Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **531,** 3336–3356 (2024) https://doi.org/10.1093/mnras/stae1234
Advance Access publication 2024 June 6

SUNBIRD: a simulation-based model for full-shape density-split clustering

Carolina Cuesta-Lazaro,[1,2,3★] Enrique Paillas,[4,5] Sihan Yuan,[6,7,8] Yan-Chuan Cai,[9] Seshadri Nadathur,[10] Will J. Percival,[4,5,11] Florian Beutler,[9] Arnaud de Mattia,[12] Daniel J. Eisenstein,[1] Daniel Forero-Sanchez,[13] Nelson Padilla,[14] Mathilde Pinon,[12] Vanina Ruhlmann-Kleider,[12] Ariel G. Sánchez,[15] Georgios Valogiannis[16,17] and Pauline Zarrouk[18]

[1] *Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA*
[2] *The NSF AI Institute for Artificial Intelligence and Fundamental Interactions*
[3] *Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[4] *Waterloo Centre for Astrophysics, University of Waterloo, Waterloo, ON N2L 3G1, Canada*
[5] *Department of Physics and Astronomy, University of Waterloo, Waterloo, ON N2L 3G1, Canada*
[6] *Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, 452 Lomita Mall, Stanford, CA 94305, USA*
[7] *Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305, USA*
[8] *SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA*
[9] *Institute for Astronomy, University of Edinburgh, Blackford Hill, Edinburgh, EH9 3HJ, UK*
[10] *Institute of Cosmology and Gravitation, University of Portsmouth, Burnaby Road, Portsmouth, PO1 3FX, UK*
[11] *Perimeter Institute for Theoretical Physics, 31 Caroline St North, Waterloo, ON N2L 2Y5, Canada*
[12] *IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France*
[13] *Institute of Physics, Laboratory of Astrophysics, École Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, CH-1290 Versoix, Switzerland*
[14] *Instituto de Astronomía Teórica y Experimental (IATE), CONICET-Universidad Nacional de Córdoba, Laprida 854, X5000BGR, Córdoba, Argentina*
[15] *Max-Planck-Institut für Extraterrestrische Physik, Postfach 1312, Giessenbachstr., D-85748 Garching, Germany*
[16] *Department of Physics, Harvard University, Cambridge, MA, 02138, USA*
[17] *Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL, 60637, USA*
[18] *Sorbonne Université, Université Paris Diderot, Sorbonne Paris Cité, CNRS, Laboratoire de Physique Nucléaire et de Hautes Energies (LPNHE), 4 place Jussieu, F-75252, Paris Cedex 5, France*

Accepted 2024 May 8. Received 2024 May 3; in original form 2023 September 28

ABSTRACT

Combining galaxy clustering information from regions of different environmental densities can help break cosmological parameter degeneracies and access non-Gaussian information from the density field that is not readily captured by the standard two-point correlation function (2PCF) analyses. However, modelling these density-dependent statistics down to the non-linear regime has so far remained challenging. We present a simulation-based model that is able to capture the cosmological dependence of the full shape of the density-split clustering (DSC) statistics down to intra-halo scales. Our models are based on neural-network emulators that are trained on high-fidelity mock galaxy catalogues within an extended-ΛCDM framework, incorporating the effects of redshift-space, Alcock–Paczynski distortions, and models of the halo–galaxy connection. Our models reach sub-per cent level accuracy down to $1\,h^{-1}$Mpc and are robust against different choices of galaxy–halo connection modelling. When combined with the galaxy 2PCF, DSC can tighten the constraints on $\omega_{\rm cdm}$, σ_8, and n_s by factors of 2.9, 1.9, and 2.1, respectively, compared to a 2PCF-only analysis. DSC additionally puts strong constraints on environment-based assembly bias parameters.

Key words: cosmological parameters – large-scale structure of Universe.

1 INTRODUCTION

The 3D clustering of galaxies contains a wealth of information about the contents and evolution of the universe; from the properties of the early universe to the nature of dark energy and dark matter, and to information on how galaxies form and evolve. Galaxy clustering provided some of the first evidence of the accelerated universe (Maddox et al. 1990), helped establish the standard model of cosmology through the detection of baryon acoustic oscillations (Percival et al. 2001; Cole et al. 2005; Eisenstein et al. 2005), and has yielded accurate cosmological constraints (Anderson et al. 2014). Upcoming surveys such as DESI (DESI Collaboration 2016), *Euclid* (Laureijs et al. 2011), and Roman (Green et al. 2012) will probe unprecedented volumes, enabling more stringent constraints that may reveal inconsistencies challenging the standard cosmological model or our understanding of how galaxies form and evolve.

★ E-mail: cuestalz@mit.edu

The spatial distribution of galaxies is commonly summarized by its two-point functions, the so-called two-point correlation function (2PCF) or its Fourier space equivalent, the power spectrum. For a Gaussian random field, this compression would be lossless. As the distribution of density fluctuations evolves through gravitational collapse, it becomes non-Gaussian: although overdensities can grow freely, underdensities are always bounded from below, as the density contrast in regions devoid of matter can never go below $\delta = -1$. As a consequence, the density field develops significant skewness and kurtosis, departing from Gaussianity (Einasto et al. 2021).

The induced non-Gaussianity in galaxy clustering deems the correlation function a lossy summary. For this reason, cosmologists have developed a wealth of summary statistics that may be able to extract more relevant information from the 3D clustering of galaxies. Examples include the three-PCF (Slepian & Eisenstein 2017) or bispectrum (Gil-Marín et al. 2017; Sugiyama et al. 2019; Philcox & Ivanov 2022), the four-PCF (Philcox, Hou & Slepian 2021) or trispectrum (Gualdi, Gil-Marín & Verde 2021), counts-in-cells statistics (Szapudi & Pan 2004; Klypin et al. 2018; Jamieson & Loverde 2020; Uhlemann et al. 2020), non-linear transformations of the density field (Neyrinck, Szapudi & Szalay 2009; Neyrinck 2011; Wang et al. 2011, 2024), the separate universe approach (Chiang et al. 2015), the marked power spectrum (Massara & Sheth 2018; Massara et al. 2023), the wavelet scattering transform (Valogiannis & Dvorkin 2022a, b), void statistics (Hawken et al. 2020; Nadathur et al. 2020; Correa et al. 2020; Woodfinden et al. 2022), k-nearest neighbours (Banerjee & Abel 2020; Yuan, Zamora & Abel 2023b), and other related statistics. Alternatively, one could avoid the use of summary statistics completely and attempt to perform inference at the field level (Lavaux, Jasche & Leclercq 2019; Schmidt 2021; Dai & Seljak 2022, 2024).

However, utilizing these summary statistics has been limited by our inability to model them analytically over a wide range of scales, difficulty compressing their high dimensionality, or due to a lack of accurate perturbation theory predictions or the difficulty in modelling the effect that observational systematics have on arbitrary summary statistics (Yuan, Hadzhiyska & Abel 2023a). This has now drastically changed due to (i) advancements in simulations: we now can run large suites of high-resolution simulations in cosmological volumes DeRose et al. (2019); Nishimichi et al. (2019); Maksimova et al. (2021), which enable us to forward model the relation between the cosmological parameters and the summary statistics with greater accuracy; and (ii) progress in machine learning techniques that allow us to perform inference on any set of parameters, θ, given any summary statistic, s, provided we can forward model the relation $s(\theta)$ for a small set of θ values (Cranmer, Brehmer & Louppe 2020). Examples of the latter in cosmology are emulators, that model $s(\theta)$ mainly through neural networks or Gaussian processes (Heitmann et al. 2009; DeRose et al. 2019; Zhai et al. 2023) and assume a Gaussian likelihood, or density estimators used to model directly the posterior distribution $p(\theta|s(x))$ (Jeffrey, Alsing & Lanusse 2020; Hahn et al. 2023) and make no assumptions about the likelihood's distribution.

While these advancements allow us to constrain cosmology with remarkable accuracy, our primary focus extends beyond just finding the most informative summary statistics. We are interested in statistics that could lead to surprising results revising our understanding of how the universe formed and evolved. Notably, models beyond Einstein gravity that add degrees of freedom in the gravitational sector must screen themselves from local tests of gravity, and can therefore only deviate from general relativity in regions of low-density or low-gravitational potential (Joyce et al. 2015; Hou et al. 2023). Therefore, surprises in this direction could be found in statistics that explore the dependency of galaxy clustering to different density environments. Moreover, previous work (Paillas et al. 2021; Bonnaire et al. 2022; Paillas et al. 2023b) has demonstrated that these statistics also have a large constraining power on the cosmological parameters.

Although we have mentioned earlier that we can now run large suites of simulations in cosmological volumes, this is only true for N-body, dark matter-only simulations. We still need a flexible and robust galaxy–dark matter connection model that allows us to populate dark matter simulations with realistic galaxy distributions. In this work, we employ halo occupation distribution (HOD) models, which use empirical relations to describe the distribution of galaxies in a halo based on the halo's mass and other secondary halo properties. In particular, recent studies have found the halo local density to be a good tracer of dark matter halo secondary properties, both in hydrodynamical simulations (Hadzhiyska et al. 2020) and semi-analytical models of galaxy formation (Xu, Zehavi & Contreras 2021).

Here, we present a full-shape theory model for galaxy clustering in different density environments that can be used to infer the cosmological parameters from observations in a robust manner. In a companion paper (Paillas et al. 2023a), we present the first cosmological constraints resulting from density-split clustering (DSC) using the model presented in this manuscript that we apply to the BOSS DR12 CMASS data (Reid et al. 2016; Dawson et al. 2016).

The paper is organized as follows. We define the observables and how we model them in Section 2. In Section 3, we demonstrate that the model can accurately recover the parameters of interest in a range of mock galaxy observations. We discuss our results and compare them to previous findings in the literature in Section 4.

2 A SIMULATION-BASED MODEL FOR DENSITY-SPLIT STATISTICS

We are interested in modelling the connection between the cosmological parameters, $\mathcal{C}$, the additional parameters describing how galaxies populate the cosmic web of dark matter, $\mathcal{G}$, and clustering as a function of density environment, $\mathrm{X}^{\mathrm{obs}}$. To solve the inverse problem and constrain $\mathcal{C}$ and $\mathcal{G}$ from data, we could use simulated samples drawn from the joint distribution $p(\mathcal{C}, \mathcal{G}, \mathbf{X}^{\mathrm{obs}})$ to either, (i) model the likelihood of the observation $p(\mathbf{X}^{\mathrm{obs}}|\mathcal{C}, \mathcal{G})$, subsequently sampling its posterior using Monte Carlo methods, or (ii) directly model the posterior distribution $p(\mathcal{C}, \mathcal{G}|\mathbf{X}^{\mathrm{obs}})$, as demonstrated in Jeffrey et al. (2020); Hahn et al. (2023), thus circumventing assumptions about the likelihood's functional form. Due to the Central Limit Theorem, we anticipate the likelihood of galaxy pair counts to approximate a Gaussian distribution. In this section, we validate that this holds true specifically for density-split statistics and elucidate how simulations can model its mean and covariance. Additionally, modelling the likelihood implies that we can use it as a measure of goodness-of-fit, vary the priors of the analysis at will, and combine our constraints with those of other independent observables.

In this section, we will proceed as follows: we begin by detailing our method for estimating density-dependent clustering. Subsequently, we discuss our approach for simulating the observable for a CMASS-like mock galaxy sample. We conclude by introducing our neural network model of the observable's likelihood.

2.1 The observables

2.1.1 *Two-point clustering*

The information contained on 3D galaxy maps is commonly summarized in terms of the 2PCF $\xi^{\mathrm{gg}}(r)$(or the power spectrum in Fourier space), which measures the excess probability dP of finding a pair of galaxies separated by a scale $\mathbf{r}$ within a volume dV, relative to an unclustered Poisson distribution

$$\mathrm{d}P = \overline{n}\left[1 + \xi^{\mathrm{gg}}(\mathbf{r})\right]\mathrm{d}V\,, \tag{1}$$

where $\overline{n}$ denotes the mean galaxy density. While the spatial distribution of galaxies is isotropic in real space, there are two main sources of distortions that induce anisotropies in the clustering measured from galaxy surveys: redshift-space distortions (RSDs) and Alcock–Paczynski (AP) distortions, which are dynamical and geometrical in nature, respectively.

RSDs arise when converting galaxy redshifts to distances ignoring the peculiar motion of the galaxies. A pair of galaxies that is separated by a vector $\mathbf{r}$ in real space, will instead appear separated by a vector $\mathbf{s}$ in redshift space (to linear order in velocity):

$$\mathbf{s} = \mathbf{r} + \frac{\mathbf{v}\cdot\hat{\mathbf{x}}}{a(z)H(z)}\hat{\mathbf{x}}\,, \tag{2}$$

where $\hat{\mathbf{x}}$ is the unit vector associated with the observer's line-of-sight, $\mathbf{v}$ is the peculiar velocity of the galaxy, $a(z)$ is the scale factor, and $H(z)$ is the Hubble parameter.

AP distortions arise when the cosmology that is adopted to convert angles and redshifts to distances, denoted as fiducial cosmology, differs from the true cosmology of the universe. This effect is partially degenerate with RSD. For close pairs, the true pair separation is related to the observed pair separation via the parameters $q_\perp$ and $q_\parallel$, which distort the components of the pair separation across and along the observer's line-of-sight

$$r_\perp = q_\perp r_\perp^{\mathrm{fid}}\,;\quad r_\parallel = q_\parallel r_\parallel^{\mathrm{fid}}\,, \tag{3}$$

where the $^{\mathrm{fid}}$ superscript represents the separations measured in the fiducial cosmology. The distortion parameters are given by

$$q_\parallel = \frac{D_{\mathrm{H}}(z)}{D_{\mathrm{H}}^{\mathrm{fid}}(z)}\,;\quad q_\perp = \frac{D_{\mathrm{M}}(z)}{D_{\mathrm{M}}^{\mathrm{fid}}(z)}\,, \tag{4}$$

where $D_{\mathrm{M}}(z)$ and $D_{\mathrm{H}}(z)$ are the comoving angular diameter and Hubble distances to redshift z, respectively.

Due to RSD and AP, the 2PCF is no longer isotropic but depends on s, the pair separation, and μ, the cosine of the angle between the galaxy pair separation vector and the mid-point line-of-sight. The two-dimensional correlation function can be decomposed in a series of multipole moments

$$\xi_\ell(s) = \frac{2\ell+1}{2}\int_{-1}^{1}\mathrm{d}\mu\,\xi(s,\mu)\mathrm{P}_\ell(\mu), \tag{5}$$

where P_ℓ is the ℓ-th order Legendre polynomial.

2.1.2 DSC

The density-split method (Paillas et al. 2023b) characterizes galaxy clustering in environments of different local densities. Instead of calculating the two-point clustering of the whole galaxy sample at once, one first splits a collection of randomly placed query points in different bins or 'quantiles', according to the local galaxy-overdensity at their locations. The two-point clustering is then calculated for each environment separately, and all this information is then combined in a joint likelihood analysis. The algorithm can be summarized as follows:

(i) Redshift-space galaxy positions are assigned to a rectangular grid with a cell size R_{cell}, and the overdensity field is estimated using a cloud-in-cell interpolation scheme. The field is smoothed using a Gaussian filter with radius R_s, which is performed in Fourier space for computational efficiency.

(ii) A set of N_{query} random points are divided into N_Q density bins, or quantiles, according to the overdensity measured at each point.

(iii) Two summary statistics are calculated for each quantile: the autocorrelation function (DS ACF) of the query points in each quantile, and the cross-correlation function (DS CCF) between the quantiles and the entire redshift-space galaxy field. These correlation functions are then decomposed into multipoles (equation (5)).

