 1.0],\ \sigma_{\log M} \sim \mathcal{U}[0.3, 0.5],$$
$$A_{\rm c}, A_{\rm s} \sim \mathcal{N}(0, 0.2) \text{ over } [-1, 1],$$
$$\eta_{\rm conc} \sim \mathcal{U}[0.2, 2.0],\ \eta_{\rm c} \sim \mathcal{U}[0., 0.7],\ \eta_{\rm s} \sim \mathcal{U}[0.2, 2.0].$$

We estimate the power spectrum multipoles and bispectrum for each of the 20,000 simulated galaxy catalogs.

Posterior Inference. We use the `sbi`[57] package to train masked autoregressive flows as conditional neural density estimators and learn the posterior, $q_\phi(\boldsymbol{\theta}|\boldsymbol{x}) \sim p(\boldsymbol{\theta}|\boldsymbol{x})$. To minimize stochasticity in training, we use the `Weights-and-Biases`[58] package and train 200 networks for each data statistic by varying hyperparameters such as the width and number of layers, learning rate, and batch size. After training, we collect 10 neural density estimators with the best validation loss and use them as an ensemble, i.e., we construct a mixture distribution with uniform weighting to approximate the posterior.

5.4.4. Analysis Choices

Scale Cuts. The QUIJOTE simulations are only $(1\,{\rm Gpc}/h)^3$ in volume, while the data challenge simulations are $(2\,{\rm Gpc}/h)^3$ in volume. Thus, in addition to the typical small-scale cut, we also impose a large-scale cut on the power spectrum and bispectrum measurements. We use only the scales between $k_{\rm min} = 0.009h\,{\rm Mpc}^{-1}$ and $k_{\rm max} = 0.5h\,{\rm Mpc}^{-1}$. We point out that the smaller volume of our training simulations than the mock data results in our learned-likelihood, and hence the posterior, being overdispersed owing to larger cosmic variance. This is further exacerbated for $\Omega_{\rm m}$ since we miss the informative large-scale modes owing to larger $k_{\rm min}$ than the fundamental wavenumber of the challenge mocks, as discussed in Section 2.1. Thus, we expect our results to be more conservative than if the simulations had the same volume. We also note that in the future such limitations on the volume of the training simulations can be overcome using hybrid SBI techniques to combine EFT analysis on the large-scale analysis with SBI analysis only on the small scales (C. Modi & O. H. E. Philcox 2023).

Validation. We use our trained ensemble to predict the cosmological parameters over the held-out test data set to do coverage tests as described in S. Talts et al. (2018) and C. Hahn et al. (2023a). We verify that all the rank histograms are uniformly distributed within the rank scatter, which suggests (but does not guarantee) that the posterior marginal distributions are accurate. These histograms are shown for the five cosmology parameters in Figure 11. To test for model misspecification, we repeat the exercise with QUIJOTE-FoF LH and find consistent results.

[57] Available at https://github.com/sbi-dev/sbi.
[58] Available at https://wandb.ai/site.

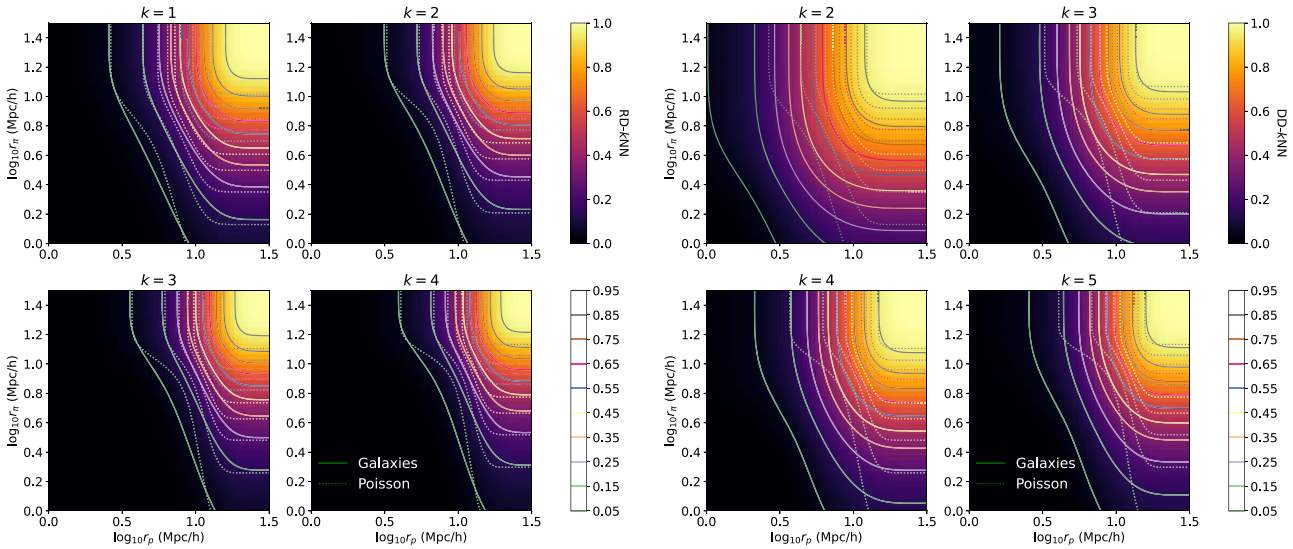

Figure 12. Visualizations of the RD-*k*NN statistics (left) and DD-*k*NN statistics (right), calculated on the redshift-space mock, averaged over 10 realizations. The dotted lines show the contours of an unclustered Poisson random sample. We only showcase the first four *k*-orders for brevity. Note that we have defined DD-*k*NN = 1; thus, $k = 2$ is the first meaningful order.

Caveats. The primary source of error with SBI is model misspecification (P. Cannon et al. 2022; C. Modi et al. 2025). Previous analyses (C. Hahn et al. 2023a; C. Modi et al. 2025) suggest that inference with power spectrum up to $k_{\rm max} = 0.5h\,{\rm Mpc}^{-1}$ should be robust to these modeling differences. However, more powerful statistics such as bispectrum and combined power spectrum and bispectrum analysis can exacerbate model misspecification, due to both increased sensitivity and smaller error bars. This parameter-masked challenge provides a unique opportunity to investigate this aspect.

5.5. k*NNs with* `AbacusSummit` *Emulator*[59]

In this section, we summarize the methodology of analyzing the 2D (*k*NN) statistics with an extended-HOD emulator based on the `AbacusSummit` simulation suite. The detailed methodology is described in a dedicated supporting paper (S. Yuan et al. 2024a).

5.5.1. Data and Estimators

For this analysis, we use two different 2D *k*NN statistics: (1) the random-data *k*NN statistics (RD-*k*NN), and (2) the data −data *k*NN statistics (DD-*k*NN). For a detailed description of these statistics, we refer the readers to S. Yuan et al. (2023b, 2024a). Conceptually, the *k*NN statistics captures the probability distribution of the separation between random query points (or in the case of DD-*k*NN, the query points are galaxies) and their *k*th-nearest galaxies. Thus, in principle, *k*NNs are parameterized as a function of k and the separation, which we further decompose in 2D to a transverse component r_p and an LOS component r_π. Figure 12 visualizes the RD and DD-*k*NN statistics up to the first four orders, where each panel shows how the *k*NN-cumulative distribution function (CDF) depends on the separation (r_p, r_π). While the heatmap and the solid contours showcase the *k*NN measurement on the provided redshift-space mocks, the dashed contours visualize the measurement on an unclustered Poisson random sample of the same size for comparison. The difference between the solid and dashed contours represents the informative features in the statistics.

The exact compressions we use to construct our data vectors for this analysis are as follows: We use $k = 1, 2, 3,\ldots,9$, and for each k we use eight logarithmic bins along the r_p direction between 0.63 and $63\,h^{-1}$ Mpc and five logarithmic bins along the r_π direction between 0.5 and $32\,h^{-1}$ Mpc. We also remove bins where the CDF is less than 0.05 or greater than 0.95 to increase the overall signal-to-noise ratio of our statistics. As a result, the final data vector is summarized with 114 bins in the RD case and 144 bins in the DD case. Our r_p and r_π binning choices are designed to expose our analysis to the nonlinear and quasi-nonlinear scales. We are not sensitive to large-scale features such as the BAO by design. Our k choices are fairly arbitrary and are mostly chosen to confine us to nonlinear scales and computational efficiency. We reserve the discussion of optimal k choices for a future paper.

5.5.2. Model

To forward-model the galaxy *k*NNs, we employ a neural-network-based emulator that learns the summary statistics as a function of cosmology and extended-HOD parameters. In this subsection, we introduce the relevant model details and emulation techniques.

Parameterization. For the galaxy–halo connection, we employ the `AbacusHOD` package for efficiency and beyond vanilla HOD extensions (S. Yuan et al. 2022b). The model notably assigns satellites to halo particles and includes velocity bias, two types of galaxy assembly bias, and baryonic modulations. The model parameters were summarized as model HOD2 in Table 3. Here we expand on the relevant model extensions.

Velocity bias is parameterized with two parameters $\alpha_{\rm vel,c}$ and $\alpha_{\rm vel,s}$. The central velocity bias parameter $\alpha_{\rm vel,c}$ modulates the peculiar velocity of the central galaxy relative to the halo center along the LOS. Here $\alpha_{\rm vel,c} = 0$ corresponds to no central velocity bias, i.e., centrals perfectly track the velocities of halo centers. We also define $\alpha_{\rm vel,c}$ as nonnegative.

[59] Author: S. Yuan.

The satellite velocity bias parameter $\alpha_{\rm vel,s}$ modulates how the satellite galaxy peculiar velocity deviates from that of the local dark matter particle. Here $\alpha_{\rm vel,s}=1$ indicates no satellite velocity bias, i.e., satellites perfectly track the velocity of their underlying particles. Detailed descriptions of our implementation can be found in S. Yuan et al. (2022b).

The assembly bias parameters can be defined against the halo concentration or the local overdensities over an $r_{\rm env}=5$ h^{-1} Mpc tophat filter. $B_{\rm cent}=0$, $B_{\rm sat}=0$ indicate no galaxy assembly bias. In this analysis, we consider assembly bias against both halo concentration and local environment when we compare different models.

To summarize, the extended-HOD model employed in the kNN analysis consists of 10 parameters: (1) five vanilla HOD parameters $M_{\rm cut}$, M_1, σ, α, κ; (2) an incompleteness parameter $f_{\rm ic}$; (3) velocity bias parameters $\alpha_{\rm vel,c}$ and $\alpha_{\rm vel,s}$; (4) galaxy assembly bias parameters $B_{\rm cent}$ or $B_{\rm sat}$; and (5) baryonic modulation parameter s.

Simulation Details. This analysis employs the `AbacusSummit` simulation suite (N. A. Maksimova et al. 2021), a set of large, high-accuracy cosmological N-body simulations using the `Abacus` N-body code (L. H. Garrison et al. 2019), designed to meet and exceed the Cosmological Simulation Requirements of the Dark Energy Spectroscopic Instrument (DESI) survey (M. Levi et al. 2013). `AbacusSummit` consists of over 150 simulations, containing approximately 60 trillion particles at 97 different cosmologies. The `AbacusSummit` suite also uses a new specialized spherical-overdensity-based halo finder known as COMPASO (B. Hadzhiyska et al. 2022). For this analysis, we use the "base" configuration boxes for forward modeling, each of which contains 6912^3 particles within a $(2\,h^{-1}\,{\rm Gp})^3$ volume, corresponding to a particle mass of $2.1\times10^9\,h^{-1}\,M_\odot$.[60]

To model cosmology dependencies, we use the 85 emulator boxes. The cosmology parameter basis used for our emulator includes eight parameters spanning the wCDM+Neff+running parameter space: the baryon density $\omega_{\rm b}=\Omega_{\rm b}h^2$, the cold dark matter density $\omega_{\rm cdm}=\Omega_{\rm cdm}h^2$, the amplitude of structure σ_8, the spectral tilt $n_{\rm s}$, running of the spectral tilt α_s, the density of massless relics $N_{\rm eff}$, and dark energy equation-of-state parameters w_0 and w_a ($w(a)=w_0+(1-a)w_a$). The different cosmologies are indexed by `cXXX`, where `XXX` ranges from 000 to 181. The details of each cosmology are described on the `AbacusSummit` website.[61]

Emulator Details. We build a neural-network-based emulator to interpolate between the finite set of cosmologies. Specifically, we follow the approach of S. Yuan et al. (2022a), where we take advantage of the high efficiency of the `AbacusHOD` code and run MCMC chains in the HOD parameter space against the target data vector at each cosmology. We stop the MCMC chains after 20,000 evaluations (limited by computational resources) in each box and select samples whose likelihood is greater than $\log L>-9000$ (to ensure a large training sample around the maximum likelihood region) as the training set for the subsequent emulator model. This approach constrains the emulator training to a compact region in the cosmology+HOD parameter space, improving the emulator precision. We also find consistent behavior when selecting our training set with different likelihood cutoffs. Similar approaches were adopted in S. Yuan et al. (2022a, 2024a). For the emulator model, we adopt a fully connected neural network of five layers and 500 nodes per layer with Randomized Leaky Rectified Linear Units activation. We train the network following a minibatch routine with the Adam optimizer and a mean squared loss function, where we use the diagonal terms of the covariance matrix as bin weights. The performance of the emulator is characterized on nine hold-out cosmologies (`c001-004` and `c171-175`). We refer the readers to S. Yuan et al. (2024a) for details.

5.5.3. Inference

Likelihood Function. For this analysis, we assume a Gaussian likelihood function; see S. Yuan et al. (2024a) for validation of the Gaussian likelihood assumptions. To sample the likelihood function, we employ the nested sampling package `dynesty` (J. Speagle & K. Barbary 2018; J. S. Speagle 2020).

Covariance. We compute jackknife covariances for 2D-kNNs using the 10 realizations provided. We divide each realization into 125 cubic chunks, each of size $(400\ h^{-1}\,{\rm Mpc})^3$. We visualize the resulting correlation matrix in Figures A1 and A2 of S. Yuan et al. (2024a). Both covariances are invertible. The RD-kNN covariance matrix has a condition number of 1×10^7, and the DD-kNN covariance matrix has a condition number of 4×10^7. To account for emulation errors, we compute a separate emulator covariance matrix from the hold-out cosmologies and add that onto the data jackknife covariance matrix.

Priors. This analysis employs cosmology priors based on the ΛCDM parameter range covered by the `AbacusSummit` simulations,

$$\begin{aligned}&\omega_{\rm cdm}\in[0.099,\,0.140],\quad \sigma_8\in[0.68,\,0.94],\\ &n_{\rm s}\in[0.90,\,1.02],\quad \omega_{\rm b}\in[0.021,\,0.024],\end{aligned}\tag{27}$$

with the prior distribution specified by the `AbacusSummit` parameter envelope, $N_{\rm ur}=2.0328$ and $\alpha_s=0$. For the HOD parameters, we adopt the following flat priors:

$$\begin{aligned}&\log_{10}M_{\rm cut}\sim\mathcal{U}[12.0,\,14.5],\ \log_{10}M_1\sim\mathcal{U}[13.0,\,15.0],\\ &\alpha\sim\mathcal{U}[0.5,\,1.5],\\ &\alpha_{\rm vel,c}\sim\mathcal{U}[0.0,\,1.0],\ \alpha_{\rm vel,s}\sim\mathcal{U}[0.2,\,1.8],\\ &\log_{10}\sigma\sim\mathcal{U}[-3.5,\,1.5],\\ &\kappa\sim\mathcal{U}[0.0,\,2.0],\ B_{\rm cent}\sim\mathcal{U}[-1.0,\,1.0],\\ &B_{\rm sat}\sim\mathcal{U}[-1.0,\,1.0].\end{aligned}\tag{28}$$

5.5.4. Analysis Choices

Scale Cut Validation. For both kNN configurations, the first cut we apply is removing bins where the CDF value measured on the target mock is either less than 0.05 or greater than 0.95. This removes the noisiest bins and Gaussianizes the likelihood function.

We apply an additional scale cut to the DD-kNN statistics in $r_p>5\ h^{-1}$ Mpc. We refer the readers to S. Yuan et al. (2024a) for justification of this cut. Essentially, the DD-kNN statistic is highly sensitive to small-scale modeling, and we find our models to not be flexible enough to accurately predict the DD-kNN statistic at $r_p<5\ h^{-1}$ Mpc given the precision of a $(2\,h^{-1}\,{\rm Gpc})^3$ volume. We do not apply any scale cut to the RD-kNN statistic, as it is less sensitive to small-scale modeling. Unfortunately, our scale cut at $r_p<5\ h^{-1}$ Mpc comes at significant cost to our constraining power on cosmology. This

[60] For more details, see https://abacussummit.readthedocs.io/en/latest/abacussummit.html.
[61] https://abacussummit.readthedocs.io/en/latest/cosmologies.html

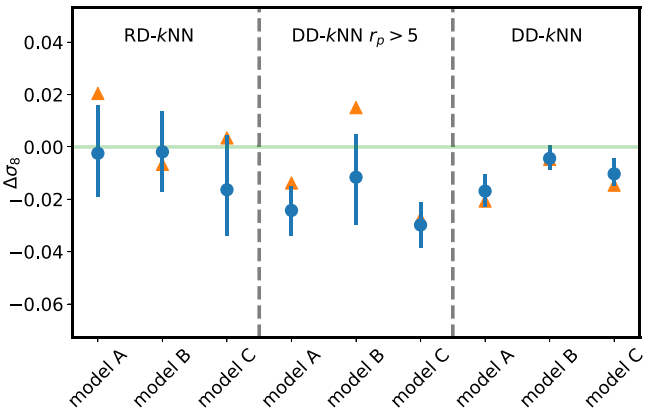

Figure 13. (Plot adapted from Figure 9 of S. Yuan et al. 2024a). The constraints of *k*NN statistics on σ_8 for different HOD models and different scale cuts. Models A, B, and C refer to the three candidate HOD models. The blue error bars show the 1σ marginalized constraints. The orange triangles show the maximum likelihood points. The green line represents the truth. The left, middle, and right blocks represent the RD-*k*NN constraints, DD-*k*NN constraints with scale cut $r_p > 5\ h^{-1}$ Mpc, and DD-*k*NN constraints without minimum scale cuts ($r_{\rm min} = 0.67\ h^{-1}$ Mpc), respectively. Our pre-unmasking consistency tests favored model B, and we see post-unmasking that model B indeed results in the least biased constraints. Comparing the middle block to the right block shows that the inclusion of small scales drastically tightens the constraints on σ_8. Unfortunately, our models were not flexible enough to self-consistently model $r_p < 5\ h^{-1}$ Mpc. This comparison motivates for more robust modeling of small scales.

highlights an important avenue of future work. We discuss this point further in Section 6.

