Galaxy clustering from the bottom up: A Streaming Model emulator I

Carolina Cuesta-Lazaro[1,2,⋆] Takahiro Nishimichi[3,4], Yosuke Kobayashi[5], Cheng-Zong Ruan[1], Alexander Eggemeier[8] †, Hironao Miyatake[6,4], Masahiro Takada[4], Naoki Yoshida[7,4], Pauline Zarrouk[9], Carlton M. Baugh,[1,2], Sownak Bose[1] and Baojiu Li[1]

[1] *Institute for Computational Cosmology, Department of Physics, Durham University, South Road, Durham DH1 3LE, UK*
[2] *Institute for Data Science, Durham University, South Road, Durham DH1 3LE, UK*
[3] *Center for Gravitational Physics, Yukawa Institute for Theoretical Physics, Kyoto University, Kyoto 606-8502, Japan*
[4] *Kavli Institute for the Physics and Mathematics of the Universe (WPI), The University of Tokyo Institutes for Advanced Study (UTIAS), The University of Tokyo, Chiba 277-8583, Japan*
[5] *Department of Astronomy/Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721-0065, USA*
[6] *Kobayashi-Maskawa Institute for the Origin of Particles and the Universe (KMI), Nagoya University, Nagoya, 464-8602, Japan*
[7] *Department of Physics, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-0033, Japan*
[8] *Argelander Institut für Astronomie der Universität Bonn, Auf dem Hügel 71, 53121 Bonn, Germany*
[9] *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 XXX. Received YYY; in original form ZZZ

ABSTRACT

In this series of papers, we present a simulation-based model for the non-linear clustering of galaxies based on separate modelling of clustering in real space and velocity statistics. In the first paper, we present an emulator for the real-space correlation function of galaxies, whereas the emulator of the real-to-redshift space mapping based on velocity statistics is presented in the second paper. Here, we show that a neural network emulator for real-space galaxy clustering trained on data extracted from the Dark Quest suite of N-body simulations achieves sub-per cent accuracies on scales $1 < r < 30\ h^{-1}$ Mpc, and better than 3% on scales $r < 1\ h^{-1}$Mpc in predicting the clustering of dark-matter haloes with number density $10^{-3.5}\ (h^{-1}\mathrm{Mpc})^{-3}$, close to that of SDSS LOWZ-like galaxies. The halo emulator can be combined with a galaxy-halo connection model to predict the galaxy correlation function through the halo model. We demonstrate that we accurately recover the cosmological and galaxy-halo connection parameters when galaxy clustering depends only on the mass of the galaxies' host halos. Furthermore, the constraining power in σ_8 increases by about a factor of 2 when including scales smaller than $5\ h^{-1}$ Mpc. However, when mass is not the only property responsible for galaxy clustering, as observed in hydrodynamical or semi-analytic models of galaxy formation, our emulator gives biased constraints on σ_8. This bias disappears when small scales ($r < 10\ h^{-1}$Mpc) are excluded from the analysis. This shows that a vanilla halo model could introduce biases into the analysis of future datasets.

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

1 INTRODUCTION

The large scale structure (LSS) of the Universe as shown by three-dimensional galaxy maps carries a wealth of information which can be used to constrain theories of gravity. In particular, we can use the clustering properties of the LSS to address some of the most pressing questions faced by the standard cosmological model, such as what drives the accelerated expansion of the Universe and what is the dark matter. Ongoing and future surveys, such as the Dark Energy Spectroscopic Instrument (DESI) (DESI Collaboration et al. 2016), the Subaru Prime Focus Spectrograph (PFS) (Takada et al. 2014), and the space-based mission Euclid (Laureijs et al. 2011) will provide LSS maps of unprecedented statistical precision. The challenge for cosmologists now is to develop statistical methods that are accurate enough to match the precision of the data, so that we can extract all of the valuable information on gravity and cosmology contained in the LSS.