(iv) The collection of correlation functions of all quantiles but the middle-one, is combined in a joint data vector, which is then fitted in a likelihood analysis to extract cosmological information.

In Fig. 1, we show the different density-split summary statistics for five quantiles and $R_s = 10\,h^{-1}$Mpc, as measured in the AbacusSummit simulations presented in Section 2.2.1. Note that the smoothing scale can be varied depending on the average density of tracers in a given survey, here we restrict ourselves to a smoothing scale appropriate for a CMASS-like survey. In the first column, we show the CCF of the different density quantiles and the entire galaxy sample. Above, the amplitude of the different correlations reflects the non-Gaussian nature of the density PDF: the most underdense regions, Q$_0$, are always constrained from below as voids cannot be emptier than empty ($\delta = -1$), meanwhile, dense regions, Q$_4$, can go well beyond 1, breaking the symmetry of the correlations. Around the scale of $100\,h^{-1}$Mpc we can distinguish the signal coming from the baryon acoustic oscillations for all density quantiles, both for the cross- and autocorrelations. Regarding the quadrupole moments, the anisotropy found is a consequence of the RSD effect on the galaxy positions, which also introduces an additional anisotropy in the distribution of quantiles when these are identified using the galaxy-redshift-space distribution, as shown in (Paillas et al. 2023b).

2.2 Forward modelling the galaxy observables

In this section, we will first present the suite of dark matter-only N-body simulations used in this work to model the cosmological dependence of DSC, and will later present the galaxy–halo connection model we adopt to build CMASS-like mock galaxy catalogues.

2.2.1 The AbacusSummit simulations

AbacusSummit (Maksimova et al. 2021) is a suite of cosmological N-body simulations that were run with the Abacus N-body code (Garrison, Eisenstein & Pinto 2019; Garrison et al. 2021), designed to meet and exceed the simulation requirements of DESI (Levi et al. 2019). The base simulations follow the evolution of 6912^3 dark matter particles in a $(2\,h^{-1}\mathrm{Gpc})^3$ volume, corresponding to a mass resolution of $2\times10^9\,\mathrm{M}_\odot/h$.

In total, the suite spans 97 different cosmologies, with varying

$$\mathcal{C} = \{\,\omega_{\mathrm{cdm}}, \omega_b, \sigma_8, n_s, \mathrm{d}n_s/\mathrm{d}\ln k, N_{\mathrm{eff}}, w_0, w_a\}, \tag{6}$$

where $\omega_{\mathrm{cdm}} = \Omega_{\mathrm{c}}h^2$ and $\Omega_{\mathrm{b}}h^2$ are the physical cold dark matter and baryon densities, dn_s/dln k is the running of the spectral tilt, N_{eff} is the effective number of ultra-relativistic species, w_0 is the present-day dark energy equation-of-state, and w_a captures the time evolution of the dark energy equation-of-state. The simulations assume a flat spatial curvature, and the Hubble constant H_0 is calibrated to match the Cosmic Microwave Background acoustic scale θ_* to the *Planck* 2018 measurement.

In this study, we focus on the following subsets of the AbacusSummit simulations:

(i) c000: *Planck* 2018 ΛCDM base cosmology (Planck Collaboration 2020), corresponding to the mean of the

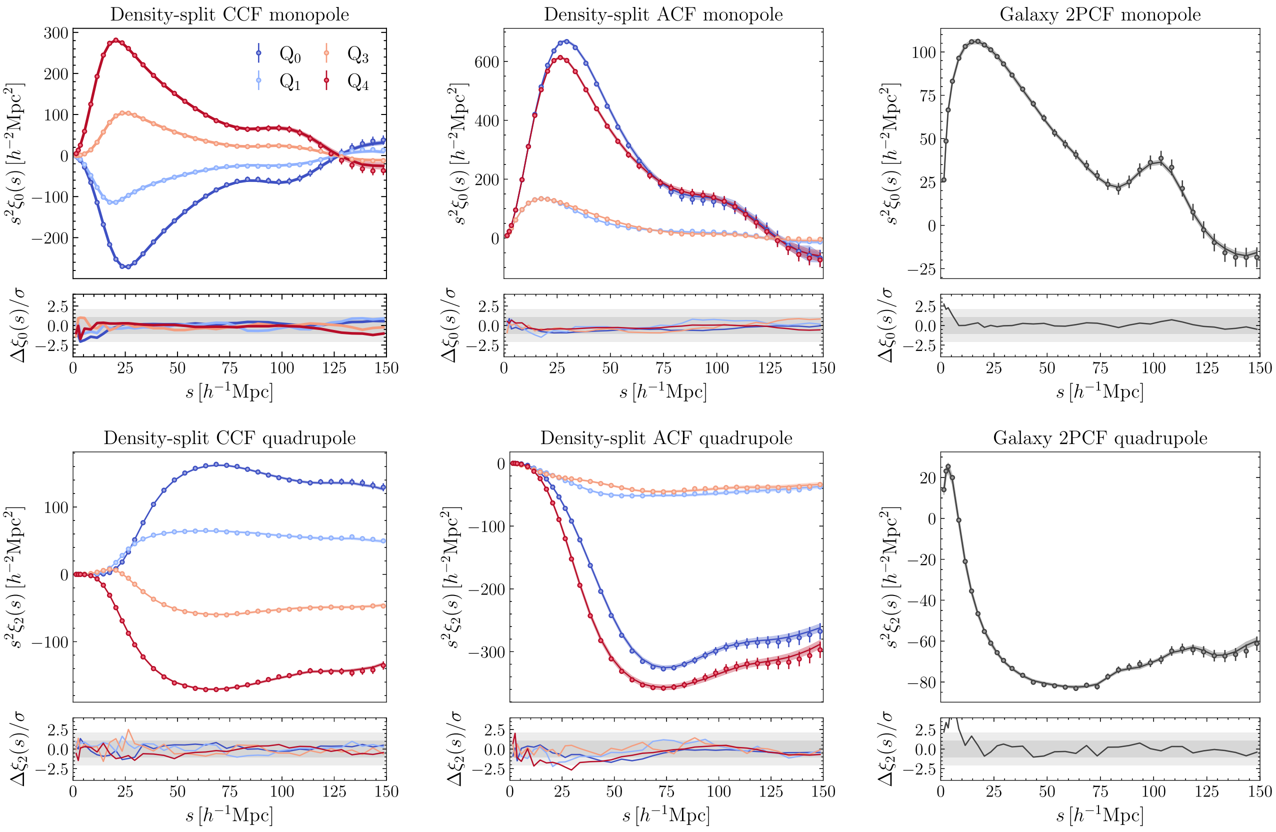

Figure 1. A visualization of the DSC data vectors from the AbacusSummit simulations, along with emulator prediction at the parameter values of the simulation. The lowest density quantile is shown in blue, Q_0, and the highest one in red, Q_4. Markers and solid lines show the data vectors and the emulator predictions, respectively, whereas the shaded area represents the emulator predicted uncertainty. Left: multipoles of the quantile–galaxy CCFs. Middle: multipoles of the quantile ACFs. Right: multipoles of the 2PCF. The upper and lower 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 colour corresponds to a different density quantile. https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/F1_data_vectors.py

base_plikHM_TTTEEE_lowl_lowE_lensing likelihood. There are 25 independent realizations of this cosmology.

(ii) `c001-004`: Secondary cosmologies, including a low $\omega_{\rm cdm}$ choice (WMAP7, Komatsu et al. 2011), a wCDM choice, a high-$N_{\rm eff}$ choice, and a low-σ_8 choice.

(iii) `c013`: Cosmology that matches *Euclid* Flagship2 ΛCDM (Castander et al., in preparation).

(iv) `c100-126`: A linear derivative grid that provides pairs of simulations with small negative and positive steps in an 8D cosmological parameter space

(v) `c130-181`: An emulator grid around the base cosmology that provides a wider coverage of the cosmological parameter space. Note that all the simulations in the emulator grid have the same phase seed. The parameter ranges in the emulator grid are shown in Table 1.

Moreover, we use a smaller set of 1643 N-body simulations denoted as AbacusSmall to estimate covariance matrices. These simulations are run with the same mass resolution as that of AbacusSummit in $500\,h^{-1}$Mpc boxes, with 1728^3 particles and varying phase seeds.

Group finding is done on the fly, using a hybrid Friends-of-Friends/Spherical Overdensity algorithm, dubbed CompaSO (Hadzhiyska et al. 2021). We use dark matter halo catalogues from snapshots of the simulations at $z = 0.5$ and populate them with galaxies using the extended HOD framework presented in Section 2.2.2.

2.2.2 *Modelling the galaxy–halo connection*

We model how galaxies populate the cosmic web of dark matter using the HOD framework, which populates dark matter haloes with galaxies in a probabilistic way, assuming that the expected number of galaxies in each halo correlates with some set of halo properties, the main one being halo mass.

In the base halo model (Zheng, Coil & Zehavi 2007), the average number of central galaxies in a halo of mass M is given by

$$\langle N_{\rm c}\rangle(M) = \frac{1}{2}\left(1 + {\rm erf}\left(\frac{\log M - \log M_{\rm cut}}{\sqrt{2}\sigma}\right)\right), \tag{7}$$

where erf(x) denotes the error function, $M_{\rm cut}$ is the minimum mass required to host a central, and σ is the slope of the transition between having zero and one central galaxy. The average number of satellite galaxies is given by

$$\langle N_{\rm s}\rangle(M) = \langle N_{\rm c}\rangle(M)\left(\frac{M - \kappa M_{\rm cut}}{M_1}\right)^{\alpha}, \tag{8}$$

where $\kappa M_{\rm cut}$ gives the minimum mass required to host a satellite, M_1 is the typical mass that hosts one satellite, and α is the power-law index for the number of galaxies. Note that these particular functional forms have been developed for the clustering of luminous red galaxies (LRGs) and should be modified for other tracers such as emission-line galaxies (ELGs).

Table 1. Definitions and ranges of the cosmological and galaxy–halo connection parameters for the simulations used to train our emulator.

	Parameter	Interpretation	Prior range
Cosmology	$\omega_{\rm cdm}$	Physical cold dark matter density	[0.103, 0.140]
	$\omega_{\rm b}$	Physical baryon density	[0.0207, 0.024]
	σ_8	Amplitude of matter fluctuations in 8 h^{-1}Mpc spheres	[0.687, 0.938]
	n_s	Spectral index of the primordial power spectrum	[0.901, 1.025]
	dn_s/dln k	Running of the spectral index	[−0.038, 0.038]
	$N_{\rm eff}$	Number of ultra-relativistic species	[2.1902, 3.9022]
	w_0	Present-day dark energy equation-of-state	[−1.27, −0.70]
	w_a	Time evolution of the dark energy equation-of-state	[−0.628, 0.621]
HOD	$M_{\rm cut}$	Minimum halo mass to host a central	[12.4, 13.3]
	M_1	Typical halo mass to host one satellite	[13.2, 14.4]
	$\log\sigma$	Slope of the transition from hosting zero to one central	[−3.0, 0.0]
	α	Power-law index for the mass dependence of the number of satellites	[0.7, 1.5]
	κ	Parameter that modulates the minimum halo mass to host a satellite	[0.0, 1.5]
	α_c	Velocity bias for centrals	[0.0, 0.5]
	α_s	Velocity bias for satellites	[0.7, 1.3]
	$B_{\rm cen}$	Environment-based assembly bias for centrals	[−0.5 0.5]
	$B_{\rm sat}$	Environment-based assembly bias for satellites	[−1.0, 1.0]

Alternatively, one could model the connection between dark matter haloes and galaxies through more complex models of galaxy formation such as semi-analytical models or hydrodynamical simulations. In these scenarios, the simplified assumptions of HOD models whose occupation parameters solely depend on halo mass have been found to break down. In particular, recent studies have found the halo local density to be a good tracer of dark matter halo secondary properties that control galaxy occupation, both in hydrodynamical simulations (Hadzhiyska et al. 2020) and semi-analytical models of galaxy formation (Xu et al. 2021). There is however no direct observational evidence of this effect so far, and we are interested in using density-split statistics to more accurately constrain the role that environment plays in defining the halo–galaxy connection.

In this work, we implement the HOD modelling using AbacusHOD (Yuan et al. 2021), which is a highly efficient python package that contains a wide range of HOD variations. In AbacusHOD, the environment-based secondary bias parameters, $B_{\rm cen}$ and $B_{\rm sat}$, effectively modulate the mass of a dark matter halo during the HOD assignment, so that it depends on the local matter overdensity δ_m

$$\log_{10} M_{\rm cut}^{\rm eff} = \log_{10} M_{\rm cut} + B_{\rm cen}(\delta_m - 0.5)$$
$$\log_{10} M_1^{\rm eff} = \log_{10} M_1 + B_{\rm sat}(\delta_m - 0.5)\,. \quad (9)$$

Here, δ_m is defined as the mass-density within a 5 h^{-1}Mpc top-hat filter from the halo centre, without considering the halo itself. More details about the exact implementation of this extension can be found in Yuan et al. (2021).

Moreover, we include velocity bias parameters to increase the flexibility of the model to describe the dynamics of galaxies within dark matter haloes, that ultimately influence galaxy clustering through RSDs. There is in fact observational evidence pointing towards central galaxies having a larger velocity dispersion than their host dark matter haloes (Guo et al. 2014; Yuan et al. 2021) for CMASS galaxies (dominated by LRGs), evidence for other tracers is not established yet. In the AbacusHOD implementation, the positions and velocities of central galaxies are matched to the most-bound particle in the halo, whereas the satellites follow the positions and velocities of randomly selected dark matter particles within the halo. The velocity bias parameters, $\alpha_{\rm vel,\,c}$ and $\alpha_{\rm vel,\,s}$, allow for offsets in these velocities, such that the centrals do not perfectly track the velocity of the halo centre, and the satellites do not exactly match the dark matter particle velocities. The exact velocity match is recovered when $\alpha_{\rm vel,\,c} = 0$ and $\alpha_{\rm vel,\,c} = 1$.

The extended-HOD framework used in this study is then comprised of 9 parameters

$$\mathcal{G} = \{M_{\rm cut}, M_1, \sigma, \alpha, \kappa, \alpha_{\rm vel,c}, \alpha_{\rm vel,s}, B_{\rm cen}, B_{\rm sat}\}\,. \quad (10)$$

Note that we are here not including additional parameters that may help marginalize over the effect that baryons have on halo density profiles. Although this has been shown to be a small effect (Bose et al. 2019), Yuan et al. (2021) presented an extended parametrization that could be use to marginalize over this effect.

2.2.3 *Generating mock galaxy catalogues*

We generate a Latin hypercube with 8500 samples from the 9D HOD parameter space defined in equation (10), with parameter ranges as listed in Table 1. Each of the 85 cosmologies is assigned 100 HOD variations from the Latin hypercube, which are then used to generate mock galaxy catalogues using the AbacusHOD. This number of HOD variations was chosen as a compromise between reducing the emulator error and increasing the computational cost of these measurements. In the future, we plan to develop a more efficient HOD sampling strategy to resample those HOD parameter values where the emulator error is large.

Our target galaxy sample is the DR12 BOSS CMASS galaxy sample (Reid et al. 2016) at $0.45 < z < 0.6$. If the resulting number density of an HOD catalogue is larger than the observed number density from CMASS, $n_{\rm gal} \approx 3.5 \times 10^{-4}\,(h/{\rm Mpc})^{-3}$, we invoke an incompleteness parameter $f_{\rm ic}$ and randomly downsample the catalogue to match the target number density.