Consistency Checks. Our supporting paper (S. Yuan et al. 2024a) details how we test and compare different HOD models invoking different extensions. We leverage both goodness-of-fit metrics and Bayesian evidence. We also conduct cross-validation tests where we fit one subset of the data vectors and predict a different set of data vectors.

Specifically, we set RD-*k*NN and DD-*k*NN as separate data vectors, as they summarize different information of the density field. RD-*k*NN captures the density information, and DD-*k*NN captures the clustering information. We also include the standard 2PCF as a third statistic that shares some information with DD-*k*NN but also captures clustering out to larger scales. We start with three different HOD models that include some combinations of baryonic effect treatment and two different types of galaxy assembly bias in addition to the vanilla HOD. We fit all three different HOD models to one of the three statistics and predict one of the other two statistics.

Our tests clearly favor one HOD model, which results in good fits and no tension between different subsets of the data vectors. We show the constraints of the three models on σ_8 in Figure 13, where we see that the favored model B indeed achieves the least biased constraints post-unmasking. We do not reveal the exact details of the favored HOD model to keep the challenge for future participants masked. Note that we submitted only the cosmology constraints for the favored model B for the official unmasking.

Unmasking Criteria. Our unmasked criteria consist of the series of tests that we have described so far. To summarize, we first test the Gaussianity of our summary statistics. As a result, we removed *k*NN bins where the CDF value is less than 0.05 or greater than 0.95. Then, we construct our covariance matrix and make sure the matrices are invertible and have reasonable condition numbers.

The second set of tests were associated with the emulators. Specifically, we ensure that the mean emulator error computed on the hold-out tests is subdominant compared to the statistical error computed from covariances. In principle, because we are adding the emulator error to the final covariance matrix, the emulator error should not bias our results. However, because the emulator error is intrinsically a function of model parameter values and our treatment of emulator errors (only looking at mean error) is approximate, we would like to make sure potential issues with the emulator error do not dominate our final error budget and lead to significant biases.

Finally, we conduct a suite of consistency tests outlined in the *Consistency Checks* section as the final part of our unmasking criteria. We explored different scale cuts and different HOD models until we found a configuration (model B and $r_p > 5\ h^{-1}$ Mpc) that yields acceptable goodness of fits and self-consistent results in our cross-validation tests. These validation tests are critical in increasing our confidence in our results. Again, all these steps were described in detail in S. Yuan et al. (2024a, 2024b).

The 68% marginalized cosmology constraints with our final results using our favored model are shown in Figures 1 and 2. The RD-*k*NN fit achieves a $\chi^2/{\rm dof} = 0.44$, whereas the DD-*k*NN fit achieves a $\chi^2/{\rm dof} = 1.23$.

Caveats. The main limitation is that we had to employ aggressive scale cuts because our extended-HOD models were not flexible enough to reliably model scales smaller than $5\ h^{-1}$ Mpc. In S. Yuan et al. (2024) we argue that this is specifically due to the inability to model the small-scale velocities correctly. Thus, limitations with our galaxy–halo connection modeling are our main area of concern and improvement.

Another issue that we were aware of before unmasking was the potential mischaracterization of emulator errors since we only accounted for the mean errors calculated on a limited set of hold-out tests. The issue with emulator errors could be mitigated either by running coverage tests or by following a likelihood-free inference approach. Both require a large amount of training data, which requires additional simulations.

5.6. Density-split Clustering[62]

In this subsection, we describe the methodology for analyzing the DSC statistics, based on an emulator trained on the `AbacusSummit` simulations, which can learn about the dependence of these statistics on cosmology and HOD parameters. More details about the implementation of this emulator can be found in C. Cuesta-Lazaro et al. (2023)

5.6.1. Data and Estimators

DSC (E. Paillas et al. 2021, 2023) is a method that characterizes the environmental dependence of galaxy clustering, exploiting the sensitivity of each environment to cosmology. We focus our analysis on measuring the DSC statistics on the 10 ΛCDM mocks in redshift space. Here we briefly summarize the algorithm to measure the DSC data vector.

We begin by painting the galaxy overdensity field to a rectangular grid of $5\ h^{-1}$ Mpc resolution. We smooth the field with a Gaussian kernel of width $R_s = 10\ h^{-1}$ Mpc, and then we

[62] Authors: E. Paillas, C. Cuesta-Lazaro.

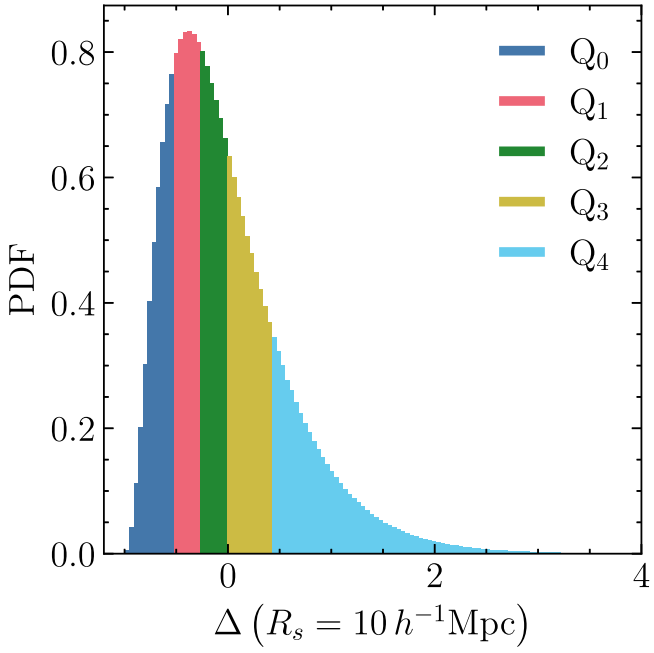

Figure 14. Probability distribution function of the galaxy overdensity measured at random query points from one of the ΛCDM simulations, where the galaxy density field has been smoothed with a Gaussian filter of radius $R_s = 10\,h^{-1}$ Mpc. The points are classified into five density quintiles, shown by the different colors.

sample it at $N_{\rm query}$ random query points, where $N_{\rm query}$ is equal to five times the number of galaxies. These query points are then split into five quintiles, according to the value of the overdensity at their locations. Figure 14 shows an example of the measured overdensity distribution and its splitting into quintiles, as measured from one of the ΛCDM simulations.

We proceed to measure the cross-correlation function between each quintile and the redshift-space galaxy field, as well as the ACF of the quintiles themselves. Both of these are measured in bins of s and μ, using 241 μ bins from -1 to 1 and radial bins of scale-dependent widths: $1\,h^{-1}$ Mpc width for $s \in [0,\ 4]\,h^{-1}$ Mpc, $3\,h^{-1}$ Mpc width for $s \in [4,\ 30]\,h^{-1}$ Mpc, and $5\,h^{-1}$ Mpc width for $s \in [30,\ 150]\,h^{-1}$ Mpc. The correlation functions are then decomposed into multipoles. Here we restrict the analysis to the monopole and quadrupole moments. Figure 15 shows the measured multipoles, averaged over the 10 simulations.

Finally, the data vector is constructed as a concatenation of the monopole and quadrupole moments of the quintile–galaxy cross-correlation functions, $\xi^{QG}_{0+2}(s)$, and the quintile ACFs, $\xi^{QQ}_{0+2}(s)$,

$$\widehat{\boldsymbol{d}}_{\rm DSC} = (\xi_0^{Q=(0,1,3,4)G}(s),\ \xi_2^{Q=(0,1,3,4)G}(s),\ \xi_0^{QQ=(00,11,33,44)},\ \xi_2^{QQ=(00,11,33,44)}(s)), \tag{29}$$

using the same radial binning as described above. When doing so, we discard the middle quintile, Q_2, before concatenating the data vector. The reason for this is that the information from all five quintiles is redundant, since the sum of all cross-correlation functions averages to zero by construction. Therefore, all the information available from this method is already contained in the remaining four quintiles.

5.6.2. Model

We model the DSC multipoles using a neural-network-based emulator that learns about the dependence of these multipoles on cosmology and HOD parameters. The emulator is trained on the `AbacusSummit` suite of simulations, using an extended-HOD parameterization, as described in detail in Section 5.5.

Parameterization and Simulations. We start from the dark matter halo catalogs from the base `AbacusSummit` simulation boxes at 85 different cosmologies, which span an eight-dimensional w_0w_aCDM parameter space, including the physical cold dark matter density $\omega_{\rm cdm}$, the physical baryon density $\omega_{\rm b}$, the rms of density fluctuations σ_8, the spectral index $n_{\rm s}$, the running of the spectral index α_s, the effective number of massless relics $N_{\rm eff}$, and the dark energy equation-of-state parameters w_0 and w_a. However, when we perform the cosmological inference, we restrict ourselves to a base-ΛCDM parameter space, using the priors on cosmological parameters defined in Equation (32).

We generate an LH with 100,000 samples of an extended-HOD parameter space, including velocity bias and environment-based secondary bias, using the prior range defined in Equation (33). Each cosmology is assigned 1000 HOD parameters from the LH, and dark matter halos are populated with HOD galaxies from those parameters, using the `AbacusHOD` code (S. Yuan et al. 2022b). When the resulting number density of an HOD catalog is larger than the average number density of the ΛCDM simulations ($n_{\rm gal} \approx 4.5 \times 10^{-4} (h\ \mathrm{Mpc}^{-1})^{-3}$), we invoke an incompleteness parameter $f_{\rm ic}$ and subsample the catalog to match the target number density. If the number density is lower than the target, we simply keep it as is. Galaxy positions are converted to redshift space by perturbing their real-space positions with their peculiar velocities along the LOS (taken to be one of the axes of the simulation box).

Neural Network Emulator. We split the collection of HOD catalogs into training, validation, and test sets. We construct two neural network emulators, one for each correlation type (quintile–galaxy cross-correlation and quintile autocorrelation). The inputs to the neural network are the cosmological and HOD parameters, whereas the outputs are the concatenated monopole and quadrupole for the four quintiles.

The networks are fully connected, using SiLU activation functions, a mean absolute error loss function, and an AdamW optimizer. To avoid overfitting due to the limited size of our training set, we also introduce dropout. To improve the performance of the model and reduce training time, we decrease the learning rate by a factor of 10 every five epochs over which the validation loss does not improve, until reaching a minimum learning rate of 10^{-6}. We use the `optuna` framework[63] to optimize the hyperparameters of the neural network, aiming at minimizing the validation loss. More details related to the architecture and its optimization can be found in our public repository.[64]

5.6.3. Inference

Likelihood. We fit our emulator to the mean measurement of the 10 realizations of the ΛCDM redshift-space mocks. We assume a Gaussian data likelihood with covariance $\mathsf{C}_{\rm tot}$, which incorporates the total error budget of our analysis.[65]

[63] Available at https://github.com/optuna/optuna.
[64] Available at https://github.com/florpi/sunbird,
[65] The validity of the Gaussian likelihood assumption for DSC is explored in Appendix A of C. Cuesta-Lazaro et al. (2024).

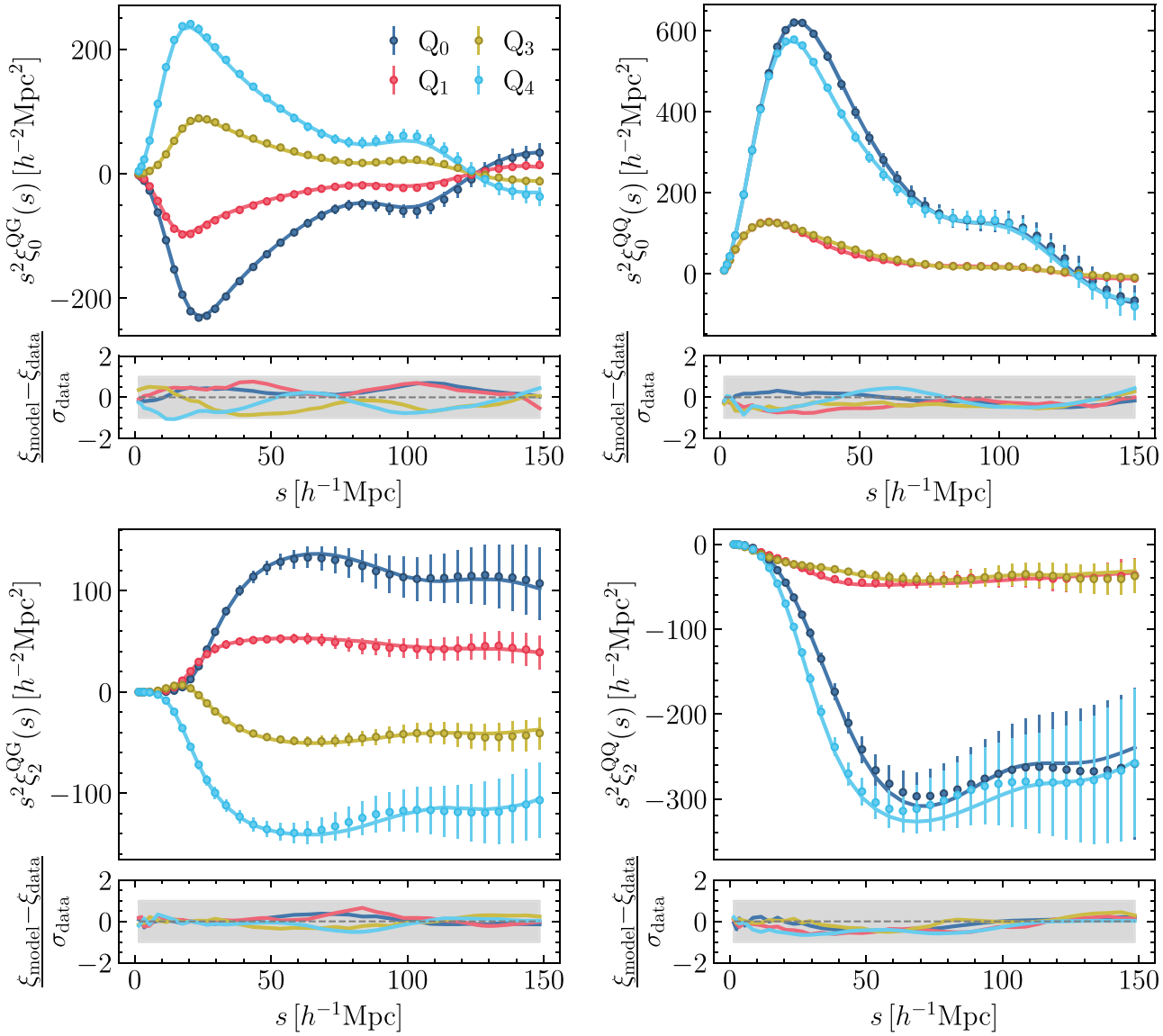

Figure 15. A visualization of the DSC data vectors, along with the best-fit model from our emulator. Markers and solid lines show the data vectors and the emulator predictions, respectively. Left: multipoles of the quintile–galaxy cross-correlation functions. Right: multipoles of the quintile ACFs. Top and bottom panels show the monopole and quadrupole moments, respectively. We also display the difference between the model and the data, in units of the data error. Each color corresponds to a different density quintile, as exemplified in Figure 14.

This covariance includes contributions from sample variance in the data vector ($\mathsf{C}_{\mathrm{data}}$), sample variance in the simulations used for training ($\mathsf{C}_{\mathrm{train}}$), and the intrinsic emulator error ($\mathsf{C}_{\mathrm{emu}}$):

$$\mathsf{C}_{\mathrm{tot}} = \mathsf{C}_{\mathrm{data}} + \mathsf{C}_{\mathrm{train}} + \mathsf{C}_{\mathrm{emu}}. \tag{30}$$

Covariance. To calculate $\mathsf{C}_{\mathrm{data}}$, we measure the DSC multipoles from 1500 realizations of the small `AbacusSummit` simulations, which are $500\,h^{-1}$ Mpc on a side and use the baseline `AbacusSummit` cosmology. We choose a combination of HOD parameters that produce multipoles that minimize the χ^2 with respect to the data vector measured from the ΛCDM simulations. We then divide the covariance by a factor of 64 to compensate for the volume difference between the small `AbacusSummit` boxes and the $2000\,h^{-1}$ Mpc ΛCDM boxes. We note that here we have assumed that the noise in the data vector is that associated with a single box of the ΛCDM simulations. To account for the finite volume of the simulations used for training, we repeat the same procedure to calculate $\mathsf{C}_{\mathrm{train}}$, so effectively $\mathsf{C}_{\mathrm{train}} = \mathsf{C}_{\mathrm{data}}$ for this analysis. Finally, to estimate $\mathsf{C}_{\mathrm{emu}}$, which accounts for the error due to an imperfect emulation of the multipoles, we compare the emulator predictions $\boldsymbol{m}(\boldsymbol{\Omega}^{\mathrm{test}})$ against the multipoles measured from our test set of simulations, $\widehat{\boldsymbol{d}}^{\mathrm{test}}$, and compute the covariance of the absolute emulator error $\Delta^{\mathrm{test}} = \boldsymbol{m}(\boldsymbol{\Omega}^{\mathrm{test}}) - \widehat{\boldsymbol{d}}^{\mathrm{test}}$ as

$$\boldsymbol{C}_{\mathrm{emu}} = \frac{1}{n_{\mathrm{test}} - 1}\sum_{k=1}^{n_{\mathrm{test}}}(\Delta^{\mathrm{test}_k} - \overline{\Delta^{\mathrm{test}}})(\Delta^{\mathrm{test}_k} - \overline{\Delta^{\mathrm{test}}})^{\top}, \tag{31}$$

where the overline denotes an average across test samples.

Figure 16 shows the correlation matrices associated with $\mathsf{C}_{\mathrm{data}}$ and $\mathsf{C}_{\mathrm{emu}}$, highlighting how the different quintiles, multipoles, and correlation types complement each other. We also display the ratio between the emulator and data covariance, where we can see that while the data error is the dominant source of error at large scales, the emulator error becomes larger for small separation bins.