Galaxy clustering provides us with a means to constrain the cosmological model through late-Universe measurements. This enables us to carry out a consistency check by comparing cosmological constraints derived from observations of the early and late Universe and determining whether or not the results are consistent with the evolution expected in a Λ-cold dark matter (ΛCDM) model. Inconsistencies of more than 2-3 σ have been found when comparing the matter clustering strength, σ_8, inferred from the early Universe through cosmic microwave background (CMB) measurements, with the estimate from the late Universe, as deduced from both weak gravitational lensing and galaxy clustering (Joudaki et al. 2016; Hikage et al. 2019; Abbott et al. 2022; Philcox & Ivanov 2022). Late Universe probes prefer a smaller value of σ_8 and hence a lower degree of structure formation than is expected from cosmic microwave background (CMB)

⋆ E-mail: carolina.cuesta-lazaro@durham.ac.uk
† Argelander Fellow

observations (see Abdalla et al. 2022 for a detailed discussion of the so-called $\sigma_8 - S_8$ tension). Reducing the uncertainties on the estimated cosmological parameters would help to determine if the observed tension is the result of systematics, statistical bad luck, or even the imprint of new physics that is yet to be discovered.

Given a 3-D galaxy field, one could aim to infer the cosmological parameters directly at the field level by comparing the gridded number density of the observed galaxies with that expected from a model (Leclercq & Heavens 2021; Elsner et al. 2020). There are two different sets of variables that play a role in determining the expected number density. On the one hand, the random phases of the initial conditions determine where the initial seeds that gave rise to the observed large scale structure were located. On the other hand, the cosmological parameters influence how these seeds will collapse through gravitational evolution. However, jointly constraining the initial random phases and the cosmological parameters is a very challenging task. To avoid this difficulty, we define summary statistics of the 3-D galaxy maps that aim to reduce the stochasticity of sampling the initial conditions, whilst preserving as much information as possible about the cosmological parameters.

If the galaxy field were a Gaussian random field, its two-point statistics (the power spectrum or the two-point correlation function) would contain all information there is in the full 3-D maps. But while the density field at high redshift is indeed close to Gaussian over a wide range of scales, nonlinear gravitational evolution produces non-Gaussianity. Given that the mass overdensity $(\rho(x) - \bar{\rho}(x))/\bar{\rho}(x)$, where ρ is the mass density, is bounded at low values by -1, since a region of the Universe cannot have a negative density, the distribution of δ values must develop skewness as the density contrast grows. Finding alternative summary statistics to supplement the constraints obtained from the two-point functions is currently an active area of research (see, for instance, studies on the bispectrum, Hahn et al. 2020, the scattering transform, Valogiannis & Dvorkin 2022, and density split statistics for galaxy clustering Paillas et al. 2021).

An alternative way to maximise the information that is extracted from cosmological surveys is by modelling the cosmological dependence of small scale clustering. Although the statistical precision of data on small scales is higher than that on large scales, most studies that rely on perturbation theory (e.g. Chen et al. 2021) to model the dependence of two-point functions on cosmology restrict their analysis to pair separations larger than $\approx 30\ h^{-1}$Mpc. On smaller scales, the accuracy of perturbation theory breaks down rapidly, and its use introduces biases in the inferred cosmological parameters. The additional constraining power of small scales was demonstrated by Zhai et al. (2019) who showed how the constraints on the growth rate of structure, f, and the clustering amplitude, σ_8, increase monotonically as smaller scales are added to the analyses.

To obtain fully non-linear predictions for the properties of the large-scale structure and recover all the cosmological information contained in the small-scale clustering, we must resort to N-body simulations (Kuhlen et al. 2012). N-body simulations have been widely used as cosmic laboratories to test the precision and robustness of analytical methods for the large-scale structure (e.g., Carlson et al. 2009; Vlah et al. 2015; Cuesta-Lazaro et al. 2020), together with the effects of systematic errors in our measurements. Over the past decade, advances in computing have allowed us to produce a large enough number of dark matter only N-body simulations covering a significant fraction of the cosmological parameter space, which allows us to use the simulations themselves as predictive models that directly constrain the cosmological parameters. These simulations both need to be large enough to reduce sample variance, and have a high enough resolution to resolve the tracers that will be surveyed.