The resulting HOD catalogues consist of the real-space galaxy positions and velocities. Under the distant-observer approximation, we map the positions of galaxies to redshift space by perturbing their coordinates along the line of sight with their peculiar velocities along the same direction (equation (2)). For each mock catalogue, we build three redshift-space counterparts by adopting three different lines-of-sight, taken to be the x, y, and z axes of the simulation, which can be averaged out in the clustering analysis to increase the signal-to-noise ratio of the correlation functions (Smith et al. 2020).

Since the end goal of our emulator is to be able to model galaxy clustering from observations, we adopt the same fiducial cosmology as in our CMASS clustering measurements (Paillas et al. 2023a)

$$\omega_{\rm cdm} = 0.12 \quad \omega_{\rm b} = 0.02237 \quad h = 0.6736$$
$$\sigma_8 = 0.807952 \quad n_s = 0.9649\,, \tag{11}$$

and infuse the mocks with the AP distortions that would be produced if we were to analyse each mock with this choice of fiducial cosmology. We do so by scaling the galaxy positions[1] and the simulation box dimensions with the distortion parameters from equation (4), which depend on the adopted fiducial cosmology and the true cosmology of each simulation. Since, in general, $q_\perp$ and $q_\parallel$ can be different, the box geometry can become non-cubic, but it still maintains the periodicity along the different axes. This is taken into account when calculating the clustering statistics, as explained in the further section.

2.2.4 *Generating the training sample*

We run the DSC pipeline on the HOD mocks using our publicly available code[2] redshift-space-galaxy positions are mapped onto a rectangular grid of resolution $R_{\rm cell} = 5\,h^{-1}$Mpc, smoothed with a Gaussian kernel of width $R_s = 10\,h^{-1}$Mpc. The overdensity field[3] is sampled at $N_{\rm query}$ random locations, where $N_{\rm query}$ is equal to five times the number of galaxies in the box. We split the query positions into five quantiles according to the overdensity at each location. We plan to explore the constraining power of the statistic based on different values of the smoothing scale and the number of quantiles in future work.

We measure the DS ACFs and CCFs of each DS quantile in bins of μ and s using PYCORR, which is a wrapper around a modified version of CORRFUNC (Sinha & Garrison 2020). We use 241 μ bins from −1 to 1, and radial bins of different widths depending on the scale: 1 Mpc h^{-1} bins for $0 < s < 4\ h^{-1}$Mpc, 3 Mpc h^{-1} bins for $4 < s < 30\ h^{-1}$Mpc, and 5 Mpc h^{-1} bins for $30 < s < 150\ h^{-1}$Mpc. Additionally, we measure the galaxy 2PCF adopting these same settings. All the correlation functions are then decomposed into their multipole moments (equation (5)). In this analysis, we decided to omit the hexadecapole due to its low-signal-to-noise ratio, restricting the analysis to the monopole and quadrupole. The multipoles are finally averaged over the three lines-of-sight.

Due to the addition of AP distortions, whenever the true cosmology of a mock does not match our fiducial cosmology, the boxes will have non-cubic dimensions while still maintaining the periodicity along the three axes. Both the DENSITYSPLIT and PYCORR codes can handle non-cubic periodic boundary conditions. In the case of DENSITYSPLIT, we choose to keep the resolution of the rectangular grid fixed, so that $R_{\rm cell} = 5\,h^{-1}$Mpc remains fixed irrespectively of the box dimensions (which, as a consequence, can change the number of cells that are required to span the different boxes). The smoothing scale R_s is also kept fixed to $10\,h^{-1}$Mpc, but since the underlying galaxy positions are AP-distorted, this mimics the scenario we would encounter in observations, where we make a choice of smoothing kernel and apply it to the distorted galaxy-overdensity field.

An example of the density-split summary statistics for `c000` and one of the sampled HOD parameters from the latin hypercube is shown in Fig. 1.

2.3 Defining the observable's likelihood

The data vector for DSC is the concatenation of the monopole and quadrupole of the ACFs and CCFs of quantiles Q$_0$, Q$_1$, Q$_3$, and Q$_4$. In the case of the galaxy 2PCF, it is simply the concatenation of the monopole and quadrupole. In Appendix A, we show that the likelihood of these data vectors is well-approximated by a Gaussian distribution as also demonstrated in Paillas et al. (2023b). We therefore define the log-likelihood as

$$\log \mathcal{L}(\mathbf{X}^{\rm obs}|\mathcal{C},\mathcal{G}) = \left(\mathbf{X}^{\rm obs} - \mathbf{X}^{\rm theo}\left(\mathcal{C},\mathcal{G}\right)\right) \mathbf{C}^{-1}\left(\mathbf{X}^{\rm obs} - \mathbf{X}^{\rm theo}\left(\mathcal{C},\mathcal{G}\right)\right)^{\top}\,, \tag{12}$$

where $\mathbf{X}^{\rm obs}$ is the observed data vector, $\mathbf{X}^{\rm theo}$ is the expected theoretical prediction dependent on $\mathcal{C}$, the cosmological parameters, and $\mathcal{G}$, the parameters describing how galaxies populate the cosmic web, referred to as galaxy bias parameters throughout this paper, and $\mathbf{C}$ the theoretical covariance of the summary statistics. We will here assume that the covariance matrix is independent of $\mathcal{C}$ and $\mathcal{G}$, and use simulations with varying random seeds to estimate it. This assumption has been shown to have a neglibible impact in parameter estimation for two-point functions (Kodwani, Alsono & Ferreira 2019), although it will need to be revised as the statistical precision of future surveys increases.

In the following section, we demonstrate how we can use neural networks to model the mean relation between cosmological and HOD parameters and the density-split statistics in the generated galaxy mocks.

2.3.1 *Emulating the mean with neural networks*

We split the suite of mocks of different cosmologies (and their corresponding HOD variations) into training, validation, and test sets. We assign cosmologies `c000`, `c001`, `c002`, `c003`, `c004`, and `c013` to the test set, while 80 per cent of the remaining cosmologies are randomly assigned to the training and 20 per cent to the validation set. See Section 2.2.1 for the definition of the different cosmologies.

We construct separate neural-network emulators for the galaxy 2PCF, the DS ACF, and the DS CCF. The inputs to the neural network are the cosmological and HOD parameters, normalized to lie between 0 and 1, and the outputs are the concatenated monopole and quadrupole of each correlation function, also normalized to be between 0 and 1.We train fully connected neural networks with Sigmoid Linear Units as activation functions (Elfwing, Uchibe & Doya 2018) and a negative Gaussian log-likelihood as the loss function

$$\mathcal{L}(\mathbf{X}|\mu_{\rm pred}(\mathcal{C},\mathcal{G}),\sigma_{\rm pred}(\mathcal{C},\mathcal{G})) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(X_i - \mu_{\rm pred}(\mathcal{C},\mathcal{G}))^2}{2\sigma_{\rm pred}(\mathcal{C},\mathcal{G})^2} + \log(\sigma_{\rm pred}(\mathcal{C},\mathcal{G})^2) + \frac{1}{2}\log(2\pi)\right). \tag{13}$$

[1] These distortions would have been naturally produced if we had started from galaxy catalogues in sky coordinates, and used our fiducial cosmology to convert them to comoving Cartesian coordinates. In our case, we have to manually distort the galaxy positions, since we are already starting from the comoving box.

[2] https://github.com/epaillas/densitysplit.

[3] The galaxy overdensity in each grid cell depends on the number of galaxies in the cell, the average galaxy number density, and the total number of grid cells. As we are working with a rectangular box with periodic boundary conditions, the average galaxy number density can be calculated analytically, which allows us to convert the galaxy number counts in each cell to an overdensity. When working with galaxy surveys, this has to be calculated using random catalogues that match the survey window function.

Where $\mu_{\rm pred}(\mathcal{C}, \mathcal{G})$, the mean of the log likelihood, emulates the theory predictions from the N-body simulations, $\sigma_{\rm pred}(\mathcal{C}, \mathcal{G})$ models the network's uncertainty in its prediction, and n is the batch size.

We use the AdamW optimization algorithm to optimize the weights of the neural network, together with a batch size of 256. In contrast to Adam, AdamW includes L2 regularization to ensure that large weights are only allowed when they significantly reduce the loss function. To further prevent overfitting, given the limited size of our data set, we also introduce a dropout factor (Srivastava et al. 2014). Finally, to improve the model's performance and reduce training time, we decrease the learning rate by a factor of 10 every 5 epochs over which the validation loss does not improve, until the minimum learning rate of 10^{-6} is reached.

We use Optuna[4] to find the hyperparameters of the neural network that produce the best validation loss. We optimize the following hyperparameters: learning rate, weight decay controlling the strength of L2 regularization, number of layers, number of hidden units in each layer, and the dropout rate, over 200 trials. More details related to the neural network architecture and its optimization can be found on our GitHub repository.[5]

In Section 3, we present an extensive validation of the emulator's accuracy.

2.3.2 Estimating the covariance matrix

The likelihood function in equation (12) requires defining the data vector, expected theoretical mean, and covariance matrix of the summary statistics. The total covariance matrix includes contributions from three sources (i) the intrinsic error of the emulator in reproducing simulations with identical phases to those of the training set ($\mathbf{C}_{\rm emu}$); (ii) the error related to the difference between the fixed-phase simulations used for training and the true ensemble mean ($\mathbf{C}_{\rm sim}$); and (iii) the error between the observational data and the mean ($\mathbf{C}_{\rm data}$)

$$\mathbf{C} = \mathbf{C}_{\rm data} + \mathbf{C}_{\rm emu} + \mathbf{C}_{\rm sim} \, . \tag{14}$$

Because the test sample is small and covers a range of cosmologies, to estimate the contribution from the emulator's error to the covariance matrix, we are limited to either assume a diagonal covariance matrix whose diagonal elements are the emulator's predicted uncertainties as a function of cosmological and HOD parameters, $\sigma_{\rm pred}(\mathcal{C}, \mathcal{G})$, or we can estimate the emulator error from the test set simulations and ignore its parameter dependence. For the latter, we compute the difference between measurements from the test set and the emulator predictions, $\Delta\mathbf{X} = \mathbf{X}^{\rm emu} - \mathbf{X}^{\rm test}$, and we estimate a covariance matrix as

$$\mathbf{C}_{\rm emu} = \frac{1}{n_{\rm test} - 1} \sum_{k=1}^{n_{\rm test}} \left(\Delta\mathbf{X}_k - \overline{\Delta\mathbf{X}_k}\right) \left(\Delta\mathbf{X}_k - \overline{\Delta\mathbf{X}_k}\right)^\top \, , \tag{15}$$

where the overline denotes the mean across all 600 test set mocks.

To estimate $\mathbf{C}_{\rm sim}$, we do a χ^2 minimization to choose an HOD catalogue from the fiducial `c000` cosmology that matches the density-split multipoles measured from BOSS CMASS (Paillas et al. 2023a). We then use those HOD parameters to populate dark matter haloes and measure the multipoles from multiple independent realizations of the small AbacusSummit boxes ran with different phases. The covariance is calculated as

$$\mathbf{C}_{\rm sim} = \frac{1}{n_{\rm sim} - 1} \sum_{k=1}^{n_{\rm sim}} \left(\mathbf{X}^{\rm sim}_k - \overline{\mathbf{X}^{\rm sim}}\right) \left(\mathbf{X}^{\rm sim}_k - \overline{\mathbf{X}^{\rm sim}}\right)^\top \, , \tag{16}$$

where $n_{\rm sim} = 1643$. Each of these boxes is $500\,h^{-1}$Mpc on a side, so we rescale the covariance by a factor of 1/64 to match the $(2\,h^{-1}{\rm Gpc})^3$ volume covered by the base simulations. See Howlett & Percival (2017) for an in-depth discussion on rescaling the covariance matrix by volume factors. For a volume such as that of CMASS, the contribution of $\mathbf{C}_{\rm sim}$ will be almost negligible. However, this will not be true for larger data sets such as those from the upcoming DESI galaxy survey (DESI Collaboration 2016). Alternatively, the phase correction routine introduced in appendix B of Yuan et al. (2022) could be used to reduce this contribution.

The calculation of $\mathbf{C}_{\rm data}$ depends on the sample that is used to measure the data vector. In this work, we estimate it from multiple realizations of the small AbacusSummit boxes, in the same way as we compute $\mathbf{C}_{\rm sim}$. Thus, in the current setup, $\mathbf{C}_{\rm data} = \mathbf{C}_{\rm sim}$. When fitting real observations, however, $\mathbf{C}_{\rm data}$ would have to be estimated from mocks that match the properties of the specific galaxy sample that is being used, or using other methods such as jackknife resampling. Importantly, the volume of AbacusSummit is much larger than the volume of the CMASS galaxy sample that we are targetting, and therefore we are providing a stringent test of our emulator framework.

In Fig. 2, we show the correlation matrix for both data and emulator. The full data vector, which combines DSC and the galaxy 2PCF, is comprised by 648 bins. This results in covariance matrices with 648^2 elements, showing significant (anti) correlations between the different components of the data vector. The horizontal and vertical black lines demarcate the contributions from different summary statistics. Starting from the bottom left, the first block along the diagonal represents the multipoles of the DS CCF, for all four quantiles. The second block corresponds to the DS ACF, and the last block corresponds to the galaxy 2PCF. The non-diagonal blocks show the cross-covariance between these different summary statistics.

3 VALIDATING THE NEURAL NETWORK EMULATOR

In this section, we present an exhaustive evaluation of the emulator's accuracy by, (i) assessing the network's accuracy at reproducing the test set multipoles, (ii) ensuring that the emulator recovers unbiased cosmological constraints when the test set is sampled from the same distribution as the training set, (iii) testing the ability of the emulator to recover unbiased cosmological constraints when applied to out-of-distribution data.

3.1 Testing the accuracy of the emulated multipoles

We first compare the multipoles measured from the test simulations against the emulator predictions. Fig. 1 shows the density-split and the 2PCF multipoles as measured from one of the HOD catalogues corresponding to the `c000` cosmology. The HOD catalogue is chosen among the prior samples to maximize the likelihood of the CMASS data set presented in Paillas et al. (2023a). The model predictions, which are overplotted as solid lines, show excellent agreement with the data on a wide range of scales. These theory lines are the emulator prediction for the true cosmology and HOD parameters from the mock catalogue.

[4] https://github.com/optuna/optuna
[5] https://github.com/florpi/sunbird

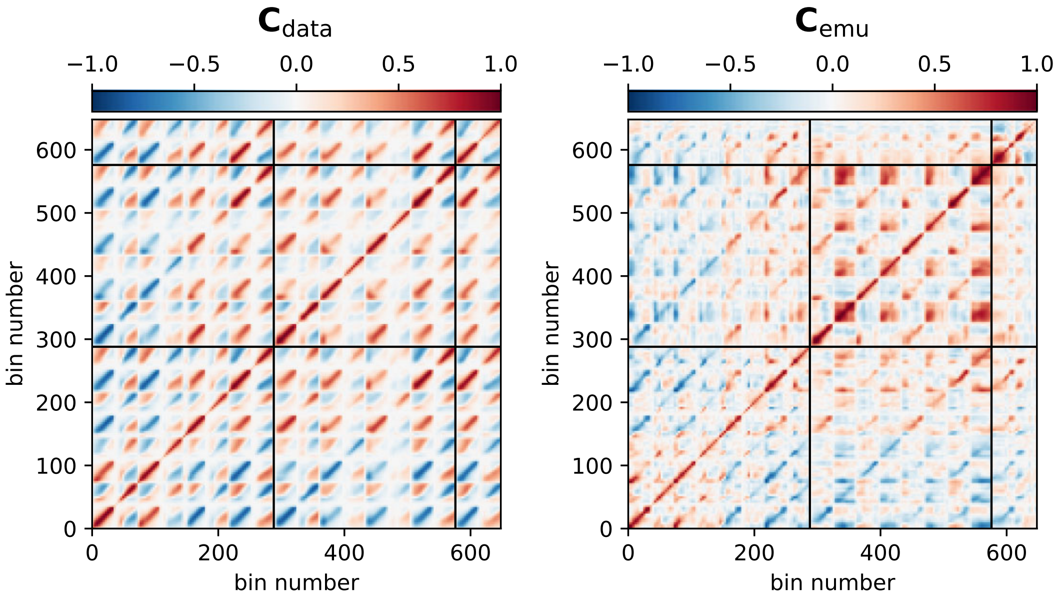

Figure 2. Correlation matrices of the data and model vectors in our clustering analysis. $\mathbf{C}_{\rm data}$ corresponds to errors associated with the sample variance of the data vector, while $\mathbf{C}_{\rm emu}$ is associated with the systematic or intrinsic error of the model due to an imperfect emulation. The black horizontal and vertical lines demarcate contributions from the three summary statistics included in the data vector: the density-split cross-correlations and ACFs, and the galaxy 2PCF (listed in the same order as they appear along the diagonal of the correlation matrices).