Sampler and Priors. We sample the posterior distribution using the `Dynesty` nested sampler (J. S. Speagle 2020). We

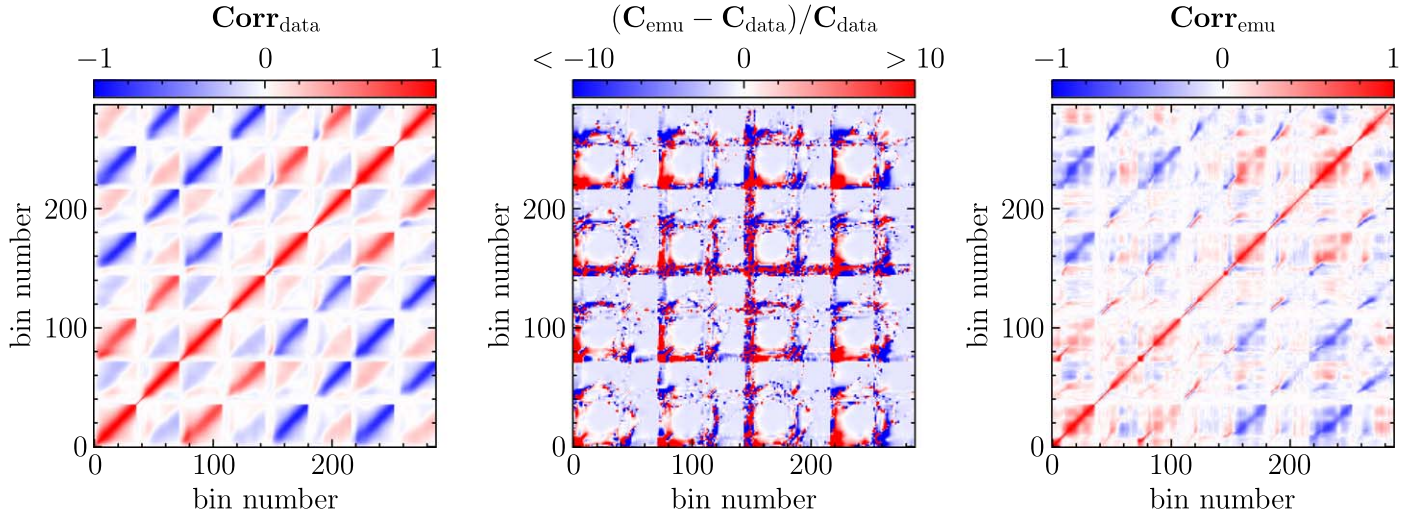

Figure 16. Error sources in the DSC analysis. Left: correlation matrix associated with the sample variance of the data vector. Right: correlation matrix associated with the imperfect emulation of the multipoles. Middle: relative change between the emulator and data covariance. The error associated with the finite size of the training sample, $\boldsymbol{C}_{\rm train}$, is the same as the data error in this case, so we do not display it.

assume the following flat priors on cosmological parameters:

$$\omega_{\rm cdm} \sim \mathcal{U}[0.1032, 0.140], \quad \sigma_8 \sim \mathcal{U}[0.678, 0.938],$$
$$n_s \sim \mathcal{U}[0.9012, 1.025], \quad \omega_{\rm b} \sim \mathcal{U}[0.0207, 0.0243]. \quad (32)$$

We fix those parameters beyond the base-ΛCDM model to the following default values: $w_0 = -1$, $w_a = 0$, $N_{\rm ur} = 2.0328$, and $\alpha_s = 0$. For the HOD parameters, we adopt the following flat priors:

$$\log_{10} M_{\rm cut} \sim \mathcal{U}[12.0, 13.5], \log_{10} M_1 \sim \mathcal{U}[12.5, 15.0],$$
$$\alpha \sim \mathcal{U}[0.3, 1.5],$$
$$\alpha_{v,\rm cen} \sim \mathcal{U}[0.0, 1.0], \alpha_{v,\rm sat} \sim \mathcal{U}[0.0, 2.0],$$
$$\log_{10} \sigma \sim \mathcal{U}[-7.0, 0.0],$$
$$\kappa \sim \mathcal{U}[0.0, 1.5], B_{\rm cen} \sim \mathcal{U}[-0.5, 0.5],$$
$$B_{\rm sat} \sim \mathcal{U}[-1.0, 1.0]. \quad (33)$$

5.6.4. Analysis Choices

To validate the baseline setup of our analysis, we run our inference pipeline adopting different choices of scales, summary statistics, HOD models, and treatment of errors. The results are summarized in Figure 17. Our baseline configuration, shown at the top of each panel, uses scales between $1.0\,h^{-1}$ Mpc $< s <$ $151\,h^{-1}$ Mpc, including the monopole and quadrupole of the quintile–galaxy CCF and the quintile ACF, using quintiles Q_0, Q_1, Q_3, and Q_0. The extended-HOD model includes both velocity and assembly bias, and the emulator systematic error, as well as the variance associated with our finite training sample, is included in the likelihood calculation. We see that adjusting these choices one by one produces changes in the recovered parameter constraints, but, overall, all such variations are consistent within 1σ of the values inferred from our fiducial configuration, highlighting the robustness of the recovered constraints against the tuning of these settings. Below, we comment on each of these aspects, based on the results from Figure 17.

Scale Cut Validation. We observe that the recovered constraints do not change significantly when excluding the very small scale information at $s < 30\,h^{-1}$ Mpc. This is a direct consequence of the inclusion of the systematic emulator error in our analysis, which starts to dominate the error budget on those scales. We also see that the inclusion of large-scale information adds a significant amount of constraining power. Moreover, we have run independent recovery tests on our test set from the `AbacusSummit` simulations, which further confirm that our fiducial range $1\,h^{-1}$ Mpc $< s < 151\,h^{-1}$ Mpc seems robust at recovering unbiased cosmological constraints for the volume of the ΛCDM simulations. Based on this, we choose to adopt this scale range for the baseline analysis.

Choice of Data Vector. Concerning the DSC statistics, the quintile–galaxy CCF amounts for most of the information, with the quintile ACF only adding a small contribution to the budget. In terms of the multipoles, the constraints are mostly driven by the monopole. We also see that the quintiles at the extrema, Q_0 and Q_4, carry most of the information. This is in line with previous findings based on Fisher forecasts for DSC statistics (E. Paillas et al. 2023).

Choice of HOD Model. The choice of the HOD model has little impact on $\omega_{\rm cdm}$ but becomes noticeable for σ_8 and $n_{\rm s}$. More specifically, not including central and satellite velocity bias in the model shifts the mean of the posterior for σ_8 and $n_{\rm s}$ to smaller values, although well within the 1σ region of the baseline setup. A similar trend is seen when excluding assembly bias. Given the findings of previous studies, which suggest that velocity and assembly bias are important for accurately fitting the small-scale clustering from observational data and simulations (H. Guo et al. 2015; S. Yuan et al. 2022a, 2022b), and to ensure that our emulator is robust against the potential presence of such effects in the ΛCDM simulations fitted in this study, we choose to include them in the default analysis.

Choice of Error Budget. We see that the exclusion of the emulator systematic error has an important effect in the analysis. Removing this error results in a dramatic increase of the precision at which we recover the spectral tilt $n_{\rm s}$, and a smaller effect is observed for $\omega_{\rm cdm}$ and σ_8. A similar trend is spotted for the exclusion of the error associated with the training sample, although in a lesser extent. Based on separate tests with the `AbacusSummit` simulations, we have concluded that this increase of constraining power comes at the expense of potential biases in the recovered cosmological parameters (C. Cuesta-Lazaro et al. 2024). Therefore, to ensure the robustness of our inference analysis, we include both error contributions in the likelihood calculation.

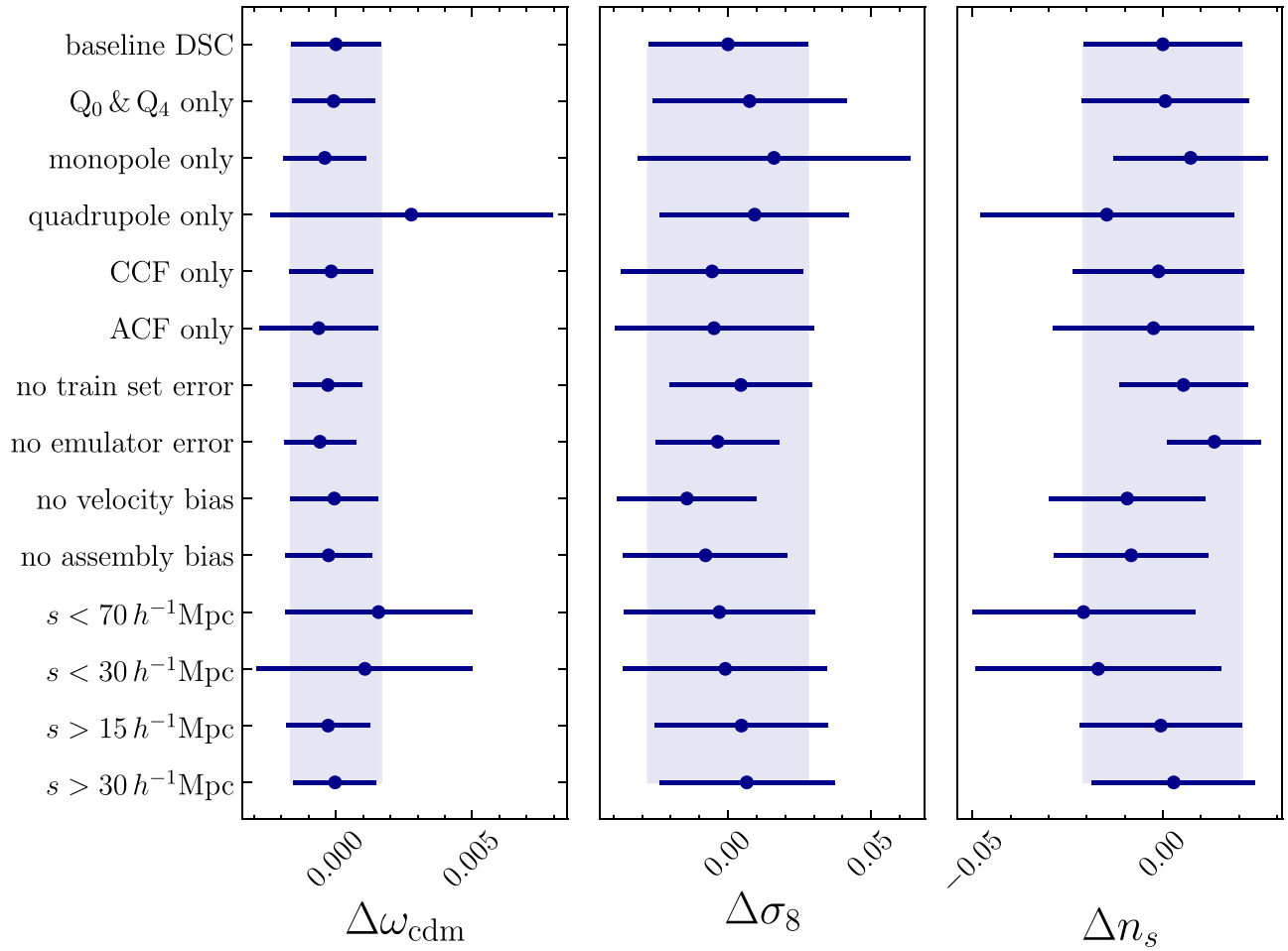

Figure 17. Masked constraints from DSC, assuming different configurations for the analysis. Blue circles show the mean of the marginalized posterior for each parameter (relative to the mean of the baseline constraints), and the error bars show the associated 68% confidence interval. The blue shaded region shows the constraints for the baseline configuration adopted in the paper.

Choice of Likelihood. The choice of a Gaussian likelihood in our analysis is motivated by previous tests we have carried out with the `Quijote` and `AbacusSummit` simulations, which show that the likelihood associated with the DSC data vector is indeed close to Gaussian for the scales and the volume of interest (see, e.g., Figure C1 in E. Paillas et al. 2023).

Unmasking Criteria. Having explored the different choices of settings in our inference pipeline, we have found that the resulting cosmological constraints are robust against such choices. Moreover, we have verified that our best-fit model provides a good χ^2 to the measured data vector. Therefore, we allow our results to be unmasked, adopting the baseline configuration for our analysis described in the previous section.

Caveats. Possible caveats that could lead to unexpected results in our analysis after unmasking include the following:

1. HOD parameters lying outside the range of our priors.
2. An HOD model including effects that were not implemented in our emulator, such as satellite profile flexibilities to account for baryonic effects.
3. An overly conservative error budget in our likelihood, mainly related to the systematic error coming from the emulator predictions. We have chosen to add the emulator error calculated across the full prior range of our parameters, which could lead to an overestimation of the error bars at small scales and, consequently, to an unnecessary degradation of the precision of our constraints.
4. Previous analyses have shown that DSC is particularly sensitive to AP distortions (E. Paillas et al. 2021), which can increase the relative gain of information with respect to standard 2pt measurements. Here we have worked directly with the true galaxy positions in redshift space, ignoring the AP effect, which could potentially degrade the relative improvement in parameter constraints found in previous studies.
5. Due to computational constraints, we have not explored changing the number of quantiles that the galaxy density probability distribution function (PDF) is split into, or the width of the kernel that is used to smooth the density field. We followed choices made in previous analyses that were calibrated for the number density and redshift of BOSS CMASS. These choices could be suboptimal for the galaxy sample used in this study, leading to weaker constraints than what could be achieved when varying those settings.

5.7. Cosmic Voids[66]

Cosmic voids, the extended underdense regions of the cosmic web, are becoming an increasingly important probe in the galaxy clustering and cosmology communities (A. Pisani et al. 2019; M. Moresco et al. 2022). While originating in the

[66] Authors: S. Contarini, G. Verza, N. Hamaus, A. Pisani.

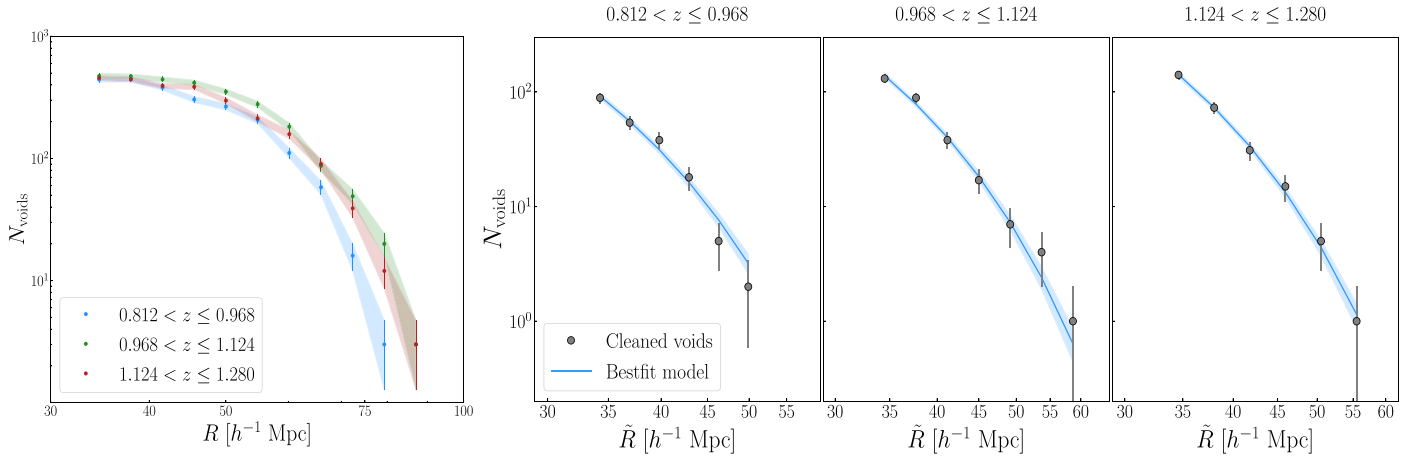

Figure 18. The first panel shows the number of voids as a function of their effective radius R (pre-cleaned VSF), identified with VIDE in three equispaced redshift bins (see legend). The second, third, and fourth panels show the number counts of voids used for the VSF analysis, after the cleaning procedure, as a function of the cleaned radius $\tilde{R}$ (for details on the procedure, see Section 5.7.2), measured in three equispaced redshift bins. We represent the data with circles and Poisson error bars. We show in blue the best-fit model derived with the Bayesian analysis described in Section 5.7.3. The solid line indicates the median of the model, while the shaded area covers from the 16th to the 84th percentile.

3D distribution of matter, voids can also be identified in tracers of the latter, such as galaxies. In this work we focus on two summary statistics: the VSF and the VGCF. The VSF specifies the number of voids as a function of their size and is sometimes referred to as void abundance. The VGCF describes the spatial clustering between the centers of voids and their surrounding galaxies in the form of a cross-correlation function. In the case where all spatial separations are expressed in units of individual void sizes, the VGCF is sometimes simply denoted as a void *stack*. Both techniques have already been applied to SDSS BOSS data (K. S. Dawson et al. 2013) to derive constraints on cosmological parameters (e.g., N. Hamaus et al. 2016, 2020; S. Contarini et al. 2023, 2024). In this challenge we consider both the VSF and the VGCF measured in the mock light cone.

5.7.1. Data Vector and Estimators

We rely on the popular void-finding algorithm VIDE[67](P. M. Sutter et al. 2014b), which is based on ZOBOV (M. C. Neyrinck 2008) to estimate a tracer density field via Voronoi tessellation and makes use of the watershed technique (E. Platen et al. 2007) to identify local depressions therein. VIDE has been used extensively for void analyses in both simulations (e.g., K. C. Chan et al. 2014; N. Hamaus et al. 2014b; G. Pollina et al. 2017; S. Contarini et al. 2019; N. Schuster et al. 2019; G. Verza et al. 2019; C. D. Kreisch et al. 2022) and observational data (e.g., P. M. Sutter et al. 2012; N. Hamaus et al. 2016; Y. Fang et al. 2019; G. Pollina et al. 2019; S. Contarini et al. 2023, 2024). In order to run on light-cone data, VIDE requires a mask to specify the (mock) survey footprint. For this we generate a healpix mask of resolution $n_{\rm side} = 128$ from the angular distribution of mock galaxy coordinates provided in the catalog. This resolution is sufficient to capture the simple rectangular geometry with the given ranges in R.A. and decl. The final data product of VIDE consists of various catalogs containing different void properties, such as their sky coordinates, redshifts, effective radii, shape, and density information. Because VIDE identifies voids in comoving space, angles and redshifts are transformed to comoving coordinates assuming a fiducial cosmology. Therefore, void properties defined in comoving space, such as effective radii, volumes, and densities, depend on this assumption. The same fiducial cosmology is used to transform the comoving coordinates of voids back to sky coordinates, which are therefore less affected by model assumptions. Here we focus our analysis on the ΛCDM light-cone mock catalog, where VIDE identifies more than 14,000 voids.

VSF. The data vector for the VSF is obtained by measuring the number of voids found by VIDE in logarithmic bins of radius. In our analysis we split the void sample into three equispaced bins of redshift, i.e., $0.812 < z \leqslant 0.968$, $0.968 < z \leqslant 1.124$, and $1.124 < z \leqslant 1.280$. The division into multiple bins allows us to capture the redshift evolution. While different binning options have been tested (finding consistent results), the subdivision in three bins provides a good balance in terms of number of bins and void statistics within each bin.