Moreover, in order to compare the outcomes of dark matter only simulations to the observed distribution of galaxies we have to model the connection between dark matter halos and galaxies (see Wechsler & Tinker 2018 for a review on this topic). Uncertainties in the galaxy-halo connection can limit the amount of information that we can extract from small scale clustering. We would like to use flexible models that can reproduce clustering in different scenarios of galaxy formation, whilst still being able to recover cosmological information after marginalising over the free parameters of the galaxy-halo connection model. Here, we use the empirical model of the halo occupation distribution (HOD) (Benson et al. 2000; Zheng et al. 2005), based on estimating the probability that a given halo hosts a galaxy.

Over the past few years, several studies (Zhai et al. 2019; Lange et al. 2019; Kobayashi et al. 2020b; Miyatake et al. 2021) have shown how N-body simulations can be leveraged to extract small scale information. Solving the inverse problem, estimating the posterior over the cosmological parameters given the observed clustering, would require the order of $O(10^6)$ N-body simulations to perform Bayesian inference with Markov Chain Monte Carlo. Therefore, most studies rely on modelling the dependence of the two-point correlation function on cosmology with surrogate models that are trained on a small set of $O(100)$ N-body simulations (Zhai et al. 2019; Lange et al. 2019; Kobayashi et al. 2020b). The surrogate models are orders of magnitude faster than the original N-body simulations and can then be used to sample the posterior of cosmological parameters.

For instance, Kobayashi et al. (2020b) developed an N-body version of the halo model for the galaxy power spectrum by training a neural network to reproduce the dark matter halo clustering properties in Fourier space. Zhai et al. (2019) and Yuan et al. (2022) followed a different route by emulating galaxy clustering as both a function of cosmology and galaxy-halo connection parameters with Gaussian processes (Rasmussen & Williams 2005). Alternatively, Lange et al. (2019) developed the so-called cosmological evidence modelling (CEM) method. Lange et al. (2019) used N-body simulations to compute the evidence of the data as a function of cosmology after marginalising over the HOD parameters, which can then be used to sample the posterior distribution over the cosmological parameters. In this way, the authors do not have to account for the errors introduced by the surrogate model although errors in the emulation of the likelihood function would still impact the inference. However, this approach does not yield joint constraints on the galaxy-halo connection and cosmological parameters, since the HOD parameters are marginalised over.

These simulation-based methods currently produce the tightest constraints on the combination $f\sigma_8$ (Lange et al. 2021; Kobayashi et al. 2022; Yuan et al. 2022; Zhai et al. 2022) when confronted with observations. Interestingly, all studies find values for the combination $f\sigma_8$ that are lower than those obtained from the CMB. The current challenge for emulator-based approaches is to both make sure that theoretical predictions are on a par with the statistical errors expected from future surveys, and that the modelling of how galaxies populate dark matter halos does not introduce biases into the analysis from small-scale clustering.

In this series of papers we build emulators for both real space clustering and pairwise velocity statistics (Peebles 1980; Fisher 1995); the latter determine the mapping between real and redshift space clustering. In this way, we will be able to combine constraints from clustering measurements and estimates of peculiar velocities, obtained through either the kinetic Sunyaev-Zeldovich effect (Sunyaev & Zeldovich 1980) (see Calafut et al. (2021) for a recent measurement) or through peculiar velocity surveys (Dupuy et al. 2019), to obtain more precise constraints on the cosmological parameters. Pe-

culiar velocity surveys and redshift space distortions have been shown to be specially complementary to test gravity theories (Kim & Linder 2020).

In this first paper of the series we focus on modelling small scale galaxy clustering in real space, improving on the emulators presented in Nishimichi et al. (2019) in terms of both accuracy and speed. We show how a combination of neural networks trained using the predictions of N-body simulations and the halo model can produce extremely accurate predictions for the clustering of galaxies over a wide range of pair separations, $0.01 < r < 150\,h^{-1}$ Mpc, as opposed to the range $r < 30\ h^{-1}$ Mpc, covered by previous emulators in configuration space (Zhai et al. 2019; Kobayashi et al. 2020b). This allows us to compute the likelihood using the full shape of the two-point correlation function, spanning the behaviour of the one- and two-halo terms. Finally, we demonstrate the limitations of the current implementation of the halo model to recover unbiased constraints when an assembly bias signal (Wechsler & Tinker 2018) is present in the data to be analysed.