In the lower sub-panels, we compare the emulator accuracy to the data errors. In this paper, we want to present a stringent test of the emulator and therefore compare its accuracy to that of the ABACUSSUMMIT simulations with a volume of $(2\,h^{-1}{\rm Gpc})^3$, which is about 8 times larger than that of the CMASS galaxy sample we are targetting (Paillas et al. 2023a). The data errors are estimated from the covariance boxes of the ABACUSSMALL simulations and are rescaled to represent the expected errors for a volume of $(2\,h^{-1}{\rm Gpc})^3$ as explained in Section 2.3.2. In Fig. 1, we show that the model prediction is mostly within 1σ of the data vector for this particular example, for both multipoles, and cross-correlations and autocorrelations.

For a quantitative assessment of the emulator accuracy in predicting multipoles over a range of cosmological parameters, we show in Fig. 3 the median absolute emulator error (taken to be the difference between the prediction and the test data), calculated across the entire test sample, in units of the data errors. The errors always lie within 2σ of the errors of the data for all scales and summary statistics, and peak at around the smoothing scale.

In Appendix B, we show a similar version of this plot where instead of rescaling the vertical axis by the errors of the data, we express everything in terms of the fractional emulator error. While the monopoles of all different density-split summary statistics are accurate within 5 per cent, and mostly well within 1 per cent on small scales, the quadrupoles tend to zero on very small scales, blowing up the fractional error.

Among all the multipoles, the error is generally larger for the monopole of the DS CCFs. This is in part due to the sub-per cent errors on the data vector below scales of $\sim 40\,h^{-1}$Mpc, but also due to the fact that the sharp transition of the CCFs below the smoothing scale is overall harder to emulate. The DS autocorrelation emulator errors are almost always within 1σ of the data errors, with the exception of the quadrupole of Q_5. In Appendix Fig. B1, we see that the emulator accuracy is at sub per cent level for the majority of the summary statistics in the analysis.

3.1.1 *Sensitivity to the different cosmological parameters*

After corroborating that the emulator is sufficiently accurate, we explore the dependency of the different summary statistics with respect to the input parameters through the use of derivatives around the fiducial *Planck* 18 cosmology (Planck Collaboration 2020).

In Fig. 4, we show the derivatives of the quantile–galaxy cross-correlations for the different density environments with respect to the cosmological parameters. In Appendix C, we show the corresponding derivatives respect with respect to the HOD parameters, together with those of the quantile autocorrelations. These are estimated by computing the gradient between the emulator's output and its input through `jax`'s autograd functionality[6] which reduces the errors that numerical derivative estimators can introduce.

In the first column of Fig. 4, we show that increasing $\omega_{\rm cdm}$ reduces the amplitude of the cross-correlations for all quantiles, possibly due to lowering the average halo bias. Increasing $\omega_{\rm cdm}$ also produces shifts in the acoustic peak on large-scales for all quantiles. Moreover, the effect on the quadrupole is to reduce its signal for the most extreme quantiles (note that the quadrupole of Q_0 is positive, whereas that of Q_4 is negative. Note that there are two different RSD effects influencing the quadrupole: on one hand, identifying the density quantiles in redshift space introduces an anisotropy in the quantile distribution, as was shown in Paillas et al. (2023b), and on the other hand, there will be an additional increase in anisotropy in the cross-correlations due to the RSD of the galaxies themselves.

Regarding σ_8, shown in the second column of Fig. 4, the effect on the monopoles is much smaller than that on the quadrupole due to enhancing velocities and therefore increasing the anisotropy caused by RSD.

Finally, the effect of n_s on the monopole is similar to that of $\omega_{\rm cdm}$, albeit without the shift at the acoustic scale. Interestingly, the

[6] https://github.com/google/jax

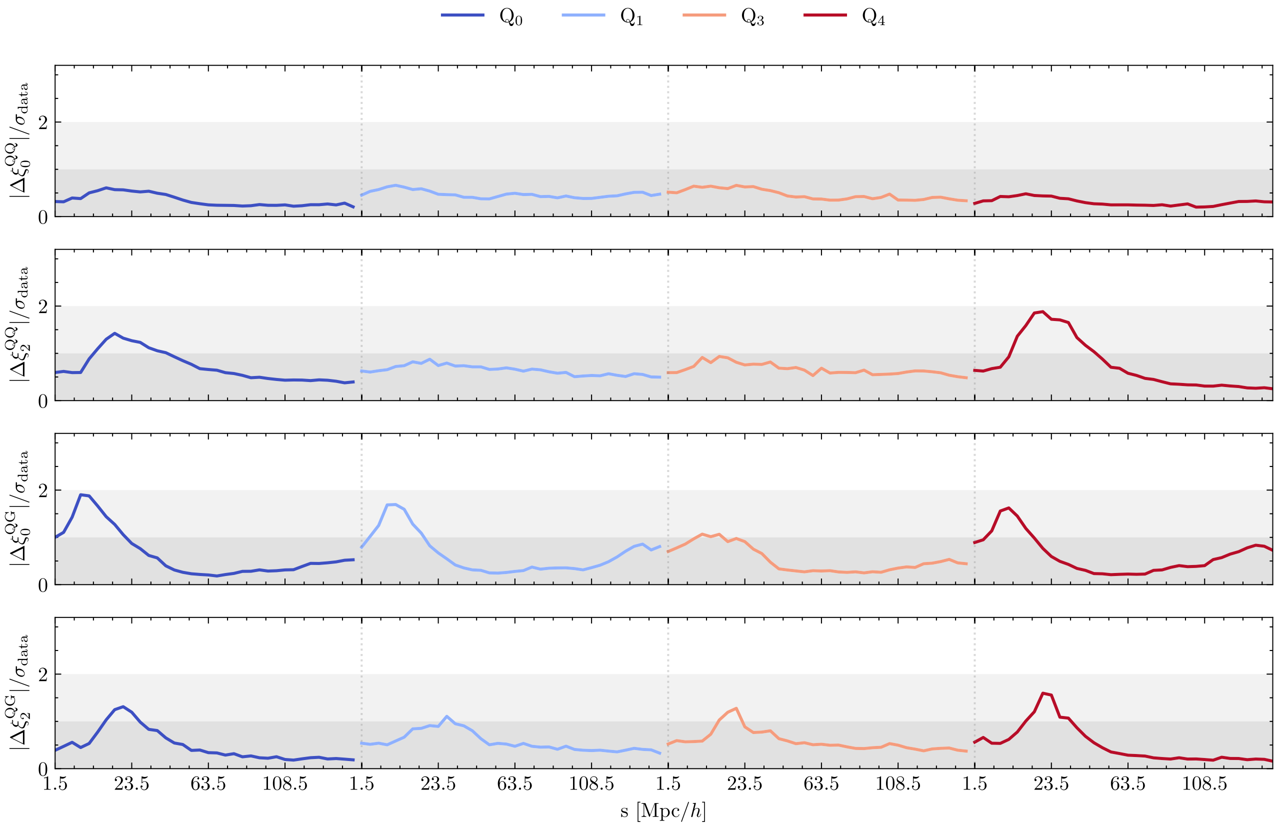

Figure 3. Median absolute emulator errors in units of the data errors, which are estimated for a volume of 2 h^{-1}Gpc. We show the monopole ACFs, quadrupole ACFs, monopole CCFs, and quadrupole CCFs in each row. The different density quantiles are shown in different colours. In Appendix B, we show that even though the emulator can be as far as 2σ away from the data for the monopole of quantile–galaxy cross-correlations, these are sub per cent errors. https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/F3_emulator_errors.py

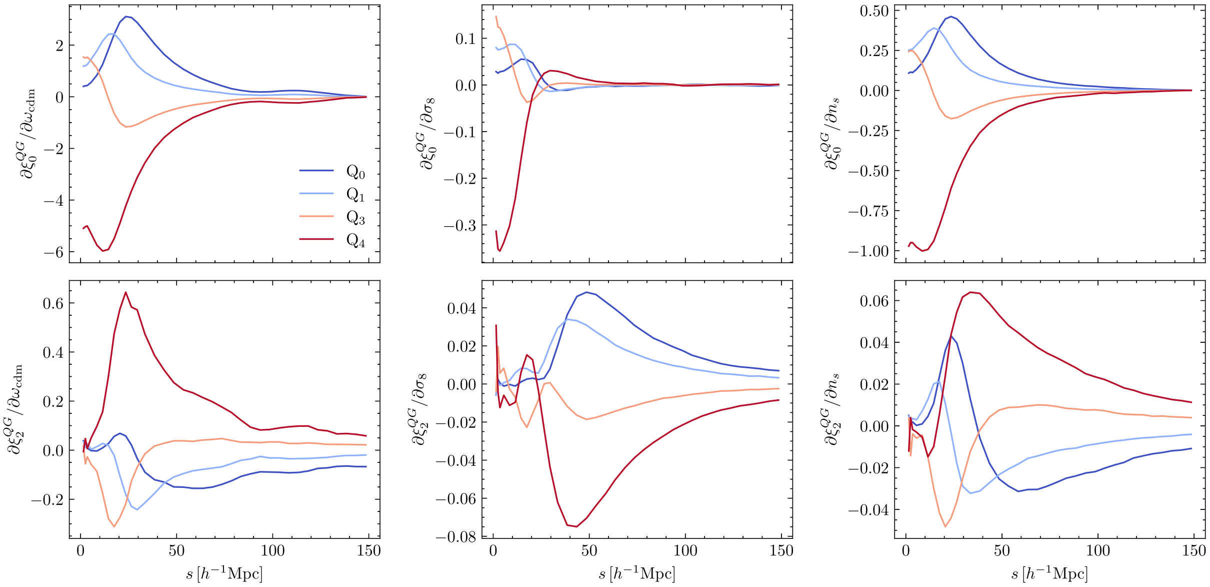

Figure 4. We show the sensitivity of the density-split statistics to each cosmological parameter by computing the derivatives of the different quantile–galaxy cross-correlations with respect to the cosmological parameters. From left to right, we show the derivatives with respect to ω_{cdm}, σ_8, and n_s, respectively. The upper panel shows the monopole derivatives, whereas the lower panel shows the derivatives of the quadrupole. We note that derivatives are estimated via automatic differentiation, as opposed to finite differences. However, these can still be noisy if the data vector itself is noisy. At small scales, the shot noise that dominates the quadrupole measurements induces noise in the estimation of the derivatives, which manifests itself as the sudden spikes on the smallest separation bins. https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/F4_derivatives.py

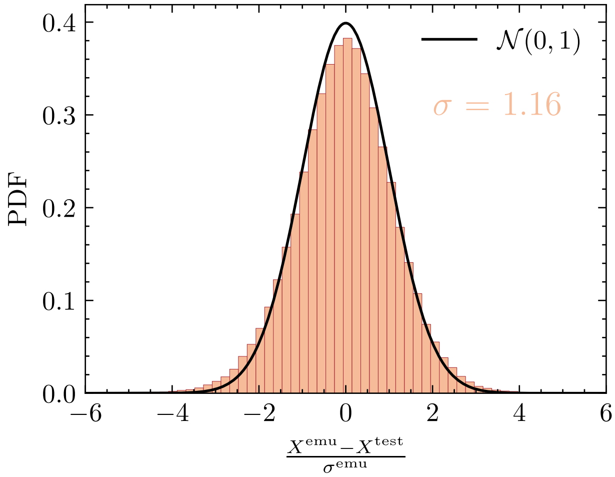

Figure 5. Z-scores of the emulator uncertainty predictions, compared to a standard normal distribution, $\mathcal{N}(0, 1)$, for the test set of the density-split cross-correlation functions. The emulator predicted uncertainty is over-confident, meaning that this predicting smaller uncertainties than those observed empirically on the test set. https://https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/F5_zscores_uncertainties.py

derivative of the quadrupole may change sign near the smoothing scale.

3.1.2 Evaluating the uncertainty estimates

While the emulator offers precise mean predictions, its uncertainty estimations present challenges. Specifically, the uncertainty estimates, $\sigma_{\rm pred}(\mathcal{C}, \mathcal{G})$, derived from training the emulator to optimize the Gaussian log-likelihood as per equation (13), tend to underestimate the true uncertainties. This underestimation is problematic as it might introduce biases in our derived cosmological parameter constraints.

To illustrate this, we present the z-score of the emulator's predictions in Fig. 5 for the monopole and quadrupole of the DS CCFs, defined as $z_k = \frac{X_k^{\rm emu} - X_k^{\rm test}}{\sigma_k^{\rm emu}}$. Given that the emulator errors are modelled as Gaussian, the emulator uncertainties would be well-calibrated if the distribution of z_k's followed a standard normal distribution. Fig. 5 shows that this is not the case, since the z-scores show a variance larger than 1 by about a 15 per cent. One possible reason for this discrepancy could be the limited size of our data set. In the remainder of the paper, we will ignore the emulator's predicted uncertainties and quantify its errors by directly estimating them from the test set instead, as described in equation (15). In the future, we aim to refine the calibration of uncertainty predictions for simulation-based models.

3.2 Solving the inverse problem: recovering the cosmological parameters

In this section, we focus on the inverse problem, that is, recovering the mocks' true cosmological parameters from their summary statistics. We will show that the emulator can recover unbiased parameter constraints on the test AbacusSummit HOD catalogues, as well as on a different N-body simulation with a galaxy–halo connection model that is based on another prescription than HOD. We also demonstrate where the density-split information comes from by varying various choices of settings in the inference analysis pipeline.

3.2.1 Recovery tests on AbacusSummit

In this section, we show the results of using the emulator to infer the combined set of cosmological and HOD parameters, a total of 17 parameters, on the test set we reserved from the AbacusSummit simulations, namely those mocks that were not used during the training of the emulator.

First, for each cosmology from the test set we select the mock catalogue with HOD parameters that maximize the likelihood with respect to a realistic data vector, taken to be the observed density-split multipoles from the BOSS CMASS galaxy sample (Paillas et al. 2023a), and infer the posterior of the cosmological and HOD parameters for that particular sample.

Since our model for the mock observables is differentiable, we can take advantage of the estimated derivatives to efficiently sample the posterior distributions through Hamiltonian Monte Carlo (HMC). HMC utilizes the gradient information from differentiable models to guide the sampling process through Hamiltonian dynamics, enabling more efficient exploration of the posterior landscape. It introduces momentum variables and a Hamiltonian function to represent the total energy, then follows the gradients to deterministically evolve the parameters over time while conserving the Hamiltonian. Here, we employ the NUTS sampler implementation from numpyro. We use flat prior ranges for the parameters that match those listed in Table 1. Fitting one mock takes about 1 min on 1 CPU.