We present in Figure 18 (first panel) the number of identified voids as a function of their effective radius R, for the three aforementioned redshift bins. R is defined as the radius of a sphere whose volume equals that of the void, which is nonspherical in general. To align the measured VSF with the theoretical model and remove spurious voids, subsequent cleaning is performed. The cleaning procedure is described below. The VSF data vector is given by

$$\hat{\boldsymbol{d}}_{\rm VSF} = [\hat{N}(\tilde{R}_{i=1,\ldots,N_r}, z_0), \hat{N}(\tilde{R}_{i=1,\ldots,N_r}, z_1), \hat{N}(\tilde{R}_{i=1,\ldots,N_r}, z_2)], \quad (34)$$

where $\tilde{R}_{i=1,\ldots,N_r}$ denotes the N_r bins in cleaned void size and $z_{0,1,2}$ denotes the three tomographic bins described above.

VGCF. The data vector of the VGCF is constructed by counting the number of galaxies around each void center as a function of their separation in directions along and perpendicular to the LOS, $s_\parallel$ and $s_\perp$, measured in units of the effective radius R. A stack over a sample of voids then yields a cross-correlation function between void centers and galaxies (N. Hamaus et al. 2015). We use the Landy–Szalay estimator (S. D. Landy & A. S. Szalay 1993) to obtain a measurement of the VGCF data vector via

$$\hat{d}_{{\rm VGCF},i} \equiv \hat{\xi}_{{\rm vg},i}(s_\perp, s_\parallel), \quad (35)$$

[67] Available at https://bitbucket.org/cosmicvoids/vide_public/wiki/Home.

where we use 15 linearly spaced bins in each direction, ranging from the void center to about twice the effective void radius. The randoms are constructed by sampling from the smoothed redshift distributions of galaxies and voids identified in the mock data, and uniformly across the sky, with an oversampling factor of 10 (we have checked that a factor of 50 yields indistinguishable results). In order to impose a sky footprint identical to that in the mock data, we apply the same healpix mask to the randoms.

5.7.2. Model

VSF Parameterization. The VSF model, developed by R. K. Sheth & R. van de Weygaert (2004) and further modified in E. Jennings et al. (2013) to account for the volume conservation in the transition to nonlinearity, relies on the excursion-set theory. The so-called *Vdn* (*volume-conserving*) model is defined as

$$\frac{dn}{d\ln R} = \frac{f_{\ln\sigma_R}(\sigma_R)}{V(R)}\,\frac{d\ln\sigma_R^{-1}}{d\ln R_L}\Big|_{R_L=R_L(R)}, \tag{36}$$

where σ_R represents the rms variance of linear matter perturbations on a scale R_L and where, to compute the number density of voids, we rely on the multiplicity function:

$$f_{\ln\sigma_R}(\sigma_R) = 2\sum_{j=1}^{\infty}\exp\left(-\frac{(j\pi x)^2}{2}\right)j\pi x^2\,\sin(j\pi D),$$
$$\text{with}\quad D\equiv\frac{|\delta_v^L|}{\delta_c^L+|\delta_v^L|}\quad\text{and}\quad x\equiv\frac{D}{|\delta_v^L|}\sigma_R. \tag{37}$$

Considering linear theory, the function $f_{\ln\sigma_R}(\sigma_R)$ represents the volume fraction of the Universe occupied by cosmic voids, with radii in the range $(R, R+dR)$. The additional cosmological dependence of this function lies in the density contrasts defining the formation of dark matter halos and cosmic voids, δ_c^L and δ_v^L, respectively. To predict the number of voids identified from observations of biased tracers, a number of effects have to be considered:

(i) the link of δ_v^L to its nonlinear counterpart $\delta_c^{\rm NL}$ (corresponding to the same threshold but considering nonlinear theory) in the biased tracer field;
(ii) the dynamic distortions caused by the peculiar velocities of the tracers;
(iii) the impact of geometric distortions.

First, to account for the link of δ_v^L to its nonlinear counterpart in the biased tracer field, the sample of voids analyzed has to be prepared to align with the definition of voids given by the VSF theory. For this purpose, we apply to the void catalog built with `VIDE` the cleaning algorithm developed by T. Ronconi & F. Marulli (2017), publicly available in the libraries `CosmoBolognaLib`[68] (F. Marulli et al. 2016). The output of this procedure is a catalog of nonoverlapping spheres with cleaned void radius $\tilde{R}$, embedding a fixed density contrast in the tracer density field, $\delta_{v,\rm tr}^{\rm NL}$. We chose to fix the latter to $\delta_{v,\rm tr}^{\rm NL}=-0.7$. While different values of this threshold can be used, the selected value leads to a good compromise. Indeed, as discussed, for example, in S. Contarini et al. (2022), a more negative threshold implies a decrease in the effective void radius, which in turn reduces the void sample since only the deepest regions can fulfill the requirement. Moreover, it provides voids with fewer tracers, therefore increasing the uncertainty on the radius itself. On the other hand, a less negative threshold provides a void sample composed of larger voids (more likely to overlap and to be excluded from the cleaned sample) and including a higher number of shallower voids, enhancing the likeliness of sample contamination by spurious underdensities (see M. C. Neyrinck 2008; M. C. Cousinou et al. 2019).

We then use a bias relation to convert the density contrast in the tracer density field, $\delta_{v,\rm tr}^{\rm NL}$, to the corresponding one in the matter distribution, $\delta_v^{\rm NL}$. Galaxy bias in voids cannot be properly described by the large-scale effective galaxy bias (S. Contarini et al. 2019; G. Verza et al. 2022). To model it, we therefore rely on a linear function of the effective bias (S. Contarini et al. 2019, 2021, 2022, 2023, 2024):

$$\delta_v^{\rm NL} = \frac{\delta_{v,\rm tr}^{\rm NL}}{\mathcal{F}(b_{\rm eff},\,\sigma_8)},\ \text{with}$$
$$\mathcal{F}(b_{\rm eff}\,\sigma_8) = C_{\rm slope}\,b_{\rm eff}\,\sigma_8 + C_{\rm offset}. \tag{38}$$

Here $b_{\rm eff}$ represents the galaxy effective bias, and in this analysis, the combined quantity $b_{\rm eff}\sigma_8$ is derived from the galaxy 2pt correlation function and inserted in the VSF model, marginalizing over it. In particular, we used the S. D. Landy & A. S. Szalay (1993) estimator to measure the 2D redshift-space 2pt correlation function, and we selected its first three nonnull multipole moments. Then, we modeled the 2pt correlation function multipoles relying on the prescriptions of A. Taruya et al. (2010) (see Appendix A of S. Contarini et al. 2023, for further details). Second, tracer peculiar velocities cause an enlargement along the observer's LOS, further increasing the mean void radius (A. Pisani et al. 2015b; N. Hamaus et al. 2020; G. Verza et al. 2023). The above parameterization is used to also correct for this enlargement (S. Contarini et al. 2022), allowing us to statistically align voids observed in the redshift-space galaxy distribution to theoretical voids, identified in real space with unbiased tracers. In Equation (38), $C_{\rm slope}$ and $C_{\rm offset}$ are redshift-independent coefficients of the linear function $\mathcal{F}$ and can be calibrated by using simulations or, as in this case, marginalized over. Adopting the conservative choice of using wide uniform priors on these parameters, we are able to constrain only the parameter $\Omega_{\rm m}$. We underline, however, that exploiting the information derived from mock catalogs to limit the priors of $C_{\rm slope}$ and $C_{\rm offset}$ would allow us to provide cosmological constraints on σ_8 as well (S. Contarini et al. 2023).

Third, we account for the AP effect, i.e., geometric distortions due to mismatch between the fiducial and the true cosmology. Following the prescriptions in A. G. Sánchez et al. (2017), N. Hamaus et al. (2020), and C. M. Correa et al. (2021), we rescale the observed void radius R^* to $R = q_\parallel^{1/3}q_\perp^{2/3}R^*$, where $q_\parallel$ and $q_\perp$ are defined via

$$r_\parallel = \frac{H^*(z)}{H(z)}\,r_\parallel^* \equiv q_\parallel\,r_\parallel^*,$$
$$r_\perp = \frac{D_A(z)}{D_A^*(z)}\,r_\perp^* \equiv q_\perp\,r_\perp^*, \tag{39}$$

and the asterisk indicates quantities evaluated in the fiducial cosmology. Here $r_\parallel^*$ and $r_\perp^*$ are the observed comoving

[68] Available at https://gitlab.com/federicomarulli/CosmoBolognaLib.

distances between two points at redshift z, projected along the parallel and perpendicular LOS directions, respectively. $H(z)$ is the Hubble parameter, and $D_A(z)$ is the comoving angular diameter distance.

The final VSF model depends on cosmology through the quantities σ_R, $q_\parallel$, and $q_\perp$, i.e., those parameters determining the amplitude of the matter power spectrum and cosmological distances. Additionally, the model depends on the nuisance parameters $C_{\rm slope}$ and $C_{\rm offset}$, as well as on the effective bias of tracers. The latter is estimated from their 2pt correlation function and marginalized over. Hence, the model space that we consider for the VSF is spanned by seven parameters: $\Omega_{\rm m}$, σ_8, h, $\Omega_{\rm b}h^2$, $n_{\rm s}$, $C_{\rm slope}$, and $C_{\rm offset}$. In addition, we assume one massive neutrino species of minimal mass. As already mentioned, in this analysis we will provide constraints on $\Omega_{\rm m}$ only, because no calibration of the parameters $C_{\rm slope}$ and $C_{\rm offset}$ is available with the cosmic tracers used in this work. Figure 18 presents measurements of the pre-cleaning VSF and of the cleaned VSF including best-fit models in the considered redshift bins.

VGCF Parameterization. We rely on three fundamental assumptions to model the VGCF. First, we make use of the cosmological principle, implying that stacked voids are spherically symmetric and hence statistically isotropic in real space. This means that they can be used as *standard spheres* to measure distance ratios via the AP effect. Second, we assume that the peculiar velocity field $\boldsymbol{u}_v$ around the void center follows linear dynamics according to the linearized continuity equation (P. J. E. Peebles 1980),

$$\boldsymbol{u}_v(\boldsymbol{r}) = -f(z)\frac{H(z)}{1+z}\frac{\boldsymbol{r}}{r^3}\int_0^r \delta(r')r'^2\,dr', \tag{40}$$

where $f(z)$ is the linear growth rate of the matter density contrast δ. Peculiar velocities cause voids to appear anisotropic in redshift space, because the Doppler effect contributes an additional component to the cosmological redshift along the $\boldsymbol{r}_\parallel$ direction. In order to successfully exploit voids as standard spheres, these RSDs have to be accurately modeled. For this purpose, the validity of Equation (40) has been thoroughly tested in simulations and was found to be extremely accurate in void environments (N. Hamaus et al. 2014a; N. Schuster et al. 2023), opening up the opportunity to use the AP test with voids as a precision probe of cosmology. However, in this mock challenge we do not have access to the full matter distribution, so our third assumption relies on a linear bias relation between the galaxy and matter density contrasts, such that $\xi_{\rm vg}(r) = b\delta(r)$. This relation has been verified in various simulation studies (P. M. Sutter et al. 2014a; G. Pollina et al. 2017, 2019; S. Contarini et al. 2019; T. Ronconi et al. 2019; N. Schuster et al. 2023), finding that the proportionality constant b asymptotes to the linear large-scale tracer bias b_1 for large voids, while smaller voids tend to yield higher but scale-independent values.

With these ingredients we can model the coordinate transformation between the real-space vector $\boldsymbol{r}$ and redshift-space vector $\boldsymbol{s}$ for the separation of void centers and galaxies,

$$\boldsymbol{s} = \boldsymbol{r} + \frac{1+z}{H(z)}\boldsymbol{u}_\parallel = \boldsymbol{r} - \frac{f(z)}{b(z)}\frac{\boldsymbol{r}_\parallel}{r^3}\int_0^r \xi_{\rm vg}(r')r'^2 dr'. \tag{41}$$

This requires the VGCF in real space, $\xi_{\rm vg}(r)$, which is not directly observable. We can, however, observe the LOS-projected VGCF, $\tilde{\xi}_{\rm vg}(s_\perp) = \int \xi_{\rm vg}(\boldsymbol{s})ds_\parallel$, which is insensitive to RSDs, as it only depends on separations $s_\perp$ on the plane of the sky. From it we obtain the real-space VGCF by deprojection via the inverse Abel transform (A. Pisani et al. 2014; A. J. Hawken et al. 2017),

$$\xi_{\rm vg}(r) = -\frac{1}{\pi}\int_r^\infty \frac{d\tilde{\xi}_{\rm vg}(s_\perp)}{ds_\perp}(s_\perp^2 - r^2)^{-1/2}ds_\perp. \tag{42}$$

Equations (40)–(42) fully specify the redshift-space VGCF at linear order in perturbation theory, and a closed-form expression for $\xi_{\rm vg}(s_\perp, s_\parallel)$ can be derived with the Jacobian of Equation (41) (Y.-C. Cai et al. 2016; N. Hamaus et al. 2017, 2020). In order to account for systematic effects, such as inaccuracies in the deprojection procedure and selection effects in the void sample due to sparse sampling of tracers with nonlinear RSDs (see caveats below), we augment this model with two additional nuisance parameters $\mathcal{M}$ (for monopole-like) and $\mathcal{Q}$ (for quadrupole-like) and adopt the semiempirical expression derived in N. Hamaus et al. (2022),

$$\begin{aligned}&\xi_{\rm vg}(s_\perp, s_\parallel)\\ &= \mathcal{M}\left\{\xi_{\rm vg}(r) + \frac{f}{b}\bar{\xi}_{\rm vg}(r) + 2\mathcal{Q}\frac{f}{b}\mu^2[\xi_{\rm vg}(r) - \bar{\xi}_{\rm vg}(r)]\right\},\end{aligned} \tag{43}$$

where $\mu = r_\parallel/r$ and $\bar{\xi}_{\rm vg}(r) = 3r^{-3}\int_0^r \xi_{\rm vg}(r')r'^2\,dr'$. We make use of Equation (41) to map the coordinates from observed redshift space to real space, where the model is evaluated, and include the AP parameters to account for geometric distortions,

$$r_\perp = q_\perp s_\perp, \qquad r_\parallel = q_\parallel s_\parallel\left[1 - \frac{1}{3}\frac{f}{b}\mathcal{M}\,\bar{\xi}_{\rm vg}(r)\right]^{-1}, \tag{44}$$

which can be solved via iteration upon setting an initial value of $r = s$ (N. Hamaus et al. 2020). Because we measure void-centric distances in units of the effective void radius R, which scales as $q_\parallel^{1/3}q_\perp^{2/3}$ with the AP parameters, only ratios of $q_\perp$ and $q_\parallel$ appear in Equation (44). Hence, the final model space for the VGCF is spanned by four parameters: f/b, $q_\perp/q_\parallel$, $\mathcal{M}$, and $\mathcal{Q}$. To summarize, our modeling of the VGCF involves the following steps:

(i) estimate the real-space VGCF via deprojection of its LOS-projected counterpart using Equation (42);
(ii) account for dynamic (redshift-space) distortions along the LOS using Equation (41);
(iii) account for geometric (AP) distortions using Equation (44); and
(iv) account for systematic effects arising via the deprojection procedure, or via selection effects due to sparse sampling and nonlinear RSD, by including nuisance parameters for the monopole and quadrupole terms in Equation (43).

Figure 19 depicts a representation of our data vector, the VGCF in separations along and perpendicular to the LOS, along with the best-fit model based on Equation (43).

5.7.3. Inference

VSF Covariance. We assume the off-diagonal terms of the VSF covariance matrix to be negligible on the basis of previous works, such as A. E. Bayer et al. (2021), C. D. Kreisch et al. (2022),

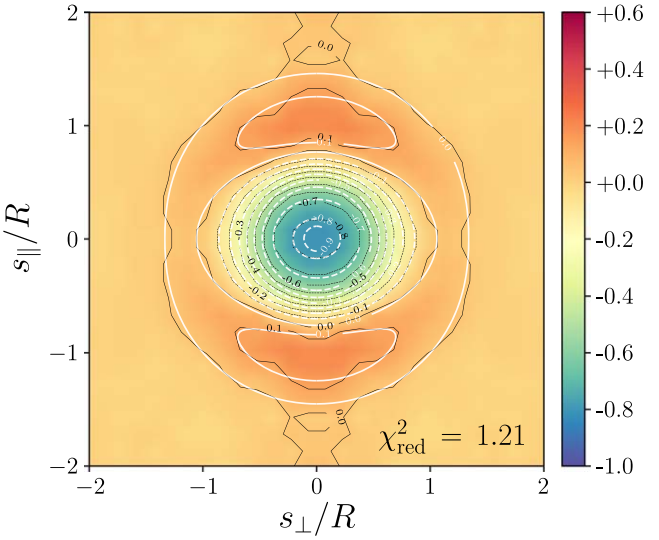

Figure 19. VGCF in separations along and perpendicular to the LOS, extracted from the entire ΛCDM light cone (color scale with black contour lines). White contours show the best-fit model with a reduced χ^2 value shown in the lower right corner.

Table 4
Parameters and Prior Bounds for the VSF and VGCF Analyses

Parameter	Prior
VSF	
$\Omega_{\rm m}$	$\mathcal{U}[0.05, 0.9]$
σ_8	$\mathcal{U}[0.2, 2]$
h	$\mathcal{U}[0.657, 0.683]$
$\Omega_{\rm b}h^2$	$\mathcal{U}[0.0216, 0.0224]$
$n_{\rm s}$	$\mathcal{U}[0.9535, 0.9763]$
$C_{\rm slope}$	$\mathcal{U}[-50, 10]$
$C_{\rm offset}$	$\mathcal{U}[-10, 50]$
VGCF	
$\Omega_{\rm m}$	$\mathcal{U}[0, 1]$
f/b	$\mathcal{U}[-10, 10]$
$q_\perp/q_\parallel$	$\mathcal{U}[-10, 10]$
$\mathcal{M}$	$\mathcal{U}[-10, 10]$
$\mathcal{Q}$	$\mathcal{U}[-10, 10]$

D. Pelliciari et al. (2023), and S. Contarini et al. (2023). For its diagonal elements, we assume Poisson statistics.

VSF Likelihood Function. Since the VSF is based on the number counts of cosmic voids, we assume the likelihood to follow the Poisson statistic (M. Sahlén et al. 2016; L. Thiele et al. 2024):

$$\mathcal{L}[\hat{\boldsymbol{d}}_{\rm VSF}|\boldsymbol{\Omega}] = \prod_i \frac{m_{{\rm VSF},i}(\boldsymbol{\Omega})^{\hat{d}_{{\rm VSF},i}} \exp[-m_{{\rm VSF},i}(\boldsymbol{\Omega})]}{\hat{d}_{{\rm VSF},i}!}, \quad (45)$$

with model parameters $\boldsymbol{\Omega}$ listed in Table 4. We use the functions implemented in the `CosmoBolognaLib` to sample the likelihood.