We first fit `c000`, the baseline cosmology of AbacusSummit. Fig. 6 shows the posterior distribution of the cosmological parameters, marginalized over the HOD parameters. Density-split clustering, the galaxy 2PCF, and their combination recover unbiased constraints with the true cosmology of the simulation lying within the 68 per cent confidence region of the marginalized posterior of every parameter. Note that in particular density-split statistics contribute to breaking the strong degeneracy between n_s and $\omega_{\rm cdm}$ observed in the 2PCF. In Table 2, we show the resulting constraints for each of the three cases tested. For the $(2\,h^{-1}{\rm Gpc})^3$ volume that is considered here, the baseline analysis recovers a 2.6 per cent, 1.2 per cent, and 1.2 per cent constraint for $\omega_{\rm cdm}$, σ_8, and n_s, respectively. These constraints are a factor of about 2.9, 1.9, and 2.1 tighter than for the 2PCF, respectively. Moreover, the parameters $N_{\rm eff}$ and w_0 are recovered with a precision of 8 per cent and 4.9 per cent in the baseline analysis. These are in turn a factor of about 2.5 and 1.9 times tighter than for the 2PCF. In an idealized Fisher analysis using simulated dark matter haloes (Paillas et al. 2023b), we found similar expected improvements for all parameters but σ_8, for which the Fisher analysis predicted a much larger improvement.

The posterior distribution of the HOD parameters, marginalized over cosmology, is shown in Appendix Fig. D1. In particular, density-split statistics can contribute to significantly tightening the constraints on the environment-based assembly bias parameters, $B_{\rm cen}$ and $B_{\rm sat}$. We expect that reducing the smoothing scale used to estimate densities with future denser data sets would help us attain even tighter constraints on these parameters that may lead to significant detections of the effect in such galaxy samples. Note that for this particular sample some of the true HOD parameters are close to the prior boundary.

Moreover, in Fig. 7 we show the marginalized constraints on $\omega_{\rm cdm}$ and σ_8 for four particular cosmologies in the test set that vary these two parameters. As before, the HOD parameters are chosen from the prior for each cosmology to maximize the likelihood of CMASS data. These cosmologies are of particular interest since they show that the model can recover lower and higher σ_8 values

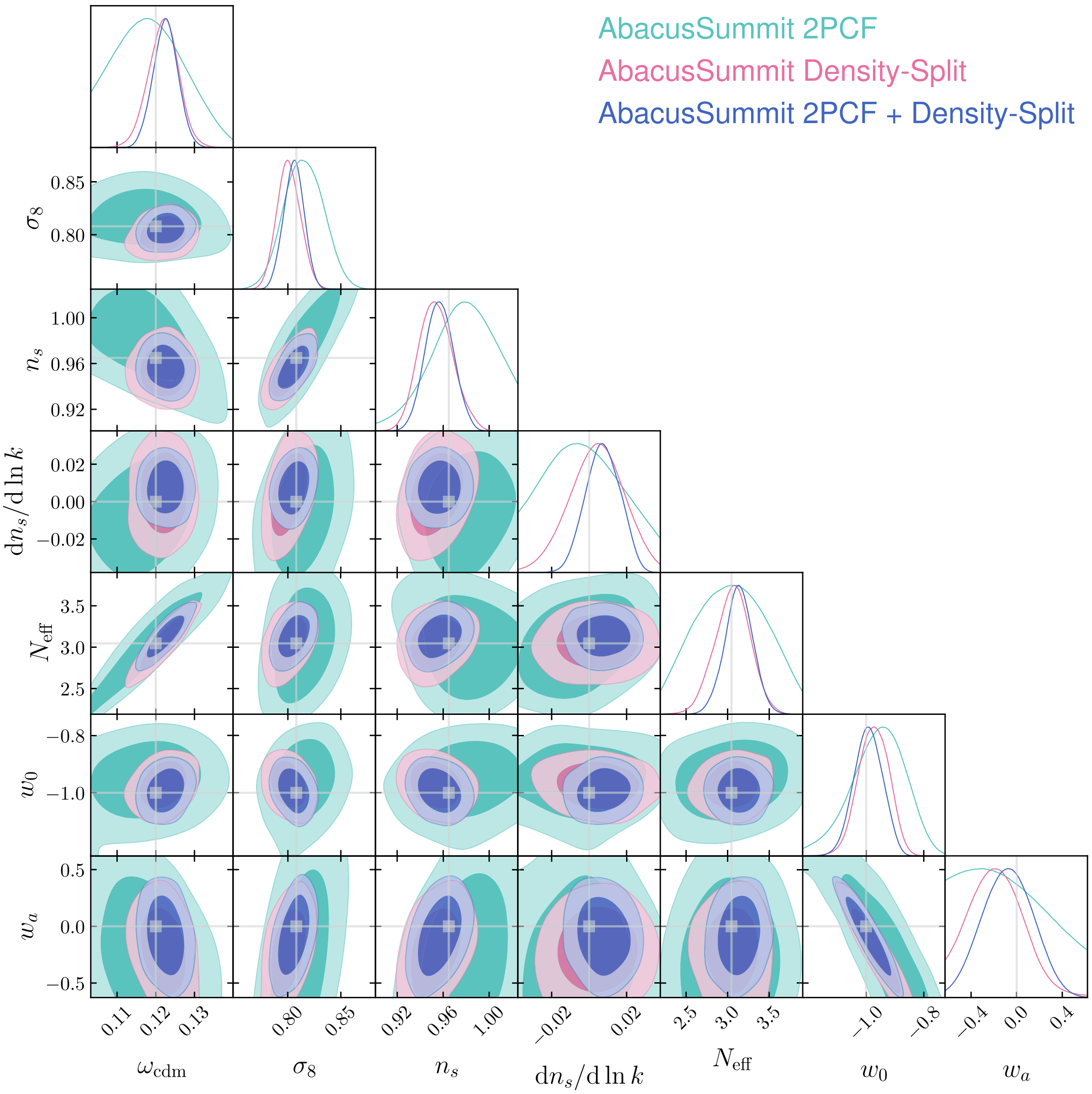

Figure 6. Recovery of AbacusSummit fiducial cosmology (c000) for the set of HOD parameters that minimize the data χ^2 error, after marginalizing over the HOD parameters. We show constraints from the 2PCF in green, Density-split statistics (Density-Split) in pink, and a combination of the two (2PCF + Density-Split) in blue. https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/F6_cosmo_inference_c0.py

than that of the fiducial *Planck* cosmology. The additional AbacusSummit cosmologies that we are analysing are, c001, based on WMAP9+ACT + SPT LCDM constraints (Calabrese et al. 2017), c003, a model with extra relativistic density ($N_{\rm eff}$) taken from the base_nnu_plikHM_TT_lowl_lowE_Riess18_post_BAO chain of (Planck Collaboration 2020) which also has both high σ_8 and $\omega_{\rm cdm}$, and c004, a model with lower amplitude clustering σ_8.

3.2.2 *Exploring the information content*

In this section, we will delve deep into the effects that removing subsets of the data when analysing the fiducial cosmology c000 have on the resulting parameter constraint to analyse what information is being used to constrain each of the parameters. The results are summarized in Fig. 8.

Let us first examine how the constraints vary as a function of the scales included in the analysis. Bear in mind however that we are not truly removing the small scales since the smoothing introduced to estimate densities leaks information from small scales into all the scales. In Fig. 8, we show first the effect of analysing only from the BAO scale, $s_{\rm min} = 80\,h^{-1}$Mpc. In that case, we still see significant gains over the full-shape 2PCF. For most parameters, however, apart from n_s, we find there is more information contained in the smaller scales.

Regarding the different quantiles, most of the information comes from the combination of void-like, Q$_0$, and cluster-like, Q$_4$, regions, whereas the intermediate quantiles barely contribute.

Moreover, we have examined the effect of removing the different error contributions on the covariance matrix. First, we show that removing the emulator error produces statistically consistent con-

Table 2. Parameter constraints from the galaxy two-point correlation (2PCF), DSC, and the baseline combination (2PCF + DSC) analyses. Each row shows the parameter name and the corresponding mean and 68 per cent confidence intervals. We note that the constraints on $\omega_{\rm b}$ are dominated by the prior used described in Table 1.

	Parameter	2PCF (68 per cent confidence interval)	DSC (68 per cent confidence interval)	2PCF + DSC (68 per cent confidence interval)
Cosmology	$\omega_{\rm b}$	–	0.02257 ± 0.00054	0.02242 ± 0.00050
	$\omega_{\rm cdm}$	$0.1187^{+0.0077}_{-0.010}$	0.1220 ± 0.0039	0.1225 ± 0.0032
	σ_8	0.815 ± 0.018	0.801 ± 0.011	0.8056 ± 0.0094
	n_s	$0.976^{+0.032}_{-0.023}$	$0.954^{+0.014}_{-0.016}$	0.957 ± 0.012
	${\rm d}n_s/{\rm d}\ln k$	$-0.003^{+0.018}_{-0.024}$	$0.004^{+0.015}_{-0.014}$	0.0074 ± 0.0090
	$N_{\rm eff}$	3.04 ± 0.40	$3.06^{+0.22}_{-0.20}$	3.13 ± 0.17
	w_0	$-0.959^{+0.10}_{-0.081}$	-0.974 ± 0.053	-0.992 ± 0.049
	w_a	<0.0662	$-0.17^{+0.22}_{-0.26}$	-0.08 ± 0.22
HOD	$\log M_1$	14.03 ± 0.15	$13.94^{+0.17}_{-0.11}$	$14.01^{+0.12}_{-0.098}$
	$\log M_{\rm cut}$	$12.588^{+0.066}_{-0.11}$	$12.621^{+0.097}_{-0.12}$	$12.581^{+0.047}_{-0.060}$
	α	$1.13^{+0.25}_{-0.19}$	$1.19^{+0.27}_{-0.11}$	$1.25^{+0.16}_{-0.11}$
	$\alpha_{\rm vel,\,c}$	$0.375^{+0.069}_{-0.054}$	$0.286^{+0.17}_{-0.089}$	$0.390^{+0.039}_{-0.033}$
	$\alpha_{\rm vel,\,s}$	>1.05	$1.08^{+0.18}_{-0.10}$	$1.09^{+0.11}_{-0.090}$
	$\log \sigma$	$-1.54^{+0.98}_{-0.56}$	$-1.61^{+0.64}_{-0.48}$	$-1.58^{+0.57}_{-0.50}$
	κ	$---$	<0.830	$0.65^{+0.22}_{-0.63}$
	$B_{\rm cen}$	<-0.404	$-0.336^{+0.059}_{-0.14}$	$-0.410^{+0.043}_{-0.060}$
	$B_{\rm sat}$	<-0.0339	-0.11 ± 0.36	-0.37 ± 0.28

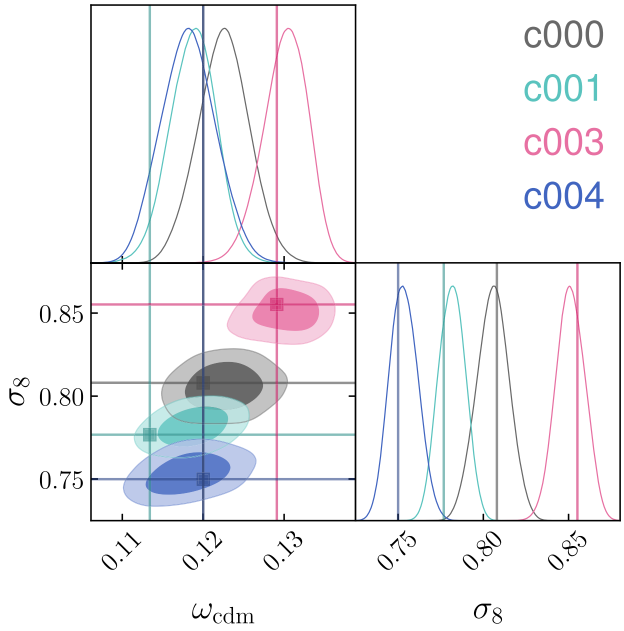

Figure 7. Marginalized constraints from DSC on $\omega_{\rm cdm}$, σ_8, and n_s, derived from fits to mock galaxy catalogues at 4 different cosmologies from our test sample. The true cosmology of each mock is shown by the horizontal and vertical-dotted coloured lines. 2D contours show the 68 and 95 per cent confidence regions around the best fit values. https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/F7_cosmo_inference_c0_c1_c3_c4.py

straints, but about a factor of 2 tighter for most parameters compared to the baseline. As we will show in the further section, our estimated uncertainties are designed to be conservative and therefore removing the emulator error does not lead in this case to extremely biased constraints. In the future, we will work on developing training sets and models that can overcome this limitation and produce more accurate predictions on small scales. This could lead to major improvements on the σ_8 constraints.

Finally, we demonstrate that cross-correlations between quantiles and galaxies (DS CCF) are on their own the most constraining statistic but there is a significant increase in constraining power obtained when combining them with auto correlations for the parameters $\omega_{\rm cdm}$, σ_8, and n_s.

3.2.3 Coverage probability test

We can test the covariance matrix and likelihood using a coverage probability test. Using repeated experiments with true values drawn from the Bayesian prior, we can test that the recovered values have the correct distribution within the likelihood using the chains sampling the posterior (Hermans et al. 2021).

In simple terms, if you have a 95 per cent confidence interval derived from the likelihood, the expected coverage is 95 per cent. That means that, theoretically, we expect that for 100 repeated trials, the true value should fall within that interval 95 times. The empirical coverage is what you actually observe when you compare the rank of the true value within the likelihood. Using the same 95 per cent confidence interval, if you applied this method to many samples and found that the true value was within the interval only 90 times out of 100, then the empirical coverage for that interval would be 90 per cent.

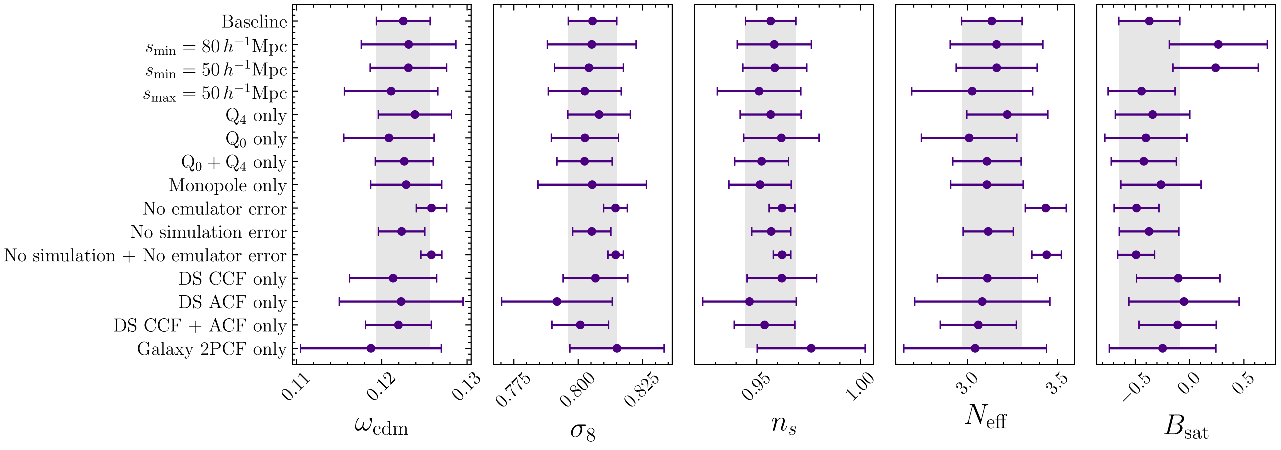

Figure 8. Marginalized constraints on $\omega_{\rm cdm}$, σ_8, n_s, $N_{\rm eff}$, and $B_{\rm sat}$ for different configurations in the inference analysis. Dots and error bars show the mean and the 68 per cent confidence interval of the parameters, respectively. The uppermost points show the baseline configuration, which consists of the combination of the monopole and quadrupole of the DS cross-correlation and autocorrelation functions for quantiles Q_0, Q_1, Q_3, and Q_4. https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/F8_whisker.py

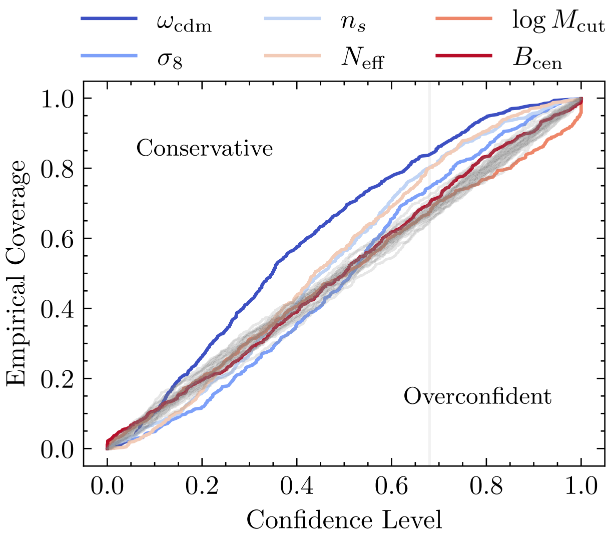

Figure 9. Comparison of the empirical coverage for a given confidence level, shown in different colours for the different cosmological parameters, to the expected coverage, shown in grey. A perfectly calibrated model would follow the one-to-one diagonal. This diagonal has some associated error bars however, given that we are only using 600 samples to estimate the coverage, we quantify this by sampling 600 points from a uniform distribution and estimating its coverage 30 times. These are the different grey lines plotted in the figure. https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/F9_posterior_coverage.py

We can use coverage to verify that our covariance estimates are indeed conservative and that we are not subsequently underestimating the uncertainties on the parameters of interest. Note that coverage is simply a measure of the accuracy of the uncertainties, and not of its information content. We estimate the empirical coverage of each parameter on the 600 test set samples of $p(\theta, X)$, extracted from six different values of the cosmological parameters and 100 different HOD values for each of them. In Fig. 9, we compare the empirical coverage to the expected one. For a perfectly well-calibrated covariance, all should match up on the diagonal line. A conservative estimator of the covariance and of the likelihood would produce curves above the diagonal, whereas overconfident error estimation would generate curves underneath the diagonal line. Fig. 9 shows that we mostly produce conservative confidence intervals from the likelihood, in particular for $\omega_{\rm cdm}$, whereas confidence intervals can be slightly overconfident for σ_8 although the deviation from the diagonal line is close to the error expected from estimating coverage on a small data set of only 600 examples. The HOD parameters are all very well-calibrated.

3.2.4 *Recovery tests on Uchuu*

One of the fundamental validation tests for our emulator is to ensure that we can recover unbiased cosmological constraints when applied to mock catalogues based on a different *N*-body simulation, and using a different galaxy–halo connection model. The latter is particularly important since the HOD model used to train the emulator makes strong assumptions about how galaxies populate dark matter haloes and its flexibility to model the data needs to be demonstrated.

To this end, we test our model on the Uchuu simulations (Ishiyama et al. 2021; Dong-Páez et al. 2024; Aung et al. 2023; Oogi et al. 2023; Prada et al. 2023) and use mock galaxies that were created by Zhai et al. (2023) using subhalo abundance matching (SHAM, e.g. Kravtsov et al. 2004; Vale & Ostriker 2006) to populate dark matter haloes with galaxies. This model assigns galaxies to dark matter haloes based on the assumption that the stellar mass or luminosity of a galaxy is correlated with the properties of dark matter halo or subhalo hosting this galaxy. Specifically, we use the method of Lehmann et al. (2017) to assign galaxies to dark matter haloes and subhaloes. In this method, the property used to rank haloes is a combination of the maximum circular velocity within the halo, $v_{\rm max}$, and the virial velocity, $v_{\rm vir}$. This model also includes a certain amount of galaxy assembly bias, further testing the flexibility of our HOD modelling.

Uchuu is a suite of cosmological *N*-body simulations that were generated with the GreeM code (Ishiyama, Fukushige & Makino 2009) at the ATERUI II supercomputer in Japan. The main simulation has a volume of $(2\,h^{-1}{\rm Gpc})^3$, following the evolution of 2.1 trillion dark matter particles with a mass resolution of $3.27 \times 10^8\,h^{-1}{\rm M}_\odot$. It is characterized by a fiducial cosmology $\Omega_{\rm m} = 0.3089$, $\Omega_{\rm b} = 0.0486$, $h = 0.6774$, $\sigma_8 = 0.8159$, and $n_s = 0.9667$. Dark matter haloes

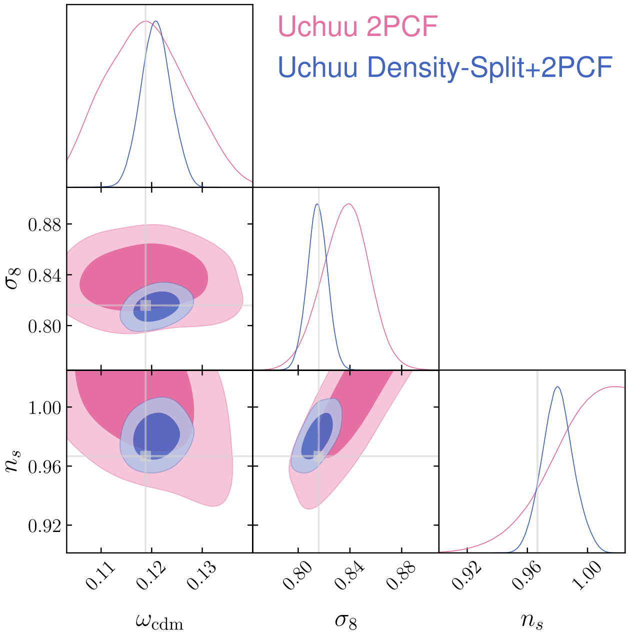

Figure 10. Marginalized posterior on the cosmological parameters when analysing the SHAM mocks based on the Uchuu simulations. We show the contours obtained when analysing only the 2PCF, compared to those found when analysing the combination of the 2PCF and density-split statistics. The true parameters that generated the mock are shown in grey. https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/F10_uchuu.py

are identified using the ROCKSTAR halo finder (Behroozi, Conroy & Wechsler 2010), which is also different from the one implemented in ABACUSSUMMIT.

Fig. 10 shows the resulting marginalized inference using our emulator for both the 2PCF, and the combination of density-split with the 2PCF. Note that the constraints on n_s from the 2PCF are in this case completely prior dominated. We can however recover unbiased constraints, even for the stringent test case of a $(2\,h^{-1}{\rm Gpc})^3$ volume.

4 DISCUSSION AND CONCLUSIONS

4.1 Comparison with previous work

4.1.1 Analytical models of density dependent statistics

Similar definitions of density-split statistics have been presented in Neyrinck et al. (2018); Repp & Szapudi (2021). In Neyrinck et al. (2018), the authors defined sliced correlation functions, by slicing the correlation function on local density. They have also presented a model with the Gaussian assumption. In Repp & Szapudi (2021), the authors introduce indicator functions by identifying regions of a given density and computing the power spectrum in density bins. This is essentially the Fourier version of the DS ACF. Our analyses have included both the DS CCF and ACF, but finding that the CCF carries most of the cosmological information. These statistics are all similar in spirit.

4.1.2 Fisher

In previous work, Paillas et al. (2023b) showed with a Fisher analysis the potential of density-split statistics to constrain cosmological parameters from dark matter halo statistics. Here, we have confirmed their findings by modelling the density-split statistics explicitly as a function of the cosmological parameters, and including the halo–galaxy connection to model the density-split statistics of galaxies.

The improved constraints over 2PCFs found here are of a similar magnitude to those in Paillas et al. (2023b) for all cosmological parameters, but σ_8, for which we find weaker constraints. Moreover, we also find that the most extreme quantiles have a similar constraining power and that it is their combination that explains most of the information content of density-split statistics. Finally, Paillas et al. (2023b) found that density-split statistics could break important degeneracies between cosmological parameters that would lead to much tighter constraints on the sum of neutrino masses. This is not something we could corroborate in this paper since variations in neutrino mass are not included in the suite of simulations used in this work, but we plan to work on this in the future by utilizing N-body simulations that can accurately simulate the effects of massive neutrinos in the large-scale structure (Elbers et al. 2021).

4.1.3 Cosmic voids

Over the past decade, there has been renewed interest in using cosmic voids to constrain cosmology (Pisani et al. 2019). They have been found to be amongst the most pristine probes of cosmology in terms of how much information is preserved in linear theory at late times (Cai et al. 2016; Hamaus et al. 2017; Nadathur & Percival 2019). However, in practice, extracting cosmological information from voids has proven to be difficult from a purely perturbation theory perspective mainly due to (i) void definitions being difficult to treat analytically and producing different results (Cautun et al. 2018), (ii) identifying voids in redshift space adds additional anisotropy to the observed void–galaxy cross-correlation (Nadathur, Carter & Percival 2019; Correa et al. 2020), a similar effect to that found here when estimating densities directly in redshift space, which is difficult to model analytically, and (iii) linear theory can only accurately model the mapping from real to redshift space, which means we still require some way of estimating the real space void profiles. In this work, we have shown how emulators can fix all of the above-mentioned issues by forward modelling each of these effects.

Moreover, we have shown here how although void–galaxy cross-correlations contain a wealth of information to constrain the cosmological parameters, it is their combination with overdense environments that would give us the tightest constraints.

4.1.4 The AEMULUS emulator

Related to this work, Storey-Fisher et al. (2024) presented an emulation framework based on the AEMULUS suite of cosmological N-body simulations to accurately reproduce two-point galaxy clustering, the underdensity probability function and the density-marked correlation function on small scales ($0.1 - 50\,h^{-1}$Mpc). We confirm that including summary statistics beyond two-point functions can improve the cosmological constraints significantly, even after marginalizing over the HOD parameters. Moreover, environment-based statistics could lead to a significant detection of assembly bias. As opposed to the marked correlations shown in Storey-Fisher et al. (2024), we estimate densities around random points spread over the survey volume which better samples underdensities in the cosmic web. In the future, it would be interesting to compare the density-split constraints to those of density-marked correlation functions, and perhaps the findings in this paper on what environments are most constraining can inform the shape of the mark function used.

4.2 Conclusions

We have presented a new simulation-based model for DSC and the galaxy 2PCF, based on mock galaxy catalogues from the ABACUSSUMMIT suite of simulations. These models allow us to extract information from the full-shape of the correlation functions on a very broad scale range, $1\,h^{-1}$Mpc $< s < 150\,h^{-1}$Mpc, including redshift-space and AP distortions to constrain cosmology and deepen our understanding of how galaxies connect to their host dark matter haloes. We have trained neural network surrogate models, or emulators, which can generate accurate theory predictions for DSC and the galaxy 2PCF in a fraction of a second within an extended $w\Lambda$CDM parameter space.

The galaxy–halo connection is modelled through an extended HOD framework, including a parametrization for velocity bias and environment-based assembly bias, but the emulator is also validated against simulations that use a SHAM framework and a different *N*-body code to demonstrate the robustness of the method. We have shown that density-split statistics can extract information from the non-Gaussian density field that is averaged out in the galaxy 2PCF.

Our emulators, which reach a sub-per cent level accuracy down to $1\,h^{-1}$Mpc, are able to recover unbiased cosmological constraints when fitted against measurements from simulations of a $(2\,h^{-1}\mathrm{Gpc})^3$ volume. The recovered parameter constraints are robust against choices in the HOD parametrization and scale cuts, and show consistency between the different clustering summary statistics.

We find that density-split statistics can increase the constraining power of galaxy 2PCFs by factors of 2.9, 1.9, and 2.1 on the cosmological parameters ω_{cdm}, σ_8, and n_s, respectively. Moreover, the precision on parameters N_{ur}, and w_0 can be improved by factors of 2.5 and 1.9 with respect to the galaxy 2PCF. Finally, we find density-split statistics to be particularly constraining the environment-based assembly bias parameters. In a companion paper, we show how all these findings result on parameter constraints from the CMASS sample of SDSS (Paillas et al. 2023a).

As we transition to the era of DESI, with its high-density galaxy samples, particularly BGS, alternative summary statistics such as density-split have a huge potential to not only increase the precision on cosmological parameter constraints, but to deepen our understanding of how galaxies connect to dark matter haloes. However, this opportunity comes with challenges. The precision that DESI promises requires that our theoretical frameworks are refined to an unprecedented degree of accuracy. It is essential to address these theoretical challenges to fully harness the potential of upcoming observational data sets in cosmological studies.

ACKNOWLEDGEMENTS

The authors thank Etienne Burtin for helpful discussions throughout the development of this project. YC acknowledges the support of the Royal Society through a University Research Fellowship. SN acknowledges support from an STFC Ernest Rutherford Fellowship, grant reference ST/T005009/2. FB is a University Research Fellow and has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 853291). WP acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference no. RGPIN-2019-03908] and from the Canadian Space Agency.

This work is supported by the National Science Foundation under Cooperative Agreement PHY- 2 019 786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/). This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics of U.S. Department of Energy under grant contract numbers DE-SC0012567, DE-SC0013718, and under DE-AC02-76SF00515 to SLAC National Accelerator Laboratory, and by the Kavli Institute for Particle Astrophysics and Cosmology. The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University, and on the Narval cluster provided by Compute Ontario (computeontario.ca) and the Digital Research Alliance of Canada (alliancecan.ca). In addition, this work used resources of the National Energy Research Scientific Computing Centre (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under contract number DE-AC02-05CH11231.

The ABACUSSUMMIT simulations were run at the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under contract number DE-AC05-00OR22725.

We thank Instituto de Astrofisica de Andalucia (IAA-CSIC), Centro de Supercomputacion de Galicia (CESGA), and the Spanish academic and research network (RedIRIS) in Spain for hosting Uchuu DR1, DR2, and DR3 in the Skies & Universes site for cosmological simulations. The Uchuu simulations were carried out on Aterui II supercomputer at Centre for Computational Astrophysics, CfCA, of National Astronomical Observatory of Japan, and the K computer at the RIKEN Advanced Institute for Computational Science. The Uchuu Data Releases efforts have made use of the skun@IAA_RedIRIS and skun6@IAA computer facilities managed by the IAA-CSIC in Spain (MICINN EU-Feder grant no. EQC2018-004366-P).

This research used the following software packages: CORRFUNC (Sinha & Garrison 2020), FLAX (Heek et al. 2023), GETDIST (Lewis 2019), JAX (Bradbury et al. 2018), JUPYTER (Kluyver et al. 2016), MATPLOTLIB (Hunter 2007), NUMPY (Harris et al. 2020), NUMPYRO (Phan, Pradhan & Jankowiak 2019), and OPTUNA.

5 DATA AVAILABILITY STATEMENT

The data underlying this article are available in https://abacusnbody.org.