VSF Priors. We impose wide uniform priors for $\Omega_{\rm m}$ and σ_8, i.e., $\mathcal{U}[0.05, 0.9]$ and $\mathcal{U}[0.2, 2]$, respectively. We impose instead restricted priors on the remaining cosmological parameters, using the uncertainty provided by Planck2018 (Planck Collaboration et al. 2020) multiplied by a factor of three. The latter prior choice is aimed at reducing the parameter-space volume on those parameters that would be weakly constrained by the VSF. This allows us to increase the speed of our MCMC without introducing biased results. Then, we marginalize over the nuisance parameters $C_{\rm slope}$ and $C_{\rm offset}$, assigning uniform priors as well. For a summary of all our parameters and priors, see Table 4.

VGCF Covariance. We estimate the covariance matrix C of the VGCF by means of jackknife resampling the selected void sample, which is spatially nonoverlapping. To this end, we apply Equation (35) after removal of one void at a time and calculate the covariance over all jackknife samples.

VGCF Likelihood Function. We adopt a Gaussian likelihood for the VGCF, which has previously been validated on the QPM (M. White et al. 2014) and PATCHY (F.-S. Kitaura et al. 2016) mocks in N. Hamaus et al. (2017, 2020). We apply the Hartlap correction (J. Hartlap et al. 2007) to estimate the inverse covariance matrix and maximize the likelihood with respect to the parameter vector $\boldsymbol{\Omega} = (f/b, q_\perp/q_\parallel, \mathcal{M}, \mathcal{Q})$. To sample the posterior probability distribution of these parameters, we use the affine-invariant MCMC ensemble sampler `emcee` (D. Foreman-Mackey et al. 2019). Two parameters of our model explicitly depend on cosmology: one via the growth rate, which can be approximated as $f(z) \simeq \Omega_m^{0.6}(z)$ (O. Lahav et al. 1991), and the other via the AP parameter $q_\perp/q_\parallel \propto D_A(z)H(z)$. The latter is particularly well constrained by the AP effect from voids and is therefore a sensitive probe of the expansion history (G. Lavaux & B. D. Wandelt 2012; N. Hamaus et al. 2015). In a flat ΛCDM cosmology (neglecting the presence of neutrinos and radiation), the product $D_A(z)H(z)$ is fully determined by the parameter $\Omega_{\rm m}$. We obtain a posterior on the latter by sampling from a Gaussian likelihood containing measurements of $q_\perp/q_\parallel$ and their uncertainty within each redshift bin.

VGCF Priors. We impose identical uniform priors for each of our model parameters, $\boldsymbol{\Omega} \sim \mathcal{U}[-10, 10]$. These prior boundaries are wide enough to be uninformative, and we have checked that our results are insensitive to this particular choice. For our cosmological parameter of interest in the ΛCDM light-cone mock, we further impose $\Omega_m \sim \mathcal{U}[0, 1]$.

5.7.4. Analysis Choices

VSF Scale Cut Validation. We select voids above the cleaned radius $\tilde{R}_{\rm min} = 33\ h^{-1}$ Mpc, for all the considered redshift bins, i.e., $0.812 < z \leqslant 0.968$, $0.968 < z \leqslant 1.124$, and $1.124 < z \leqslant 1.280$. This selection is applied to avoid those spatial scales affected by a loss of void counts and depends on the resolution of the catalog. We measure the void counts as a function of the cleaned radius fixing the value of $\tilde{R}_{\rm min}$, and the numbers of resulting nonempty radius bins are [6, 7, 6], respectively. This is shown in Figure 18, where we represent the measured void abundances and the VSF best-fit model derived by the MCMC analysis. The agreement between data and theory is such that all measurements are reproduced by the model within 68% uncertainty.

VSF Consistency Checks. We performed a number of tests to check the stability of the results with the choice of the minimum cleaned void radius, i.e., $\tilde{R}_{\rm min}$. As expected, when increasing the value of $\tilde{R}_{\rm min}$ (and thus reducing the number of voids analyzed), the constraining power is reduced. This is

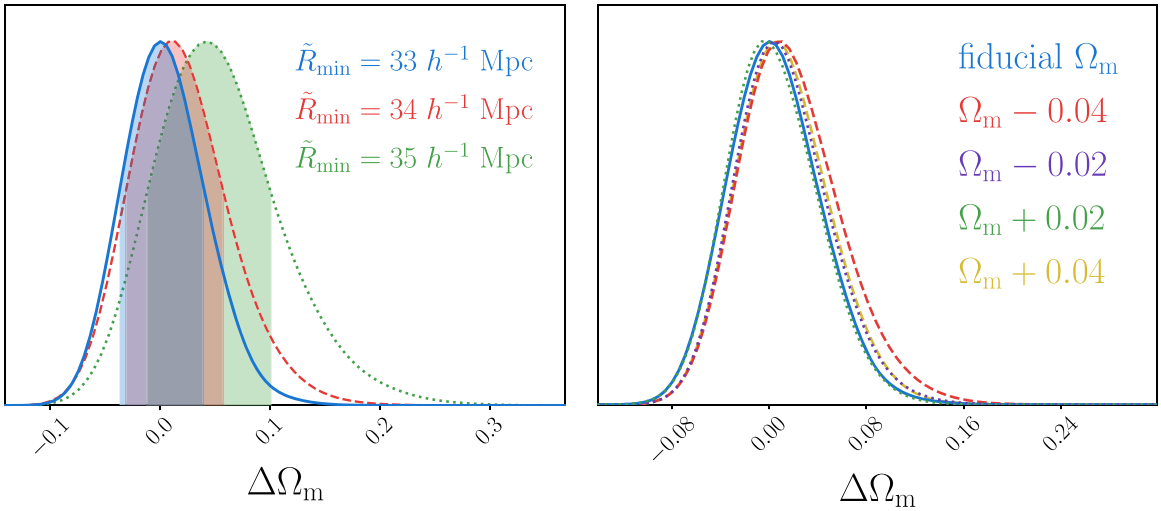

Figure 20. Posterior distribution for the value of $\Omega_{\rm m}$ based on the modeling of the VSF in the ΛCDM light cone. The mean of the reference posterior (in blue in each panel) has been subtracted to mask the results. We show in the left panel the impact of different cuts on void radius, indicating with shaded regions the 68% confidence interval around the maximum of the posterior distribution. The right panel illustrates the effect of varying the fiducial $\Omega_{\rm m}$ value in our analysis pipeline.

shown in the left panel of Figure 20, where we report the posterior distribution of $\Omega_{\rm m}$ derived by imposing $\tilde{R}_{\rm min} = [33, 34, 35]\ h^{-1}$ Mpc. Over a certain scale, that is, above the region where the void counts incompleteness is stronger, the cosmological constraints on $\Omega_{\rm m}$ are all consistent within 68% of uncertainty. However, we note a moderate dependency of the best-fit value of $\Omega_{\rm m}$ on the choice of $\tilde{R}_{\rm min}$, which could impact the robustness of our results (see caveats below).

We also assessed the impact of assuming different fiducial cosmologies, varying in particular the value of $\Omega_{\rm m}$. The results of this test are reported in the right panel of Figure 20. We find constraints consistent within 68% of uncertainty in all the analyzed cases, i.e., in a range of ± 0.04 from our fiducial value of $\Omega_{\rm m}$. Moreover, we tested different void radius and redshift binning choices, finding consistent results despite the number of cleaned voids being low, thus affected by statistical noise. Finally, we performed consistency checks on the impact of galaxy effective bias modeling and on the bias relation in voids from Equation (38), finding a negligible impact on the posterior distribution of $\Omega_{\rm m}$.

VGCF Scale Cut Validation. In contrast to most standard probes of large-scale structure, the VGCF analysis is not restricted to the largest scales. In fact, linear theory accurately describes the VGCF on all scales, even arbitrarily close to the void center. This is because the dynamics inside voids remain linear, in accordance with Equation (40), which has been tested in simulations for a wide range of void sizes down to only a few megaparsecs in effective radius (N. Schuster et al. 2023).

However, both the void identification method and estimators for the VGCF are affected by sparse sampling of tracers (P. M. Sutter et al. 2014a; M. C. Cousinou et al. 2019; N. Schuster et al. 2023). A characteristic scale of a galaxy distribution is its mean galaxy separation (mgs), below which discreteness effects from sparse sampling become important. For example, in this regime random Poisson noise can create or disrupt void detections, and counts-in-shell estimators for the VGCF can return biased results owing to low or no particle-count statistics. We therefore restrict our VGCF analysis to voids whose effective radius is larger than a multiple of the mgs. This can be implemented as a redshift-dependent cut with

$$R > N_{\rm mgs}\left(\frac{4\pi}{3}\bar{n}(z)\right)^{-1/3}, \tag{46}$$

where $N_{\rm mgs}$ is a tuning parameter and $\bar{n}(z)$ is the mean mock galaxy density as a function of redshift. The choice of $N_{\rm mgs}$ is a trade-off between maximizing the void sample size (lower $N_{\rm mgs}$) and minimizing the number of spurious voids (higher $N_{\rm mgs}$). In addition, we restrict our sample to the largest 50% of all voids passing this cut. We repeated our analysis for a range of values with $N_{\rm mgs} \in [1, 5]$ and find consistent results with decreasing uncertainty in our posterior constraints down to a value of $N_{\rm mgs} = 3$. Below that, the posteriors start shifting and stop shrinking (see the left panel of Figure 21), indicating the onset of stochastic bias or noise in the void sample. However, we select $N_{\rm mgs} = 5$ as our more conservative default, which results in a minimum void radius of $R \simeq 50.6\ h^{-1}$ Mpc and corresponds to the data vector shown in Figure 19.

Furthermore, we have the option to restrict or subdivide the provided redshift range. The choice of redshift binning is a compromise between detecting the redshift evolution of the VGCF, on the one hand, and maintaining sufficient statistics for its estimation, on the other. For example, in a *w*CDM cosmology, measuring the AP parameter in multiple redshift bins is necessary to break the degeneracy between w and $\Omega_{\rm m}$ entering $D_A(z)H(z)$. However, in ΛCDM a single redshift bin is sufficient to determine $\Omega_{\rm m}$. Hence, for the ΛCDM light-cone mock we use the entire redshift range available without binning to estimate the VGCF with reduced statistical noise.

VGCF Consistency Checks. A strategy to gain confidence in our analysis is to investigate its dependence on the assumed fiducial cosmology. We began using a flat ΛCDM cosmology with Planck2015 (Planck Collaboration et al. 2016) parameters, both in the void finder and for the construction of the VGCF data vector. The best-fit cosmology obtained from our posterior is then used to update our fiducial cosmology, and this procedure is repeated until convergence. In practice, only one such iteration was necessary in updating our void catalog to obtain a stable result within our error margins. The right panel of Figure 21 shows the impact of varying our fiducial value of

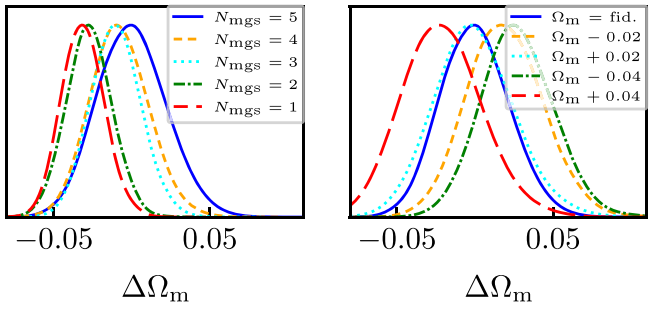

Figure 21. Posterior distributions for the value of $\Omega_{\rm m}$ based on the AP test with the VGCF in the ΛCDM light cone. The mean of the first distribution (solid blue line in each panel) has been subtracted to mask the results. The left panel shows the influence of different cuts on void effective radius in units of the mgs, while the right panel illustrates the impact of variations in the assumed fiducial values for $\Omega_{\rm m}$ in our analysis pipeline.

$\Omega_{\rm m}$ when constructing the VGCF data vector and running our analysis pipeline. It is evident that the posterior distribution moves diametrically to changes in the fiducial value of $\Omega_{\rm m}$. This allows us to identify a fiducial value that agrees most closely with its corresponding posterior mean. We select this posterior as our final result.

Unmasking Criteria. We calculate the reduced χ^2 statistic for the best-fit model of the VGCF with parameters $\mathbf{\Omega}^*$ as

$$\chi^2_{\rm red} = -\frac{2}{N_{\rm dof}} \log \mathcal{L}(\widehat{\boldsymbol{d}}_{\rm VGCF}|\mathbf{\Omega}^*), \tag{47}$$

with $N_{\rm dof} = N_{\rm bin} - N_{\rm par}$ degrees of freedom, where $N_{\rm bin}$ is the number of bins in the data and $N_{\rm par}$ is the number of model parameters. A value close to unity indicates a satisfactory model fit but does not guarantee unbiased parameter constraints. However, combined with the iteration strategy over fiducial cosmologies described above, it can be used to indicate convergence to an optimal result.

Caveats. This section describes failure modes for our analysis that could impact the final result for one or both of the considered void statistics. These can entail small shifts for the posterior, or complete failure of predicted values.

1. Void catalogs can be contaminated by spurious voids owing to sparse sampling (e.g., M. C. Neyrinck 2008; P. M. Sutter et al. 2014a). Recent work has shown that small voids are more likely to be spurious (A. Pisani et al. 2015b; M. C. Cousinou et al. 2019; N. Schuster et al. 2023); therefore, our analysis cuts based on void size should reduce the probability to include spurious voids in both the VSF and VGCF measurements. An additional purpose of this cut in the VSF analysis is to exclude spatial scales that are affected by incompleteness, i.e., loss of void counts (A. Pisani et al. 2015a). This occurs owing to the sparsity of tracers and is currently not modeled by the VSF theory. The posteriors of both analyses mildly depend on the choice of these cuts; we therefore caution that they can have an impact on our cosmological constraints. This additional source of error is so far neither quantified nor included as a systematic error in our analysis.
2. Despite the geometric distortion correction applied, our final results exhibit a mild dependence on the fiducial cosmology assumed. This indicates residual geometric effects that may play a role, but currently this dependence is not quantified as a potential systematic error.
3. Based on previous studies (S. Contarini et al. 2023; D. Pelliciari et al. 2023), the VSF analysis assumes the off-diagonal terms of the covariance matrix to be null. However, we acknowledge the possibility that this assumption could have a minor influence on our results. In addition, the assumed Poisson errors associated with the VSF are generally smaller than the true uncertainty related to this measure, so the errors on cosmological parameters may be underestimated. The jackknife approach to determine the covariance of the VGCF can be limited by the finite sample size of voids, which typically leads to an overestimation of uncertainties (N. Hamaus et al. 2022). Moreover, a supersample covariance is not included in our analysis, but we expect it to have a negligible impact on the final cosmological constraints (for the VSF, see A. E. Bayer et al. 2023a).
4. Possible systematic errors related to the void-finding and cleaning algorithms can also have an impact. A void sample identified via sparse tracers in redshift space is subject to various selection effects (e.g., C. M. Correa et al. 2021, 2022). While we can limit the impact of sparse sampling via Equation (46), residual selection biases may remain present in our sample. A full exploration of their correlation with various void properties and their mitigation is beyond the scope of this work, however.
5. The impact of HOD models on void statistics has barely been investigated in the literature so far. Because these models are usually tuned to reproduce the 2pt statistics of galaxies, their performance on void statistics is largely unknown and potentially less accurate. For example, a particular choice of HOD model may produce a strong FoG effect, which results in elongated features along the LOS. We do observe some elongation beyond the void boundaries in the VGCF shown in Figure 19. While this may be statistically insignificant, FoGs could impact the inner shape of the VGCF, resulting in overly flattened contours. In turn, this would bias the AP parameter. Furthermore, the HOD scheme may affect the VSF modeling via the bias-in-void parameterization in Equation (38).
6. The deprojection based on Equation (42) applied to the VGCF is exact for noiseless data but can amplify statistical fluctuations in noisy data by large amounts (A. Pisani et al. 2014; N. Hamaus et al. 2020). The survey characteristics, like tracer density, redshift range, and sky coverage provided in the challenge mocks, are more restricted than those available in BOSS, or expected by Euclid, for example. This limits the total number of bins (e.g., in redshift and effective radius) we

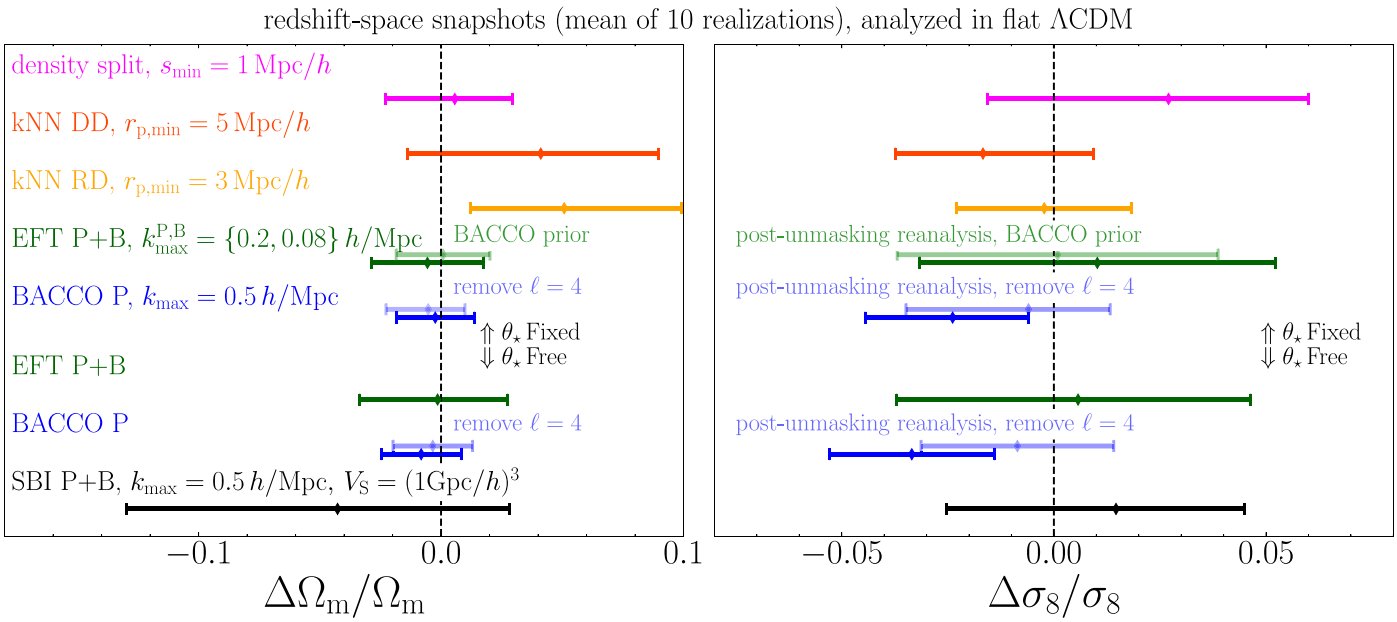

Figure 22. 1D marginalized constraints on $\Omega_{\rm m}$ and σ_8 for analyses of redshift-space mocks, including post-unmasking reanalyses shown with light opacity.

can divide our void sample by to a single one. We therefore expect stronger systematic errors due to deprojection effects in this analysis.