REFERENCES

Anderson L. et al., 2014, MNRAS, 441, 24
Aung H. et al., 2023, MNRAS, 519, 1648
Banerjee A., Abel T., 2020, MNRAS, 500, 5479
Behroozi P. S., Conroy C., Wechsler R. H., 2010, ApJ, 717, L379
Bonnaire T., Aghanim N., Kuruvilla J., Decelle A., 2022, A&A, 661, 146
Bose S., Eisenstein D. J., Hernquist L., Pillepich A., Nelson D., Marinacci F., Springel V., Vogelsberger M., 2019, MNRAS, 490, 5693
Bradbury J. et al., 2018, JAX: composable transformations of Python + NumPy programs, available at: http://github.com/google/jax
Cai Y.-C., Taylor A., Peacock J. A., Padilla N., 2016, MNRAS, 462, 2465
Calabrese E. et al., 2017, Phys. Rev. D, 95, 063525

Cautun M., Paillas E., Cai Y.-C., Bose S., Armijo J., Li B., Padilla N., 2018, MNRAS, 476, 3195

Chiang C.-T., Wagner C., Sánchez A. G., Schmidt F., Komatsu E., 2015, J. Cosmol. Astropart. Phys., 2015, 028

Cole S. et al., 2005, MNRAS, 362, 505

Correa C. M., Paz D. J., Sánchez A. G., Ruiz A. N., Padilla N. D., Angulo R. E., 2020, MNRAS, 500, 911

Cranmer K., Brehmer J., Louppe G., 2020, PNAS, 117, 30055

DESI Collaboration, 2016, preprint (arXiv:1611.00036)

Dai B., Seljak U., 2022, MNRAS, 516, 2363

Dai B., Seljak U., 2024, PNAS, 121, e2309624121

Dawson K. S. et al., 2016, AJ, 151, 44

DeRose J. et al., 2019, ApJ, 875, L69

Dong-Páez C. A. et al., 2024, MNRAS, 528, 7236

Einasto J., Klypin A., Hütsi G., Liivamägi L.-J., Einasto M., 2021, A&A, 652, 94

Eisenstein D. J. et al., 2005, ApJ, 633, L560

Elbers W., Frenk C. S., Jenkins A., Li B., Pascoli S., 2021, MNRAS, 507, 2614

Elfwing S., Uchibe E., Doya K., 2018, Neural Networks, 107, 3

Friedrich O. et al., 2021, MNRAS, 508, 3125

Garrison L. H., Eisenstein D. J., Pinto P. A., 2019, MNRAS, 485, 3370

Garrison L. H., Eisenstein D. J., Ferrer D., Maksimova N. A., Pinto P. A., 2021, MNRAS, 508, 575

Gil-Marín H., Percival W. J., Verde L., Brownstein J. R., Chuang C.-H., Kitaura F.-S., Rodríguez-Torres S. A., Olmstead M. D., 2017, MNRAS, 465, 1757

Green J. et al., 2012, preprint (arXiv:1208.4012)

Gualdi D., Gil-Marín H., Verde L., 2021, J. Cosmol. Astropart. Phys., 2021, 008

Guo H. et al., 2014, MNRAS, 446, 578

Hadzhiyska B., Bose S., Eisenstein D., Hernquist L., Spergel D. N., 2020, MNRAS, 493, 5506

Hadzhiyska B., Eisenstein D., Bose S., Garrison L. H., Maksimova N., 2021, MNRAS, 509, 501

Hahn C. et al., 2023, PNAS, 120, e2218810120

Hamaus N., Cousinou M.-C., Pisani A., Aubert M., Escoffier S., Weller J., 2017, J. Cosmol. Astropart. Phys., 2017, 014

Harris C. R. et al., 2020, Nature, 585, 357

Hawken A. J., Aubert M., Pisani A., Cousinou M.-C., Escoffier S., Nadathur S., Rossi G., Schneider D. P., 2020, J. Cosmol. Astropart. Phys., 2020, 012

Heek J., Levskaya A., Oliver A., Ritter M., Rondepierre B., Steiner A., van Zee M., 2023, Flax: A Neural Network Library and Ecosystem for JAX, available at: http://github.com/google/flax

Heitmann K., Higdon D., White M., Habib S., Williams B. J., Lawrence E., Wagner C., 2009, ApJ, 705, L156

Hermans J., Delaunoy A., Rozet F., Wehenkel A., Begy V., Louppe G., 2021, preprint (arXiv:2110.06581)

Hou J., Bautista J., Berti M., Cuesta-Lazaro C., Hernández-Aguayo C., Tröster T., Zheng J., 2023, Universe, 9, 302

Howlett C., Percival W. J., 2017, MNRAS, 472, 4935

Hunter J. D., 2007, Comput. Sci. Eng., 9, 90

Ishiyama T., Fukushige T., Makino J., 2009, PASJ, 61, 1319

Ishiyama T. et al., 2021, MNRAS, 506, 4210

Jamieson D., Loverde M., 2020, Phys. Rev. D, 102, 123546

Jeffrey N., Alsing J., Lanusse F., 2020, MNRAS, 501, 954

Joyce A., Jain B., Khoury J., Trodden M., 2015, Phys. Rep., 568, 1

Kluyver T. et al., 2016, in Loizides F., Schmidt B.eds, Positioning and Power in Academic Publishing: Players, Agents and Agendas. p. 87

Klypin A., Prada F., Betancort-Rijo J., Albareti F. D., 2018, MNRAS, 481, 4588

Kodwani D., Alsono D., Ferreira P., 2019, The Open J. Astrophys., 2

Komatsu E. et al., 2011, ApJS, 192, 18

Kravtsov A. V., Berlind A. A., Wechsler R. H., Klypin A. A., Gottlöber S., Allgood B., Primack J. R., 2004, ApJ, 609, L35

Laureijs R., Amiaux J., Arduini S., Augueres J. L., Brinchmann J., Cole R. et al., 2011, preprint (arXiv:1110.3193)

Lavaux G., Jasche J., Leclercq F., 2019, preprint (arXiv:1909.06396)

Lehmann B. V., Mao Y.-Y., Becker M. R., Skillman S. W., Wechsler R. H., 2017, ApJ, 834, L37

Levi M. et al., 2019, BAAS, 51, 57

Lewis A., 2019, preprint (arXiv:1910.13970)

Maddox S. J., Efstathiou G., Sutherland W. J., Loveday J., 1990, MNRAS, 242, 43

Maksimova N. A., Garrison L. H., Eisenstein D. J., Hadzhiyska B., Bose S., Satterthwaite T. P., 2021, MNRAS, 508, 4017

Massara E., Sheth R. K., 2018, preprint (arXiv:1811.03132)

Massara E. et al., 2023, ApJ, 951, 70

Nadathur S., Percival W. J., 2019, MNRAS, 483, 3472

Nadathur S., Carter P., Percival W. J., 2019, MNRAS, 482, 2459

Nadathur S. et al., 2020, MNRAS, 499, 4140

Neyrinck M. C., 2011, ApJ, 742, L91

Neyrinck M. C., Szapudi I., Szalay A. S., 2009, ApJ, 698, L90

Neyrinck M. C., Szapudi I., McCullagh N., Szalay A. S., Falck B., Wang J., 2018, MNRAS, 478, 2495

Nishimichi T. et al., 2019, ApJ, 884, L29

Oogi T. et al., 2023, MNRAS, 525, 3879

Paillas E., Cai Y.-C., Padilla N., Sánchez A. G., 2021, MNRAS, 505, 5731

Paillas E. et al., 2023a, preprint (arXiv:2309.16541)

Paillas E. et al., 2023b, MNRAS, 522, 606

Percival W. J. et al., 2001, MNRAS, 327, 1297

Phan D., Pradhan N., Jankowiak M., 2019, preprint (arXiv:1912.11554)

Philcox O. H. E., Ivanov M. M., 2022, Phys. Rev. D, 105, 043517

Philcox O. H. E., Hou J., Slepian Z., 2021, preprint (arXiv:2108.01670)

Pisani A. et al., 2019, preprint (arXiv:1903.05161)

Planck Collaboration, 2020, A&A, 641, 6

Prada F., Behroozi P., Ishiyama T., Klypin A., Pérez E., 2023, preprint (arXiv:2304.11911)

Reid B. et al., 2016, MNRAS, 455, 1553

Repp A., Szapudi I., 2021, MNRAS, 509, 586

Schmidt F., 2021, J. Cosmol. Astropart. Phys., 2021, 032

Sinha M., Garrison L. H., 2020, MNRAS, 491, 3022

Slepian Z., Eisenstein D. J., 2017, MNRAS, 469, 2059

Smith A., de Mattia A., Burtin E., Chuang C.-H., Zhao C., 2020, MNRAS, 500, 259

Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R., 2014, J. Mach. Learn. Res., 15, 1929

Storey-Fisher K., Tinker J., Zhai Z., DeRose J., Wechsler R. H., Banerjee A., 2024, ApJ, 961, 208

Sugiyama N. S., Saito S., Beutler F., Seo H.-J., 2019, MNRAS, 484, 364

Szapudi I., Pan J., 2004, ApJ, 602, L26

Uhlemann C., Friedrich O., Villaescusa-Navarro F., Banerjee A., Codis S. r., 2020, MNRAS, 495, 4006

Vale A., Ostriker J. P., 2006, MNRAS, 371, 1173

Valogiannis G., Dvorkin C., 2022a, Phys. Rev. D, 105, 103534

Valogiannis G., Dvorkin C., 2022b, Phys. Rev. D, 106, 103509

Wang X., Neyrinck M., Szapudi I., Szalay A., Chen X., Lesgourgues J., Riotto A., Sloth M., 2011, ApJ, 735, L32

Wang Y. et al., 2022, Commun. Phys., 7, 130

Woodfinden A., Nadathur S., Percival W. J., Radinović S., Massara E., Winther H. A., 2022, MNRAS, 516, 4307

Xu X., Zehavi I., Contreras S., 2021, MNRAS, 502, 3242

Yuan S., Garrison L. H., Hadzhiyska B., Bose S., Eisenstein D. J., 2021, MNRAS, 510, 3301

Yuan S., Garrison L. H., Eisenstein D. J., Wechsler R. H., 2022, MNRAS, 515, 871

Yuan S., Hadzhiyska B., Abel T., 2023a, MNRAS, 520, 6283

Yuan S., Zamora A., Abel T., 2023b, MNRAS, 522, 3935

Zhai Z. et al., 2023, ApJ, 948, L99

Zheng Z., Coil A. L., Zehavi I., 2007, ApJ, 667, L760

APPENDIX A: GAUSSIANITY LIKELIHOOD

In this section, we check that the likelihood of DS statistics is distributed as multivariate Gaussian, following the analysis in Friedrich et al. (2021). We first compute the χ^2 value of the summary statistic measured in each of the fiducial simulations

$$\chi^2_i = \left(\boldsymbol{d}_i(\mathbf{s}) - \bar{\boldsymbol{d}}(\mathbf{s})\right)^\top C^{-1} \left(\boldsymbol{d}_i(\mathbf{s}) - \bar{\boldsymbol{d}}(\mathbf{s})\right), \tag{A1}$$

where $\boldsymbol{d}_i$ represents the value of the summary statistic for the i-th fiducial simulation evaluated at the pair separation vector $\mathbf{s}$, $\bar{\boldsymbol{d}}(\mathbf{s})$ is the average of the summary statistic over all fiducial simulations at the pair separation vector $\mathbf{s}$, and C is the covariance matrix estimated from all the fiducial simulations.

If the likelihood of the summary statistic is Gaussian-distributed, the χ^2_i values should also follow a χ^2 distribution with degrees of freedom determined by the number of pair-separation bins.

Furthermore, if the likelihood is Gaussian, the distribution of χ^2_i should also be very close to that of sampling from a multivariate Gaussian with a mean given by $\bar{d}$ and the covariance measured from the simulations.

In Fig. A1, we show how the 2PCF and DS statistics χ^2_i calculated from the AbacusSmall data follow a very similar χ^2 distribution as that of the random samples generated from a multivariate Gaussian.

APPENDIX B: FRACTIONAL ERRORS

In this appendix, we present the emulator median fractional errors for the different multipoles of each statistic, measured on the test set simulations. Fig. B1 shows that the monopoles of all summary statistics are predicted well within 1 per cent for all statistics apart from Q_4 cross-correlations, where the errors get closer to 5 per cent. Regarding the quadrupole, the fractional errors blow up to do the quadrupole approaching zero on small scales. However, the eror on large-scales is well within 5 per cent.

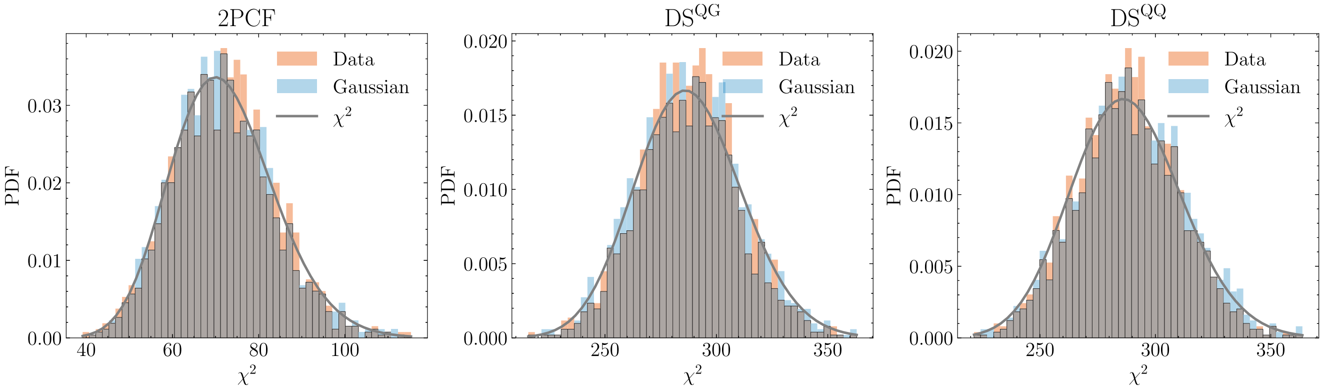

Figure A1. A qualitative assessment of the Gaussianity of the likelihoods for the 2PCF (left), DS galaxy cross-correlations (middle), and DS autocorrelations (right). The coloured histograms show the distribution of χ^2 values, as measured from the AbacusSmall simulations (orange) and a multivariate Gaussian distribution with the same mean and covariance as the simulations (blue). The solid line shows a theoretical χ^2 distribution with degrees of freedom set to the number of pair separation bins. https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/A1_gaussian_likelihood.py