Finally, we emphasize that the current analysis of the presented void statistics does not make use of any external simulation suites that can be used for (nuisance) parameter calibration, template fitting, and the training of emulators. Such tools are helpful to significantly reduce parameter biases and uncertainties but require a substantial amount of preprocessing and computing resources. We leave such more extensive approaches to future work.

Post-unmasking Studies. After unmasking, we realized that the informative prior on h, $n_{\rm s}$, and $\omega_{\rm b}$ adopted by the baseline VSF analysis excluded the true cosmology of the light-cone mock. In subsequent analyses, we repeated the VSF analysis with (a) the informative cosmology prior recentered to include the true cosmology and (b) the broad cosmology prior of the BACCO analysis (Equation (7)). The central value of $\Omega_{\rm m}$ is robust to both prior variations, and the wide prior broadens the marginalized 1σ uncertainty on $\Omega_{\rm m}$ from 12.7% to 15.6%.

Lessons Learned. One of the most relevant lessons learned in this challenge concerns the importance of gaining confidence for the treatment of small voids. In particular, performing a masked analysis reveals that when we are uncertain of the outcome, it is tempting to be overly conservative with scale cuts. Our first tests based on less conservative cuts resulted in best-fit cosmologies with a significantly different mean galaxy number density as compared to the one provided by the challenge. Although we were aware that this mismatch could have had multiple origins apart from the background cosmology (e.g., the HOD prescription), we still opted for very cautious scale cuts owing to this observation.

In hindsight, however, it is reassuring to realize that the initial results based on larger void samples that extend down to smaller effective radii actually lie closer to the truth with smaller error bars (see the left panels of Figures 20 and 21). This suggests not only that we could have been more ambitious with including smaller voids in our analyses but also that our level of confidence in them was too low for doing so. For the future, it demonstrates the importance for increasing our understanding with regard to the systematics of small voids, so that their constraining power can be fully exploited.

6. Discussion

After unmasking, several teams further investigate the accuracy and precision of their original submission. Several of the results of these post-unmasking reanalyses are shown in Figures 22–23. The discussions in this section underline the importance of accounting for systematics due to model and analysis choices.

6.1. Cosmology Priors

We show that two post-unmasking analyses illustrate the impact of informative cosmology priors on parameter constraints. For the original unmasking submission, all analysis teams chose their own priors (which all differ from the priors from which the mock cosmologies are drawn), as illustrated in Figure 5. While the posteriors on the target parameters ($\Omega_{\rm m}$, σ_8) are substantially narrower than the priors in these parameters for all analyses, several analyses adopted informative priors for some of the other ΛCDM parameters. The light-green symbols in Figure 22 show the impact of imposing the BACCO cosmology priors (Equation (7)) on the EFT P+B analysis, which originally used the widest cosmology priors among all participating analysis teams (Equation (10)). The approximately 15% improvement in 1D marginalized constraints is primarily caused by the informative prior on $n_{\rm s}$. On the other hand, the VSF analysis chose informative priors on h, $\omega_{\rm b}$, and $n_{\rm s}$ that excluded the light-cone mock cosmology at several σ. Post-unmasking analyses indicate only a minor bias in $\Omega_{\rm m}$ due to the miscentered prior and a 23% degradation in constraining power with the wider cosmology prior of the BACCO analysis. These examples highlight the importance of homogenizing cosmology priors for future detailed comparisons of different analysis methods.

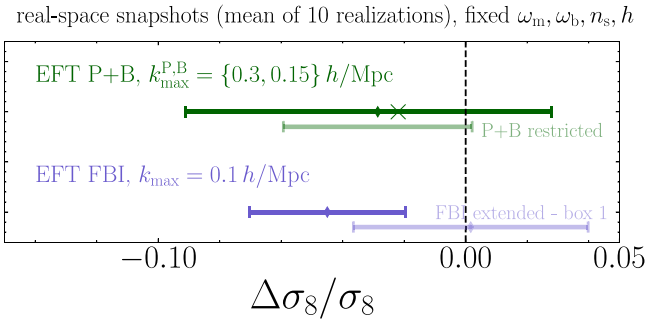

Figure 23. 1D marginalized constraints on σ_8 for analyses of real-space mocks, including post-unmasking reanalyses shown with light opacity. The "P+B restricted" posterior mean is recentered to the "EFT P+B" posterior mean in the EFT P+B baseline analysis. The "FBI extended" constraint is obtained from only one single realization (see *Post-unmasking Studies* in Section 5.3).

6.2. Post-unmasking BACCO Analysis Refinement

The BACCO team identified a previously unnoticed large emulation uncertainty of the hexadecapole that became significant owing to the increased simulation volume of the challenge mocks compared to their previous validation (see Section 5.1.3). The light-blue symbols show a post-unmasking reanalysis of monopole and quadrupole only, with all other analysis settings held fixed.

6.3. Post-unmasking EFT Analysis Comparisons

The EFT P+B and EFT FBI teams adopted different modeling choices in their pre-unmasking analyses, which are partly motivated by the complexities of numerical implementation and computation. To investigate the impact of these differences on their respective constraints, after unmasking, both teams have run new analyses with assumptions that aim to draw closer to the baseline analysis of the other team. In Figure 23, we show results of these post-unmasking studies (light opacity), next to the original pre-unmasking results (full opacity).

To first understand the difference between the two EFT pre-unmasking analyses, note that the EFT FBI pre-unmasking analysis described the matter–galaxy connection with a second-order galaxy bias expansion and assumed that the stochastic contribution to the galaxy density field, ε, is Gaussian distributed with a white power spectrum $\sigma_{\mathrm{Poisson}}$, effectively neglecting the coupling between the stochastic and deterministic fields of the form $\delta_g \supset \delta_m\varepsilon$ (e.g., V. Desjacques et al. 2018). In contrast, the EFT P+B baseline analysis included (1) all second-order bias terms plus the most relevant third-order bias term, (2) the most general galaxy stochasticity model at the given order including stochastic bispectrum and stochastic-deterministic cross bispectrum contributions (stemming from the $\delta_m\varepsilon$ term in the bias expansion), and (3) $\mathcal{O}(k^2)$ corrections to the stochasticity power spectrum.

To see how these differences between the EFT FBI and EFT P+B analyses affect their result, the EFT P+B team has run a "P+B restricted" analysis on the real-space mocks, at the same (pre-unmasking) scale cuts $k_{\mathrm{max}} = \{0.3, 0.15\}\ h\ \mathrm{Mpc}^{-1}$, and fixing all cosmological parameters to their true values—except σ_8—and with $A_{\mathrm{shot}} = B_{\mathrm{shot}} = b_{\Gamma_3} = a_0 = R_*^2 = 0$ (the higher derivative bias defined as in A. Chudaykin et al. 2021b). The first important observation is that EFT P+B restricted analysis recovered σ_8 with a significant bias. This is because all components described above, such as non-Gaussian stochasticity in the bispectrum, third-order operators in the galaxy bias expansion, and stochastic–deterministic coupling, are nonnegligible at the scales involved in the EFT P+B analyses.[69] In order to ease the comparison in the presence of this bias, the "P+B restricted" posterior mean in Figure 23 is recentered to the "EFT P+B" posterior mean in the EFT P+B baseline analysis. Thus, the "P+B restricted" results in Figure 23 should only be considered in terms of the error bars. A direct comparison to the "EFT FBI" error bars shows that the nominal variances on σ_8 in both analyses are relatively similar, suggesting that the difference between these two constraints likely comes from different analysis assumptions.

Conversely, in their post-unmasking study, the EFT FBI team has run an "FBI extended" analysis that includes the full set of third-order bias operators in the galaxy bias expansion (i.e., three additional bias terms compared to the "EFT P+B" pre-unmasking analysis), as well as higher-derivative stochasticity (but still without mode coupling in the stochastic part), at their pre-unmasking scale cut $k_{\mathrm{max}} = 0.1\ h\ \mathrm{Mpc}^{-1}$. The "FBI extended" posterior in Figure 23 is completely consistent with the ground truth σ_8, with a noticeably increased width (in between those from the "EFT P+B" and "P+B restricted" analyses, pre- and post-unmasking).

The upshot of these post-unmasking analyses and comparisons by the two EFT teams is that we find a broad consistency between the EFT P+B and FBI results, but future work is required to extend this comparison to the full cosmology parameter space and to redshift-space clustering.

6.4. Toward Correct Cosmology Constraints from the Highly Nonlinear Regime

While there is undoubtedly significant signal-to-noise ratio in galaxy clustering in the highly nonlinear regime, the conversion from signal-to-noise ratio to cosmology parameter constraints crucially relies on parameterizations of the galaxy–halo connection in the nonlinear regime being sufficiently flexible to marginalize over all modeling uncertainties that could bias cosmological parameter estimation. Furthermore, models for clustering statistics in the highly nonlinear regime are typically evaluated using simulations and emulators. This type of model evaluation is subject to uncertainties from cosmic variance, emulation errors, and finite training sample size, which may further reduce constraining power.

Four participating analyses utilize scales within the highly nonlinear regime, of which three rely on HOD-based models. Here we summarize pre-unmasking validation tests and post-unmasking reanalyses that illustrate the impact of highly nonlinear scales and discuss limitations to error quantification in the highly nonlinear regime.

The BACCO P analyses with $k_{\mathrm{max}} = 0.2\ h\ \mathrm{Mpc}^{-1}$ and $k_{\mathrm{max}} = 0.5\ h\ \mathrm{Mpc}^{-1}$ (Figure 7) indicate a substantial gain in constraining power from strongly nonlinear scales. The BACCO analysis relies on a "hybrid" Lagrangian galaxy bias expansion, which has been validated against HOD and SHAMe (S. Contreras et al. 2021a, 2021b) techniques.

The SBI P+B analysis recovers the input cosmology to better than 1σ with a scale cut $k_{\mathrm{max}} = 0.5\ h\ \mathrm{Mpc}^{-1}$. While the

[69] It is possible that these terms are less important on large scales, e.g., at $k \sim 0.1\ h\ \mathrm{Mpc}^{-1}$ corresponding to k_{max} in the EFT FBI analyses.

SBI P+B team did not perform scale cut variations for the challenge mock catalogs, other analyses using the same model for galaxy samples with similar number density show substantial gains in constraining power from scales $0.25\ h\,\mathrm{Mpc}^{-1} < k \leqslant 0.5\ h\,\mathrm{Mpc}^{-1}$ for the power spectrum (C. Hahn et al. 2023a) and $0.3\ h\,\mathrm{Mpc}^{-1} < k \leqslant 0.5\ h\,\mathrm{Mpc}^{-1}$ for the bispectrum (C. Hahn et al. 2024).

The DSC analysis includes measurements down to $1\ h^{-1}$ Mpc. However, the gain in constraining power from 30 to $1\ h^{-1}$ Mpc scales is limited (see Figure 17), as the density-split correlation functions are not sensitive to variations in the clustering below the smoothing kernel scale ($R_s = 10\ h^{-1}$ Mpc) and the small-scale measurement only contains information about the smoothed density PDF. Additionally, the emulator and training set errors contribute significantly to the total covariance on small scales, further suppressing the information gain. Finally, previous studies (E. Paillas et al. 2023) showed that density-split statistics can efficiently extract information from AP distortions, which were not included in the current redshift-space mocks.

The DD-kNN baseline analysis employs the scale cut $r_p > 5\ h^{-1}$Mpc. Post-unmasking reanalyses with even more aggressive scale cuts shown in Figure 13 do not pass robustness tests, indicating insufficient model flexibility or uncertainty modeling. Hence, further model refinements are required to quantify the cosmological constraining power of aggressive analyses in the one-halo regime.

6.4.1. Galaxy–Halo Connection Models

We reiterate that the organizers communicated to all participants that the challenge mock catalogs are based on the HOD formalism, and it is important to acknowledge that HODs, like any other model for the galaxy–halo connection, come with their own set of limitations due to assumptions about galaxy formation. While the specific HOD parameterization is not revealed, this information gives HOD-based modeling approaches an inherent advantage. Among the participating analyses, the modeling of DSC, kNN statistics, and SBI P+B are explicitly based on HODs. We refer to C. Cuesta-Lazaro et al. (2024) for parameter recovery validation of the HOD-based density-split statistics model on a SHAM-based galaxy mock and to C. Hahn et al. (2023b) for parameter recovery validation of the HOD-based SBI P+B model against different HOD models and different N-body halo catalogs. The kNN post-unmasking reanalyses with different HOD models (Figure 13) show that kNN-DD analyses with aggressive one-halo scale cuts are currently subject to either error miscalibration or model misspecification, even within the HOD framework. Future research on more flexible parameterizations will be required to pass comprehensive parameter recovery tests (see R. M. Reddick et al. 2014, for an early example of comprehensive non-HOD recovery tests) and to obtain well-calibrated error bars. Such increased modeling flexibility will likely be accompanied by a degradation in constraining power, especially when accounting for non-HOD galaxy–halo connection models (see R. H. Wechsler & J. L. Tinker 2018, for a review). Hence, the advance of novel analysis methods that exploit the highly nonlinear regime requires developing and validating more realistic mocks to establish a more accurate challenge framework that can encapsulate a broader set of models.

Validation of the uncertainties reported by individual methods will ultimately require a suite of (ideally, parameter-masked) mock catalogs with variations in cosmology and galaxy–halo connection model. However, not just any suite will do, and the design of such a suite is essential for enabling Bayesian quantification of model performance and uncertainty calibration. Ideally, such a suite would be generated by drawing from posteriors of galaxy–halo connection model(s) that satisfy the following:

1. The mock-generating model(s) should be consistent with observations within some measurement uncertainty.
2. The mock-generating model(s) should be predicated upon different assumptions from the models in the analysis being tested.

For this program to be successful, realism matching simulations to observations is essential. This is a challenging task because we do not yet have a suite of simulations available that fully matches all observational data. Still, paying attention to developing accurate and flexible mock sky surveys that have the right set of included systematics should be a high priority so that the field is not misled by either missing important systematics or trying to account for unphysical systematics. Furthermore, any Bayesian quantification of model sufficiency and uncertainty validation critically depends on the (hyper) parameter priors for the test suite, which should be determined in data space rather than through arbitrary parameterization choices. This development will benefit from continued improvements in cosmological hydrodynamical simulations (for a review, see, e.g., R. A. Crain & F. van de Voort 2023), as well as from the increasing diversity of semiempirical methods for generating realistic mock galaxy populations (e.g., P. Behroozi et al. 2019; A. P. Hearin et al. 2021, 2023; R. H. Wechsler et al. 2022; K. J. Kwon et al. 2023).

7. Conclusion

It is well established that the galaxy density field contains valuable information beyond the power spectrum, and many "novel" statistics and analysis methods have emerged and matured over the past decade to exploit this information. This makes it timely to survey the state of these analyses. In this paper, we present a parameter-masked mock challenge for beyond-2pt galaxy clustering statistics. The challenge data set consists of mock catalogs created from N-body simulations with a flat ΛCDM cosmology and HOD galaxy–halo connection models, with parameter values known only to the organizers. While all parameter values are masked, analysis teams optimize scale cuts and determine other analysis choices (e.g., nuisance parameters and their priors) for each analysis method and submit one result per method for unmasking. Upon unmasking, the organizers share plots of the relative parameter biases and uncertainties of the target parameters (Ω_m, σ_8) but do not share information on other cosmological parameters, HOD parameterization, or HOD parameters. Post-unmasking analyses are encouraged to enable continued method development but need to be clearly labeled as such. The main results of this mock challenge can be summarized in three themes:

1. *Design of pre-unmasking analysis strategies and validation studies.* The priors for the mock catalogs (e.g., $\sigma_8 \sim \mathcal{U}[0.68, 0.9]$ and any HOD parameters; see Section 3) are significantly broader than implicit priors

in (nonmasked) data analyses from previous observations. The inability to iterate on Section 2 provided an incentive for the analysis teams to develop consistency checks and unmasking criteria, summarized in Section 5, even if this robustness comes at some cost in constraining power. These validation studies are an essential ingredient for future analyses to meet the accuracy requirements of next-generation data sets.

2. *Constraints from parameter-masked mock challenge.* The unmasking results presented in Section 2 showcase the competitive constraining power of multiple beyond-2pt statistics and novel analysis methods in a parameter-masked mock challenge. This performance of multiple statistics, as well as the associated modeling and inference frameworks, lends credibility to obtaining accurate and precise cosmology constraints via these methods. Further, the consistency across different analysis approaches enables a level of cross-validation on real data that is impossible to achieve for a single method on its own. The combination of multiple statistics and modeling, in principle, will enable the most precise constraints (e.g., P+B; see A. Banerjee & T. Abel 2021; C. D. Kreisch et al. 2022; A. E. Bayer et al. 2023a; J. Hou et al. 2023; E. Massara et al. 2023; K. Storey-Fisher et al. 2024; G. Valogiannis et al. 2024, for other combinations of beyond-2pt and 2pt statistics). Accurate joint covariances and consistent models of the galaxy–halo connection across different methods will, however, require further research and pipeline developments.
3. *Post-unmasking method refinements.* The unmasking results assess the performance of an analysis based on previous method development and masked validation tests. After unmasking, several teams further investigate the accuracy and precision of their original submission (see Section 6). With caution against excessive fine-tuning, post-unmasking reanalyses identify directions for model refinements and methodological improvements for future analyses. Additionally, post-unmasking comparisons between different analysis methods provide a starting point for future studies contrasting different approaches to understanding the source of cosmological information beyond the conventional, linear, and quasi-linear galaxy 2pt analyses.

While most of these findings can in principle be obtained by each team on individually generated mock catalogs, an externally organized parameter-masked challenge provides a common benchmark and naturally separates the validation of analysis choices and post-unmasking refinements, enabling a clearer assessment of constraining power. Hence, we invite future submissions from other analysis teams and offer to update summary results plots in the Beyond-2pt challenge repository with new submissions ✎.

From a participant's perspective, the setup of the challenge and unmasking procedure, as well as the exchanges with the organizers and other teams, allowed a beneficial constructive atmosphere and strongly encouraged scientific interactions. Overall, the positive work environment throughout the challenge encouraged the collaborative development of validation tests and contributed to a better understanding of the tools used by the various teams.