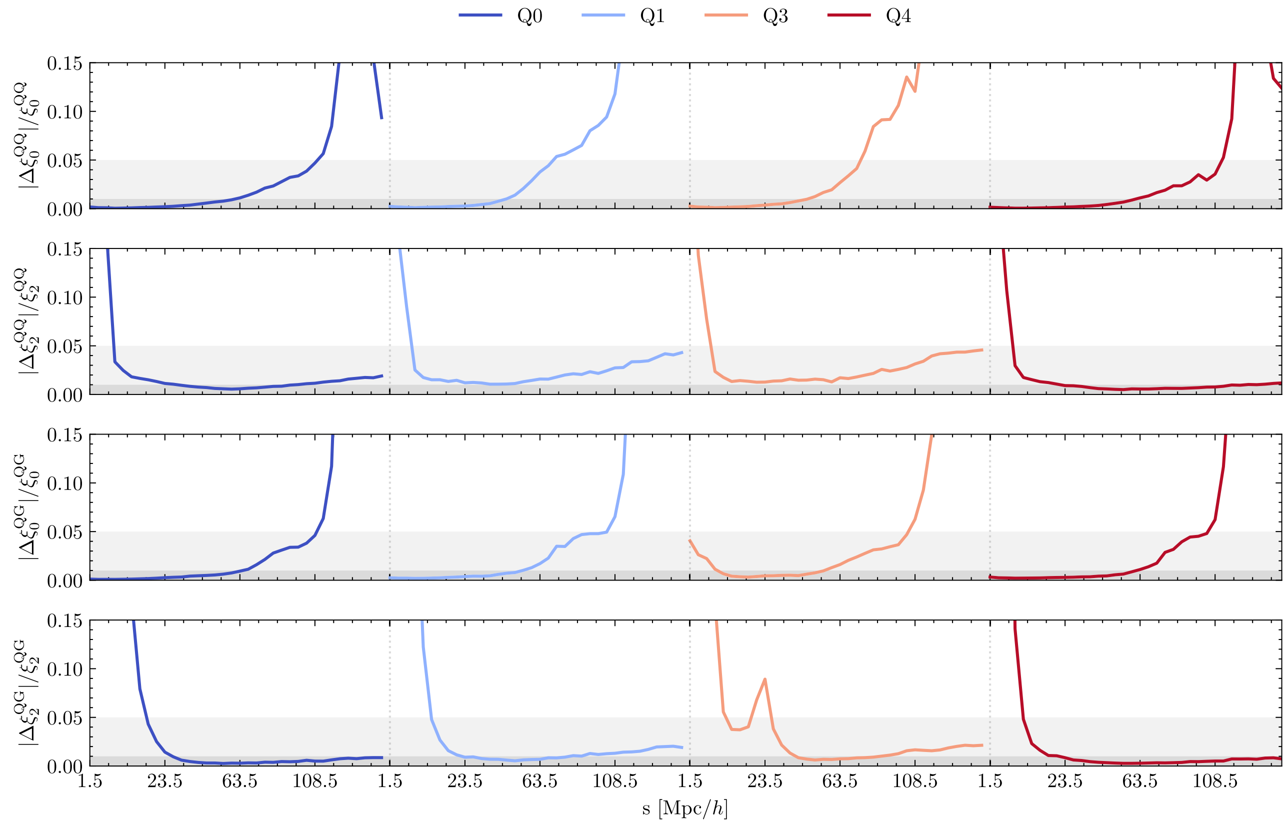

Figure B1. Median absolute fractional errors of the emulator. We show the monopole ACFs, quadrupole ACFs, monopole CCFs, and quadrupole CCFs in each row, estimated from the test set simulations with varying cosmologies and HOD parameters. The different density quantiles are shown in different colours. https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/F3_emulator_errors.py

APPENDIX C: EMULATOR DERIVATIVES WITH RESPECT TO THE COSMOLOGICAL PARAMETERS

In this section, we showcase the cosmological dependence of the different summary statistics by computing the derivative of the statistic with respect to the different cosmological and HOD parameters.

In particular, we show the DS CCFs derivatives with respect to different HOD parameters in Fig. C1. As expected, the impact of the HOD parameters is stronger on small scales.

In Fig. C2, we show the derivatives of the DS ACFs with respect to the cosmological parameters. As seen on the first panel, changes in $\omega_{\rm cdm}$ shift the BAO position of the different density quantiles. Moreover, Fig. C3 shows the derivatives of the same statistic with respect to the HOD parameters.

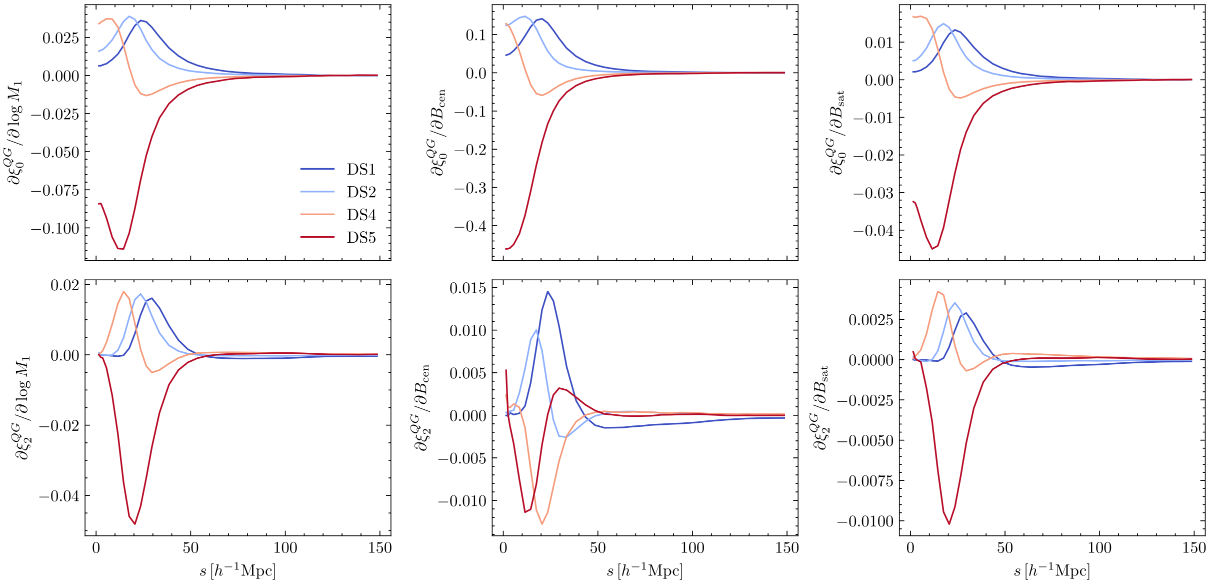

Figure C1. Derivatives of the different quantile–galaxy cross-correlations with respect to the HOD parameters. From left to right, we show the derivatives with respect to $\log M_1$, $B_{\rm cen}$ and $B_{\rm sat}$, respectively. The upper panel shows the monopole derivatives, whereas the lower panel shows the derivatives of the quadrupole. https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/F4_derivatives.py

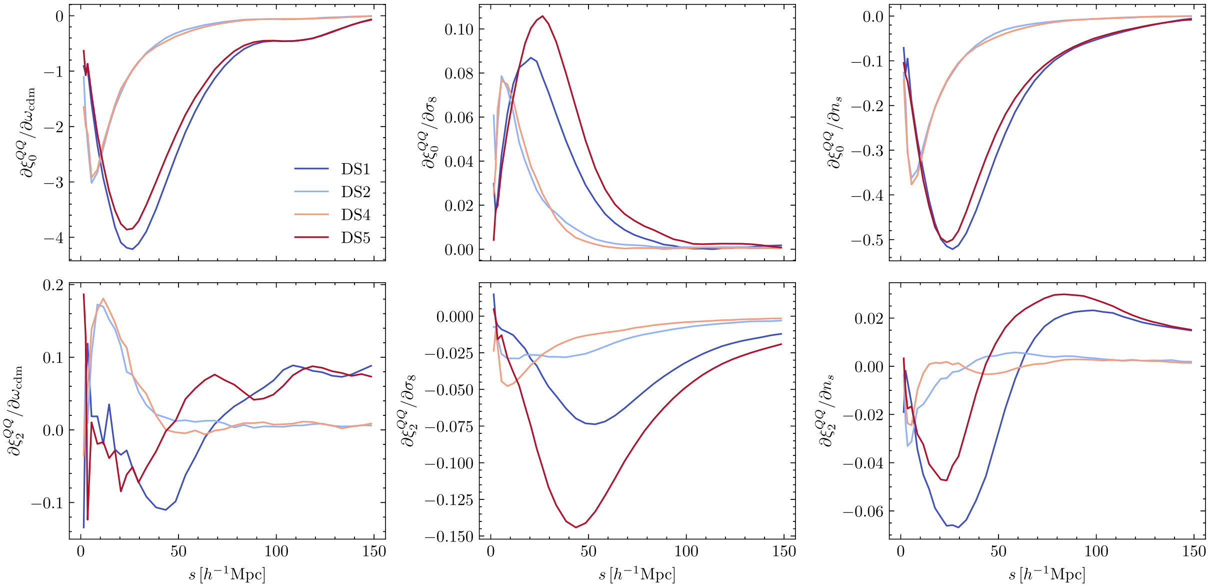

Figure C2. Derivatives of the different density-split autocorrelations with respect to the cosmological parameters. From left to right, we show the derivatives with respect to $\omega_{\rm cdm}$, σ_8, and n_s, respectively. The upper panel shows the monopole derivatives, whereas the lower panel shows the derivatives of the quadrupole. https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/F4_derivatives.py

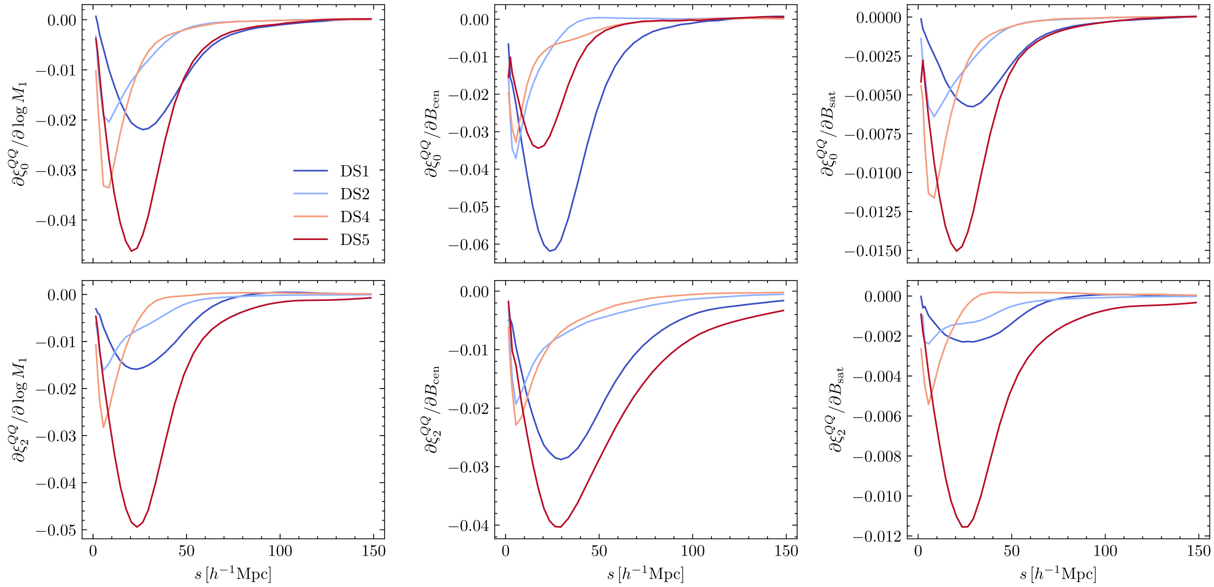

Figure C3. Derivatives of the different density-split autocorrelations with respect to the HOD parameters. From left to right, we show the derivatives with respect to log M_1, $B_{\rm cen}$, and $B_{\rm sat}$, respectively. The upper panel shows the monopole derivatives, whereas the lower panel shows the derivatives of the quadrupole. https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/F4_derivatives.py

APPENDIX D: CONSTRAINTS ON THE HOD PARAMETERS

Finally, in Fig. D1, we present the constraints on the HOD parameters obtained by the different summary statistics after marginalizing over cosmology, for the ABACUSSUMMIT fiducial cosmology. We demonstrate that the combination of 2PCF and density-split does indeed recover unbiased constraints.

Although density-split does not provide stringent constraints on those parameters that constrain the occupation of satellites (as expected, due to the choice of a large smoothing scale), it can constrain the environment-based assembly bias parameters very accurately in combination with the galaxy 2PCF.

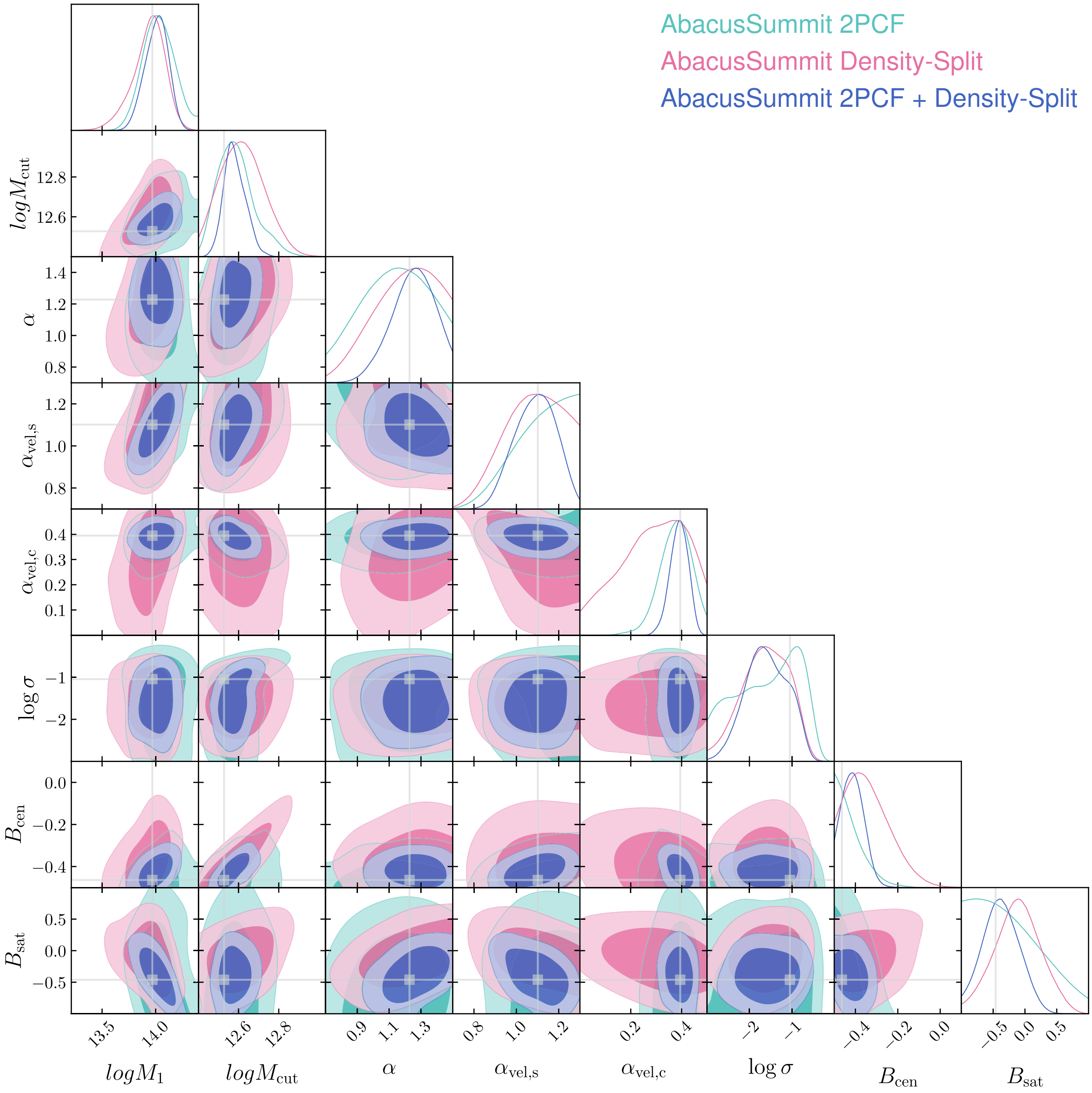

Figure D1. Recovery of the ABACUSSUMMIT fiducial cosmology for the set of HOD parameters that minimize the data χ^2 error, after marginalizing over the cosmological parameters. In green, we show the posterior distribution found when only using the galaxy 2PCF. In pink, we show those found with density-split statistics (CCFs and ACFs). In blue, we show the combination of density-split statistics and the 2PCF. https://github.com/florpi/sunbird/blob/main/paper_figures/emulator_paper/F6_cosmo_inference_c0.py

This paper has been typeset from a TEX/LATEX file prepared by the author.