A single parameter-masked mock challenge offers no panacea: models and analysis methods for different statistics evolve (e.g., due to post-unmasking refinements motivated by this study), and scale cuts and priors must be calibrated anew for different survey volumes and galaxy samples. Furthermore, this particular mock challenge was clearly labeled as consisting of HOD-based galaxies free of observational systematics. Therefore, future parameter-masked mock challenges with more realistic astrophysical and observational complications, as well as with statistical precision mirroring the increasing survey volume and galaxy density of future surveys, will be required to further method validation. While this challenge provides a vital stress test and performance benchmark at one specific point in cosmology and HOD parameter space, the current setup leaves the uncertainty calibration and validation against model misspecification to the individual analysis teams, using individually generated mock catalogs with parameter and model variations. As discussed in Section 6.4, developing a suite of mocks suitable for the validation of parameter constraints, including their uncertainties, from highly nonlinear scales at the accuracy of Stage IV surveys will be a major research and computing project that is beyond the resources of individual groups. Hence, it would be desirable for future community-wide challenges to provide suites of mock catalogs across parameter space and galaxy–halo connection models to facilitate a rigorous assessment of uncertainty quantification.

To conclude, we emphasize the maturity of multiple "novel" statistics and analysis methods that participated in this parameter-masked mock challenge. The individual constraining power of a particular statistic depends on the specific parameter space considered, and we caution against extrapolating relative constraining power in this challenge to other scenarios. The main strength of this emerging field is the complementarity of approaches, which will enable extensive cross-validations to yield reliable and competitive results.

Acknowledgments

This mock challenge was initiated during the Aspen Center for Physics 2022 Summer Program "Large-Scale Structure Cosmology beyond 2-Point Statistics," co-organized by D.J., E.K., Hiranya Peiris, and F.S. We are grateful to Hiranya Peiris for co-organizing this workshop and input on the initial design of this challenge and to the Aspen Center for Physics, supported by the National Science Foundation grant PHY-1607611. We thank Steward Observatory, University of Arizona for hosting a second workshop for participating analysis teams in spring 2023, with support from the David and Lucile Packard Foundation.

The light-cone mock galaxy catalog is based on the `AbacusSummit` simulation light cones, and we thank the `AbacusSummit` team for making their data products publicly available. The *N*-body simulations and subsequent halo catalog creation for producing the ΛCDM snapshot mocks were carried out on Cray XC50 at the Center for Computational Astrophysics, National Astronomical Observatory of Japan. We further acknowledge High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and RDI and maintained by the UA Research Technologies department.

We are grateful to Boryana Hadzhiyska, Andrew Hearin, Johannes Lange, Ariel Sánchez, and Risa Wechsler for their valuable comments on the manuscript. We further thank Camille Avestruz, Humna Awan, Andrew Hearin, Dragan Huterer, Nick Kokron, Martin Reinecke, Marko Simonović

Julia Stadler, Masahiro Takada, Kuan Wang, Risa Wechsler, and Martin White for helpful discussions.

E.K., Y.K., and A.N.S. were supported in part by the David and Lucile Packard Foundation and a research fellowship from the Alfred P. Sloan foundation. C.H. was supported by the AI Accelerator program of the Schmidt Futures Foundation. N.-M. N. acknowledges support from the Leinweber Foundation. O.H.E.P. is a Junior Fellow of the Simons Society of Fellows. The work of T.A. and S.Y. was supported by the US Department of Energy SLAC contract DE-AC02-76SF00515. K.A. acknowledges the support from Fostering Joint International Research (B) under contract No. 21KK0050. M.P.I. is supported by STFC consolidated grant No. RA5496. A.P. acknowledges support from the Simons Foundation to the Center for Computational Astrophysics at the Flatiron Institute, as well as support from the European Research Council (ERC) under the European Union's Horizon program (COSMOBEST ERC funded project, grant agreement 101078174). C.C.-L. is supported by the National Science Foundation under Cooperative Agreement PHY2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions).

The EFT FBI analyses were conducted on the COBRA and FREYA HPC clusters at the Max Planck Computing and Data Facility.

Author Contributions

All authors contributed to the interpretation of results and reviewed the manuscript. Author contributions to individual analyses and challenge organization are listed below.

Elisabeth Krause—*Challenge*: challenge conceptualization and organization, workshop organization, design of mock catalogs, coordination with analysis teams. Pre-unmasking review of analysis sections, general writing and end-to-end editing.

Yosuke Kobayashi—*Challenge*: design of mock catalogs, *N*-body simulations, creation of mock galaxy catalogs, communication with analysis teams.

Andrés N. Salcedo—*Challenge*: design of mock catalogs, creation of mock galaxy catalogs, communication with analysis teams, writing and editing of general sections.

Mikhail Ivanov—*Challenge*: customized EFT P+B analysis runs for comparisons with other methods, writing and editing of general sections. EFT P+B *analysis*: design of validation tests, analysis runs, interpretation of results, and writing of analysis section.

Tom Abel—*k*NN *analysis*: design of validation tests, interpretation of results, post-unmasking studies.

Kazuyuki Akitsu—EFT P+B *analysis*: analysis runs, interpretation of results, and writing of analysis section.

Raul Angulo—BACCO P *analysis*: design of validation tests, interpretation of results, post-unmasking studies.

Giovanni Cabass—EFT P+B *analysis*: design of validation tests and interpretation of results.

Sofia Contarini—Void *analysis*: development of analysis pipeline, design of validation tests, analysis runs, interpretation of results, and writing of analysis section.

Carolina Cuesta-Lazaro—Density-split *analysis*: design of validation tests, analysis runs, interpretation of results, and writing of analysis section.

ChangHoon Hahn—SBI P+B *analysis*: design of validation tests, analysis runs, interpretation of results, and writing of analysis section.

Nico Hamaus—Void *analysis*: development of analysis pipeline, design of validation tests, analysis runs, interpretation of results, writing of analysis section, and analysis team coordination.

Donghui Jeong—*Challenge*: challenge conceptualization and workshop organization.

Chirag Modi—SBI P+B *analysis*: design of validation tests, analysis runs, interpretation of results, and writing of analysis section.

Nhat-Minh Nguyen—*Challenge*: General writing and end-to-end editing. EFT FBI *analysis*: development of analysis pipeline, design of validation tests, analysis runs, interpretation of results, post-unmasking studies, and writing of analysis section.

Takahiro Nishimishi—*Challenge*: development of *N*-body simulation code GINKAKU, optimization of accuracy parameters for the use in the challenge, writing of simulation section.

Enrique Paillas—Density-split *analysis*: design of validation tests, analysis runs, interpretation of results, and writing of analysis section.

Marcos Pellejero Ibañez—BACCO P *analysis*: analysis team coordination, design of validation tests, analysis runs, interpretation of results, post-unmasking studies, and writing of analysis section.

Oliver H. E. Philcox—EFT P+B *analysis*: design of validation tests, analysis runs, interpretation of results, and writing of analysis section.

Alice Pisani—Void *analysis*: design of validation tests, analysis interpretation, interpretation of results, writing of analysis section, and analysis team coordination.

Fabian Schmidt—*Challenge*: challenge conceptualization and coordination, workshop organization. EFT FBI *analysis*: development of analysis pipeline, design of validation test, interpretation of results, post-unmasking studies.

Satoshi Tanaka—*Challenge*: development and tuning of *N*-body simulation code GINKAKU.

Giovanni Verza—Void *analysis*: design of validation tests, analysis runs, interpretation of results, and writing of analysis section.

Sihan Yuan—*Challenge*: writing of HOD overview. *k*NN *analysis*: design of validation tests, analysis runs, interpretation of results, and writing of analysis section.

Matteo Zennaro—BACCO P *analysis*: design of validation tests, interpretation of results, post-unmasking studies.

ORCID iDs

Elisabeth Krause https://orcid.org/0000-0001-8356-2014
Yosuke Kobayashi https://orcid.org/0000-0002-6633-5036
Andrés N. Salcedo https://orcid.org/0000-0003-1420-527X
Mikhail M. Ivanov https://orcid.org/0000-0002-6745-984X
Tom Abel https://orcid.org/0000-0002-5969-1251
Kazuyuki Akitsu https://orcid.org/0000-0001-6473-3420
Raul E. Angulo https://orcid.org/0000-0003-2953-3970
Giovanni Cabass https://orcid.org/0000-0001-9487-702X
Sofia Contarini https://orcid.org/0000-0002-9843-723X
Carolina Cuesta-Lazaro https://orcid.org/0000-0002-6069-2999
ChangHoon Hahn https://orcid.org/0000-0003-1197-0902
Nico Hamaus https://orcid.org/0000-0002-0876-2101
Donghui Jeong https://orcid.org/0000-0002-8434-979X
Chirag Modi https://orcid.org/0000-0002-1670-2248
Nhat-Minh Nguyen https://orcid.org/0000-0002-2542-7233

Takahiro Nishimichi https://orcid.org/0000-0002-9664-0760
Enrique Paillas https://orcid.org/0000-0002-4637-2868
Marcos Pellejero Ibañez https://orcid.org/0000-0003-4680-7275
Oliver H. E. Philcox https://orcid.org/0000-0002-3033-9932
Alice Pisani https://orcid.org/0000-0002-6146-4437
Fabian Schmidt https://orcid.org/0000-0002-6807-7464
Satoshi Tanaka https://orcid.org/0000-0003-2442-8784
Giovanni Verza https://orcid.org/0000-0002-1886-8348
Sihan Yuan https://orcid.org/0000-0002-5992-7586
Matteo Zennaro https://orcid.org/0000-0002-4458-1754

References

Abidi, M. M., & Baldauf, T. 2018, JCAP, 07, 029
Aghamousa, A., Aguilar, J., Ahlen, S., et al. 2016, arXiv:1611.00036
Alam, S., Aubert, M., Avila, S., et al. 2021, PhRvD, 103, 083533
Alcock, C., & Paczynski, B. 1979, Natur, 281, 358
Alsing, J., Charnock, T., Feeney, S., & Wandelt, B. 2019, MNRAS, 488, 4440
Anbajagane, D., Aung, H., Evrard, A. E., et al. 2022, MNRAS, 510, 2980
Andrews, A., Jasche, J., Lavaux, G., & Schmidt, F. 2023, MNRAS, 520, 5746
Angulo, R. E., & White, S. D. M. 2010, MNRAS, 405, 143
Angulo, R. E., Zennaro, M., Contreras, S., et al. 2021, MNRAS, 507, 5869
Babić, I., Schmidt, F., & Tucci, B. 2024, arXiv:2407.01524
Balaguera-Antolínez, A., Kitaura, F.-S., Alam, S., et al. 2023, A&A, 673, A130
Baldauf, T., Mirbabayi, M., Simonović, M., & Zaldarriaga, M. 2015, PhRvD, 92, 043514
Baldauf, T., Mirbabayi, M., Simonović, M., & Zaldarriaga, M. 2016a, arXiv:1602.00674
Baldauf, T., Schaan, E., & Zaldarriaga, M. 2016b, JCAP, 03, 017
Banerjee, A., & Abel, T. 2021, MNRAS, 500, 5479
Baumann, D., Nicolis, A., Senatore, L., & Zaldarriaga, M. 2012, JCAP, 07, 051
Bayer, A. E., Liu, J., Terasawa, R., et al. 2023a, PhRvD, 108, 043521
Bayer, A. E., Seljak, U., & Modi, C. 2023b, arXiv:2307.09504
Bayer, A. E., Villaescusa-Navarro, F., Massara, E., et al. 2021, ApJ, 919, 24
Behroozi, P., Wechsler, R. H., Hearin, A. P., & Conroy, C. 2019, MNRAS, 488, 3143
Behroozi, P. S., Wechsler, R. H., & Wu, H.-Y. 2013, ApJ, 762, 109
Beltz-Mohrmann, G. D., Szewciw, A. O., Berlind, A. A., & Sinha, M. 2023, ApJ, 948, 100
Berlind, A. A., & Weinberg, D. H. 2002, ApJ, 575, 587
Blas, D., Garny, M., Ivanov, M. M., & Sibiryakov, S. 2016a, JCAP, 07, 028
Blas, D., Garny, M., Ivanov, M. M., & Sibiryakov, S. 2016b, JCAP, 07, 052
Blot, L., Crocce, M., Sefusatti, E., et al. 2019, MNRAS, 485, 2806
Bouchet, F. R., Colombi, S., Hivon, E., & Juszkiewicz, R. 1995, A&A, 296, 575
Brieden, S., Gil-Marín, H., & Verde, L. 2021, JCAP, 12, 054
Brieden, S., Gil-Marín, H., & Verde, L. 2022, JCAP, 06, 005
Brinckmann, T., & Lesgourgues, J. 2019, PDU, 24, 100260
Brooks, S. P., & Gelman, A. 1998, Journal of Computational and Graphical Statistics, 7, 434
Buchert, T. 1992, MNRAS, 254, 729
Bullock, J. S., Wechsler, R. H., & Somerville, R. S. 2002, MNRAS, 329, 246
Cabass, G. 2021, JCAP, 01, 067
Cabass, G., & Schmidt, F. 2020a, JCAP, 04, 042
Cabass, G., & Schmidt, F. 2020b, JCAP, 2020, 051
Cabass, G., Simonović, M., & Zaldarriaga, M. 2024, PhRvD, 109, 043526
Cai, Y.-C., Taylor, A., Peacock, J. A., & Padilla, N. 2016, MNRAS, 462, 2465
Cannon, P., Ward, D., & Schmon, S. M. 2022, arXiv:2209.01845
Carrasco, J. J. M., Hertzberg, M. P., & Senatore, L. 2012, JHEP, 09, 082
Chan, K. C., Hamaus, N., & Desjacques, V. 2014, PhRvD, 90, 103521
Chen, S.-F., Vlah, Z., Castorina, E., & White, M. 2021, JCAP, 03, 100
Chen, S.-F., Vlah, Z., & White, M. 2020, JCAP, 07, 062
Chen, S.-F., Vlah, Z., & White, M. 2022, JCAP, 02, 008
Cheng, S., Marques, G. A., Grandón, D., et al. 2025, JCAP, 2025, 006
Chudaykin, A., Dolgikh, K., & Ivanov, M. M. 2021a, PhRvD, 103, 023507
Chudaykin, A., & Ivanov, M. M. 2019, JCAP, 11, 034
Chudaykin, A., & Ivanov, M. M. 2023, PhRvD, 107, 043518
Chudaykin, A., Ivanov, M. M., Philcox, O. H. E., & Simonović, M. 2020, PhRvD, 102, 063533
Chudaykin, A., Ivanov, M. M., & Simonović, M. 2021b, PhRvD, 103, 043525
Contarini, S., Marulli, F., Moscardini, L., et al. 2021, MNRAS, 504, 5021
Contarini, S., Pisani, A., Hamaus, N., et al. 2023, ApJ, 953, 46
Contarini, S., Pisani, A., Hamaus, N., et al. 2024, A&A, 682, A20
Contarini, S., Ronconi, T., Marulli, F., et al. 2019, MNRAS, 488, 3526
Contarini, S., Verza, G., Pisani, A., et al. 2022, A&A, 667, A162
Contreras, S., Angulo, R. E., Springel, V., et al. 2023a, MNRAS, 524, 2489
Contreras, S., Angulo, R. E., & Zennaro, M. 2021a, MNRAS, 504, 5205
Contreras, S., Angulo, R. E., & Zennaro, M. 2021b, MNRAS, 508, 175
Contreras, S., Chaves-Montero, J., & Angulo, R. E. 2023b, MNRAS, 525, 3149
Correa, C. M., Paz, D. J., Padilla, N. D., et al. 2022, MNRAS, 509, 1871
Correa, C. M., Paz, D. J., Sánchez, A. G., et al. 2021, MNRAS, 500, 911
Cousinou, M. C., Pisani, A., Tilquin, A., et al. 2019, A&C, 27, 53
Crain, R. A., & van de Voort, F. 2023, ARA&A, 61, 473
Cranmer, K., Brehmer, J., & Louppe, G. 2020, PNAS, 117, 30055
Crocce, M., Pueblas, S., & Scoccimarro, R. 2006, MNRAS, 373, 369
Crocce, M., & Scoccimarro, R. 2008, PhRvD, 77, 023533
Cuesta-Lazaro, C., & Mishra-Sharma, S. 2024, PhRvD, 109, 123531
Cuesta-Lazaro, C., Paillas, E., Yuan, S., et al. 2024, MNRAS, 531, 3336
D'Amico, G., Donath, Y., Lewandowski, M., Senatore, L., & Zhang, P. 2024a, JCAP, 2024, 059
D'Amico, G., Donath, Y., Lewandowski, M., Senatore, L., & Zhang, P. 2024b, JCAP, 2024, 041
D'Amico, G., Gleyzes, J., Kokron, N., et al. 2020, JCAP, 05, 005
D'Amico, G., Senatore, L., Zhang, P., & Nishimichi, T. 2024c, JCAP, 2024, 037
Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10
DeRose, J., Kokron, N., Banerjee, A., et al. 2023, JCAP, 2023, 054
DeRose, J., Wechsler, R. H., Becker, M. R., et al. 2019, arXiv:1901.02401
DESI Collaboration, Adame, A. G., Aguilar, J., et al. 2025, JCAP, 2025, 021
Desjacques, V., Jeong, D., & Schmidt, F. 2018, PhR, 733, 1
Doeser, L., Jamieson, D., Stopyra, S., et al. 2024, MNRAS, 535, 1258
Dutton, A. A., & Macciò, A. V. 2014, MNRAS, 441, 3359
Elsner, F., Schmidt, F., Jasche, J., Lavaux, G., & Nguyen, N.-M. 2020, JCAP, 01, 029
Euclid Collaboration, Pezzotta, A., Moretti, C., et al. 2024, A&A, 687, A216
Fang, Y., Hamaus, N., Jain, B., et al. 2019, MNRAS, 490, 3573
Feldman, H. A., Frieman, J. A., Fry, J. N., & Scoccimarro, R. 2001, PhRvL, 86, 1434
Feroz, F., & Hobson, M. P. 2008, MNRAS, 384, 449
Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601
Feroz, F., Hobson, M. P., Cameron, E., & Pettitt, A. N. 2019, OJAp, 2, 10
Foreman-Mackey, D., Farr, W., Sinha, M., et al. 2019, JOSS, 4, 1864
Garrison, L. H., Eisenstein, D. J., & Pinto, P. A. 2019, MNRAS, 485, 3370
Gatti, M., Jain, B., Chang, C., et al. 2022, PhRvD, 106, 083509
Gaztanaga, E., Cabre, A., Castander, F., Crocce, M., & Fosalba, P. 2009, MNRAS, 399, 801
Gil-Marín, H., Percival, W. J., Verde, L., et al. 2017, MNRAS, 465, 1757
Gonzalez-Perez, V., Comparat, J., Norberg, P., et al. 2018, MNRAS, 474, 4024
Guo, H., Zheng, Z., Behroozi, P. S., et al. 2016, MNRAS, 459, 3040
Guo, H., Zheng, Z., Zehavi, I., et al. 2015, MNRAS, 446, 578
Hadzhiyska, B., Eisenstein, D., Bose, S., Garrison, L. H., & Maksimova, N. 2022, MNRAS, 509, 501
Hahn, C., 2020 pySpectrum: Power Spectrum and Bispectrum Calculator, Astrophysics Source Code Library, ascl:2009.014
Hahn, C., Beutler, F., Sinha, M., et al. 2019, MNRAS, 485, 2956
Hahn, C., Eickenberg, M., Ho, S., et al. 2023a, PNAS, 120, e2218810120
Hahn, C., Eickenberg, M., Ho, S., et al. 2024, PhRvD, 109, 083534
Hahn, C., Eickenberg, M., Ho, S., et al. 2023b, JCAP, 2023, 010
Hamaus, N., Aubert, M., Pisani, A., et al. 2022, A&A, 658, A20
Hamaus, N., Cousinou, M.-C., Pisani, A., et al. 2017, JCAP, 7, 014
Hamaus, N., Pisani, A., Choi, J.-A., et al. 2020, JCAP, 2020, 023
Hamaus, N., Pisani, A., Sutter, P. M., et al. 2016, PhRvL, 117, 091302
Hamaus, N., Sutter, P. M., Lavaux, G., & Wandelt, B. D. 2015, JCAP, 11, 036
Hamaus, N., Sutter, P. M., & Wandelt, B. D. 2014a, PhRvL, 112, 251302
Hamaus, N., Wandelt, B. D., Sutter, P. M., Lavaux, G., & Warren, M. S. 2014b, PhRvL, 112, 041304
Hand, N., Feng, Y., Beutler, F., et al. 2018, AJ, 156, 160
Harnois-Déraps, J., Martinet, N., Castro, T., et al. 2021, MNRAS, 506, 1623
Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399
Hawken, A. J., Granett, B. R., Iovino, A., et al. 2017, A&A, 607, A54
Hearin, A. P., Chaves-Montero, J., Alarcon, A., Becker, M. R., & Benson, A. 2023, MNRAS, 521, 1741
Hearin, A. P., Chaves-Montero, J., Becker, M. R., & Alarcon, A. 2021, OJAp, 4, 7
Hearin, A. P., Zentner, A. R., van den Bosch, F. C., Campbell, D., & Tollerud, E. 2016, MNRAS, 460, 2552
Heitmann, K., Finkel, H., Pope, A., et al. 2019, ApJS, 245, 16

Heydenreich, S., Brück, B., Burger, P., et al. 2022, A&A, 667, A125
Ho, M., et al. 2024, OJAp, 7, 54
Hou, J., Moradinezhad Dizgah, A., Hahn, C., & Massara, E. 2023, JCAP, 2023, 045
Hou, J., Moradinezhad Dizgah, A., Hahn, C., et al. 2024, PhRvD, 109, 103528
Ishiyama, T., Fukushige, T., & Makino, J. 2009, PASJ, 61, 1319
Ishiyama, T., Nitadori, K., & Makino, J. 2012, arXiv:1211.4406
Ishiyama, T., Prada, F., Klypin, A. A., et al. 2021, MNRAS, 506, 4210
Ivanov, M. M. 2021, PhRvD, 104, 103514
Ivanov, M. M. 2023, in Handbook of Quantum Gravity, ed. C. Bambi, L. Modesto, & I. Shapiro (Singapore: Springer),
Ivanov, M. M., Cuesta-Lazaro, C., Mishra-Sharma, S., Obuljen, A., & Toomey, M. W. 2024, PhRvD, 110, 063538
Ivanov, M. M., Philcox, O. H. E., Cabass, G., et al. 2023, PhRvD, 107, 083515
Ivanov, M. M., Philcox, O. H. E., Nishimichi, T., et al. 2022a, PhRvD, 105, 063512
Ivanov, M. M., Philcox, O. H. E., Simonović, M., et al. 2022b, PhRvD, 105, 043531
Ivanov, M. M., & Sibiryakov, S. 2018, JCAP, 07, 053
Ivanov, M. M., Simonović, M., & Zaldarriaga, M. 2020a, JCAP, 05, 042
Ivanov, M. M., Simonović, M., & Zaldarriaga, M. 2020b, PhRvD, 101, 083504
Iwasawa, M., Tanikawa, A., Hosono, N., et al. 2016, PASJ, 68, 54
Jeffrey, N., Alsing, J., & Lanusse, F. 2021, MNRAS, 501, 954
Jennings, E., Li, Y., & Hu, W. 2013, MNRAS, 434, 2167
Kaiser, N. 1987, MNRAS, 227, 1
Kitaura, F.-S., Rodríguez-Torres, S., Chuang, C.-H., et al. 2016, MNRAS, 456, 4156
Kobayashi, Y., Nishimichi, T., Takada, M., & Miyatake, H. 2022, PhRvD, 105, 083517
Kobayashi, Y., Nishimichi, T., Takada, M., Takahashi, R., & Osato, K. 2020, PhRvD, 102, 063504
Kostić, A., Nguyen, N.-M., Schmidt, F., & Reinecke, M. 2023, JCAP, 2023, 063
Kreisch, C. D., Pisani, A., Villaescusa-Navarro, F., et al. 2022, ApJ, 935, 100
Kwan, J., Saito, S., Leauthaud, A., et al. 2023, ApJ, 952, 80
Kwon, K. J., & Hahn, C. 2024, ApJ, 976, 76
Kwon, K. J., Hahn, C., & Alsing, J. 2023, ApJS, 265, 23
Lahav, O., Lilje, P. B., Primack, J. R., & Rees, M. J. 1991, MNRAS, 251, 128
Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64
Lange, J. U., Hearin, A. P., Leauthaud, A., et al. 2022, MNRAS, 509, 1779
Lange, J. U., Hearin, A. P., Leauthaud, A., et al. 2023, MNRAS, 520, 5373
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193
Lavaux, G., & Wandelt, B. D. 2012, ApJ, 754, 109
Lazeyras, T., & Schmidt, F. 2018, JCAP, 09, 008
Levi, M., Bebek, C., Beers, T., et al. 2013, arXiv:1308.0847
Lewis, A. 2019, arXiv:1910.13970
Maion, F., Angulo, R. E., Bakx, T., et al. 2024, MNRAS, 531, 2684
Maksimova, N. A., Garrison, L. H., Eisenstein, D. J., et al. 2021, MNRAS, 508, 4017
Marinoni, C., Guzzo, L., Cappi, A., et al. 2008, A&A, 487, 7
Marulli, F., Veropalumbo, A., & Moresco, M. 2016, A&C, 14, 35
Massara, E., Villaescusa-Navarro, F., Hahn, C., et al. 2023, ApJ, 951, 70
Matsubara, T. 2008, PhRvD, 78, 083519
Matsubara, T. 2015, PhRvD, 92, 023534
McEwen, J. E., & Weinberg, D. H. 2018, MNRAS, 477, 4348
Modi, C., Chen, S.-F., & White, M. 2020, MNRAS, 492, 5754
Modi, C., Pandey, S., Ho, M., et al. 2025, MNRAS, 536, 254
Modi, C., & Philcox, O. H. E. 2023, arXiv:2309.10270
Moresco, M., Amati, L., Amendola, L., et al. 2022, LRR, 25, 6
Namekata, D., Iwasawa, M., Nitadori, K., et al. 2018, PASJ, 70, 70
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493
Neyrinck, M. C. 2008, MNRAS, 386, 2101
Nguyen, N.-M., Schmidt, F., Lavaux, G., & Jasche, J. 2021, JCAP, 03, 058
Nguyen, N.-M., Schmidt, F., Tucci, B., Reinecke, M., & Kostić, A. 2024, PhRvL, 133, 221006
Nicola, A., Hadzhiyska, B., Findlay, N., et al. 2024, JCAP, 2024, 015
Nishimichi, T., D'Amico, G., Ivanov, M. M., et al. 2020, PhRvD, 102, 123541
Nishimichi, T., Takada, M., Takahashi, R., et al. 2019, ApJ, 884, 29
Nitadori, K., Makino, J., & Hut, P. 2006, NewA, 12, 169
Nunes, R. C., Vagnozzi, S., Kumar, S., Di Valentino, E., & Mena, O. 2022, PhRvD, 105, 123506
Paillas, E., Cai, Y.-C., Padilla, N., & Sánchez, A. G. 2021, MNRAS, 505, 5731
Paillas, E., Cuesta-Lazaro, C., Zarrouk, P., et al. 2023, MNRAS, 522, 606
Peebles, P. J. E. 1980, The Large-scale Structure of the Universe (Princeton, NJ: Princeton Univ. Press)
Pellejero Ibañez, M., Angulo, R. E., Jamieson, D., & Li, Y. 2024, MNRAS, 529, 89
Pellejero Ibañez, M., Angulo, R. E., Zennaro, M., et al. 2023, MNRAS, 520, 3725
Pellejero Ibañez, M., Stücker, J., Angulo, R. E., et al. 2022, MNRAS, 514, 3993
Pelliciari, D., Contarini, S., Marulli, F., et al. 2023, MNRAS, 522, 152
Petri, A., Liu, J., Haiman, Z., et al. 2015, PhRvD, 91, 103511
Philcox, O. H. E. 2021a, PhRvD, 103, 103504
Philcox, O. H. E. 2021b, PhRvD, 104, 123529
Philcox, O. H. E., & Ivanov, M. M. 2022, PhRvD, 105, 043517
Philcox, O. H. E., Ivanov, M. M., Cabass, G., et al. 2022, PhRvD, 106, 043530
Philcox, O. H. E., Ivanov, M. M., Zaldarriaga, M., Simonovic, M., & Schmittfull, M. 2021a, PhRvD, 103, 043508
Pisani, A., Lavaux, G., Sutter, P. M., & Wandelt, B. D. 2014, MNRAS, 443, 3238
Philcox, O. H. E., Sherwin, B. D., Farren, G. S., & Baxter, E. J. 2021b, PhRvD, 103, 023538
Pisani, A., Massara, E., Spergel, D. N., et al. 2019, BAAS, 51, 40
Pisani, A., Sutter, P. M., Hamaus, N., et al. 2015a, PhRvD, 92, 083531
Pisani, A., Sutter, P. M., & Wandelt, B. D. 2015b, arXiv:1506.07982
Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A&A, 594, A13
Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, A&A, 641, A6
Platen, E., Van De Weygaert, R., & Jones, B. J. T. 2007, MNRAS, 380, 551
Pollina, G., Hamaus, N., Dolag, K., et al. 2017, MNRAS, 469, 787
Pollina, G., Hamaus, N., Paech, K., et al. 2019, MNRAS, 487, 2836
Press, W. H., Flannery, B. P., & Teukolsky, S. A. 1986, Numerical Recipes. The Art of Scientific Computing (Cambridge: Cambridge Univ. Press)
Ramanah, D. K., Lavaux, G., Jasche, J., & Wandelt, B. D. 2019, A&A, 621, A69
Reddick, R. M., Tinker, J. L., Wechsler, R. H., & Lu, Y. 2014, ApJ, 783, 118
Régaldo-Saint Blancard, B., Hahn, C., Ho, S., et al. 2024, PhRvD, 109, 083535
Reid, B. A., Seo, H.-J., Leauthaud, A., Tinker, J. L., & White, M. 2014, MNRAS, 444, 476
Rimes, C. D., & Hamilton, A. J. S. 2005, MNRAS, 360, L82
Ronconi, T., Contarini, S., Marulli, F., Baldi, M., & Moscardini, L. 2019, MNRAS, 488, 5075
Ronconi, T., & Marulli, F. 2017, A&A, 607, A24
Saadeh, D., Koyama, K., & Morice-Atkinson, X. 2025, MNRAS, 537, 448
Sahlén, M., Zubeldía, Í., & Silk, J. 2016, ApJL, 820, L7
Salcedo, A. N., Weinberg, D. H., Wu, H.-Y., & Wibking, B. D. 2022a, MNRAS, 510, 5376
Salcedo, A. N., Zu, Y., Zhang, Y., et al. 2022b, SCPMA, 65, 109811
Sánchez, A. G., Scoccimarro, R., Crocce, M., et al. 2017, MNRAS, 464, 1640
Schaye, J., Kugel, R., Schaller, M., et al. 2023, MNRAS, 526, 4978
Schmidt, F. 2021a, JCAP, 04, 032
Schmidt, F. 2021b, JCAP, 04, 033
Schmidt, F., Cabass, G., Jasche, J., & Lavaux, G. 2020, JCAP, 11, 008
Schmidt, F., Elsner, F., Jasche, J., Nguyen, N. M., & Lavaux, G. 2019, JCAP, 01, 042
Schmittfull, M., Simonović, M., Assassi, V., & Zaldarriaga, M. 2019, PhRvD, 100, 043514
Schmittfull, M., Simonović, M., Ivanov, M. M., Philcox, O. H. E., & Zaldarriaga, M. 2021, JCAP, 05, 059
Schuster, N., Hamaus, N., Dolag, K., & Weller, J. 2023, JCAP, 2023, 031
Schuster, N., Hamaus, N., Pisani, A., et al. 2019, JCAP, 2019, 055
Scoccimarro, R. 1998, MNRAS, 299, 1097
Scoccimarro, R. 2004, PhRvD, 70, 083007
Scoccimarro, R. 2015, PhRvD, 92, 083532
Seljak, U., Aslanyan, G., Feng, Y., & Modi, C. 2017, JCAP, 12, 009
Senatore, L., & Zaldarriaga, M. 2015, JCAP, 02, 013
Sheth, R. K., & van de Weygaert, R. 2004, MNRAS, 350, 517
Speagle, J., & Barbary, K., 2018 dynesty: Dynamic Nested Sampling Package, Astrophysics Source Code Library, ascl:1809.013
Speagle, J. S. 2020, MNRAS, 493, 3132
Stadler, J., Schmidt, F., & Reinecke, M. 2023, JCAP, 2023, 069
Stevens, A. R. H., Sinha, M., Rohl, A., et al. 2024, PASA, 41, e053
Storey-Fisher, K., Tinker, J. L., Zhai, Z., et al. 2024, ApJ, 961, 208
Sugiyama, N. S., Yamauchi, D., Kobayashi, T., et al. 2023, MNRAS, 523, 3133
Sutter, P. M., Lavaux, G., Hamaus, N., et al. 2014a, MNRAS, 442, 462
Sutter, P. M., Lavaux, G., Wandelt, B. D., & Weinberg, D. H. 2012, ApJ, 761, 187
Sutter, P. M., Lavaux, G., Wandelt, B. D., et al. 2014b, MNRAS, 442, 3127
Takada, M., et al. 2014, PASJ, 66, R1

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. 2018, arXiv:1804.06788
Tanikawa, A., Yoshikawa, K., Nitadori, K., & Okamoto, T. 2013, NewA, 19, 74
Tanikawa, A., Yoshikawa, K., Okamoto, T., & Nitadori, K. 2012, NewA, 17, 82
Taruya, A., Nishimichi, T., & Saito, S. 2010, PhRvD, 82, 063522
Thiele, L., Massara, E., Pisani, A., et al. 2024, ApJ, 969, 89
To, C. H., DeRose, J., Wechsler, R. H., et al. 2024, ApJ, 961, 59
Tucci, B., & Schmidt, F. 2024, JCAP, 2024, 063
Valogiannis, G., Yuan, S., & Dvorkin, C. 2024, PhRvD, 109, 103503
Van Den Bosch, F. C., Weinmann, S. M., Yang, X., et al. 2005, MNRAS, 361, 1203
Vasudevan, A., Ivanov, M. M., Sibiryakov, S., & Lesgourgues, J. 2019, JCAP, 09, 037
Verde, L., Heavens, A. F., Percival, W. J., et al. 2002, MNRAS, 335, 432
Verza, G., Carbone, C., Pisani, A., & Renzi, A. 2023, JCAP, 2023, 044
Verza, G., Carbone, C., & Renzi, A. 2022, ApJL, 940, L16
Verza, G., Pisani, A., Carbone, C., Hamaus, N., & Guzzo, L. 2019, JCAP, 2019, 040
Villaescusa-Navarro, F., Hahn, C., Massara, E., et al. 2020, ApJS, 250, 2
Wadekar, D., Ivanov, M. M., & Scoccimarro, R. 2020, PhRvD, 102, 123521
Wadekar, D., & Scoccimarro, R. 2020, PhRvD, 102, 123517
Wang, K., Mao, Y.-Y., Zentner, A. R., et al. 2022, MNRAS, 516, 4003
Wang, Z., Jeong, D., Taruya, A., Nishimichi, T., & Osato, K. 2023, PhRvD, 107, 103534
Wechsler, R. H., DeRose, J., Busha, M. T., et al. 2022, ApJ, 931, 145
Wechsler, R. H., & Tinker, J. L. 2018, ARA&A, 56, 435
White, M., Tinker, J. L., & McBride, C. K. 2014, MNRAS, 437, 2594
Xu, X., Zehavi, I., & Contreras, S. 2021, MNRAS, 502, 3242
Yoshikawa, K., & Fukushige, T. 2005, PASJ, 57, 849
Yuan, S., Eisenstein, D. J., & Garrison, L. H. 2018, MNRAS, 478, 2019
Yuan, S., Garrison, L. H., Eisenstein, D. J., & Wechsler, R. H. 2022a, MNRAS, 515, 871
Yuan, S., Garrison, L. H., Hadzhiyska, B., Bose, S., & Eisenstein, D. J. 2022b, MNRAS, 510, 3301
Yuan, S., Hadzhiyska, B., & Abel, T. 2023a, MNRAS, 520, 6283
Yuan, S., Zamora, A., & Abel, T. 2023b, MNRAS, 522, 3935
Yuan, S., Abel, T., & Wechsler, R. H. 2024a, MNRAS, 527, 1993
Yuan, S., Zhang, H., Ross, A. J., et al. 2024b, MNRAS, 530, 947
Zennaro, M., Angulo, R. E., Contreras, S., Pellejero-Ibáñez, M., & Maion, F. 2022, MNRAS, 514, 5443
Zennaro, M., Angulo, R. E., Pellejero-Ibáñez, M., et al. 2023, MNRAS, 524, 2407
Zhai, Z., Percival, W. J., & Guo, H. 2023a, MNRAS, 523, 5538
Zhai, Z., Tinker, J. L., Banerjee, A., et al. 2023b, ApJ, 948, 99
Zheng, Z., Berlind, A. A., Weinberg, D. H., et al. 2005, ApJ, 633, 791
Zheng, Z., Coil, A. L., & Zehavi, I. 2007, ApJ, 667, 760