This paper is organised as follows. In Section 2, we introduce the theoretical model of redshift-space clustering. In Section 3 we describe the N-body simulations used to train the emulator. In Section 4 we describe the halo model approach for predicting galaxy clustering. In Section 5 we present neural network emulators trained to reproduce the clustering of dark matter halos and their abundance, and show how they can be combined with the halo model to accurately reproduce the clustering of galaxies. Section 6 focusses on solving the inverse problem to obtain unbiased posterior distributions over the cosmological parameters. In particular, we show the limitations of the halo model in recovering unbiased constraints on σ_8 when assembly bias is present and scales smaller than $10h^{-1}$ Mpc are included in the likelihood. Finally, we present our conclusions in Section 7.

2 THEORETICAL BACKGROUND

The two-point correlation function, $\xi^R(r)$, quantifies clustering as the excess probability of finding a pair of galaxies at a given separation, compared with a random distribution of galaxies. The two-point correlation function is defined as

$$\xi^R(r) = \langle \delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r})\rangle, \tag{1}$$

where $\delta = (\rho(x) - \bar{\rho}(x))\,/\bar{\rho}(x)$ is the density contrast and $\bar{\rho}$ is the mean density. When assuming statistical isotropy and homogeneity, ξ^R depends only on pair separation, r.

In redshift surveys, we measure the angular positions of galaxies in the sky and their redshift. Then, the angular coordinates can be converted to comoving distances by assuming a cosmology through the angular diameter distance. If we assume that galaxies are at rest, as the photons emitted by galaxies travel towards us through an expanding universe, their wavelengths stretch accordingly, producing the redshift effect. We can translate this redshift into a comoving distance by introducing the Hubble factor, $H(z)$:

$$r(z) = \int_0^z \frac{dz'}{H(z')}, \tag{2}$$

where $r(z)$ is the comoving distance to the galaxy, and we have used the natural unit where the speed of light $c = 1$.

Nevertheless, there are several effects related to the distorted way in which we observe the Universe that complicate this simple picture. In fact, much of the information used to constrain cosmology from 3-D galaxy maps does not come directly from the underlying comoving map of galaxy positions, but from the distortion effects that alter this map.

In particular, galaxies move because of the gravitational pull generated by the inhomogeneous distribution of matter around them. If a source that emits light moves, the wavelength of its light becomes further redshifted because of the Doppler effect. If we ignored this effect, then we would infer the wrong position, $\mathbf{s}$, given by

$$\mathbf{s} = \mathbf{r} + \frac{\mathbf{v}(\mathbf{r})\hat{z}}{\mathcal{H}}\hat{z}, \tag{3}$$

instead of the real position of the galaxy, $\mathbf{r}$,where $\mathbf{v}(\mathbf{r})$ is the peculiar velocity of the galaxy, $\mathcal{H} = aH(a)$ the comoving Hubble factor, where a is the expansion factor, and the inferred distance, $\mathbf{s}$, is the redshift space position of the galaxy. Note that we have assumed that the observer is far away from the sources and therefore the line-of-sight direction can be fixed to a particular direction, regardless of the angular coordinates of the galaxy, which we arbitrarily set as the $\hat{z}$ axis.

Due to peculiar motions of galaxies, we observe redshift space positions, $\mathbf{s}$, instead of the real space positions, $\mathbf{r}$, and thus we can only measure

$$\xi^S(s_\perp, s_\parallel) = \langle \delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{s})\rangle, \tag{4}$$

which depends on both the pair separation, s, and its inclination with respect to the line of sight direction. Throughout, we denote the separations perpendicular and parallel to the line of sight by $s_\perp$ and $s_\parallel$, respectively.

The two-point correlation function of galaxies in redshift space has been used to obtain tight constraints on the cosmological parameters in a ΛCDM universe (e.g. Dawson et al. 2013, 2016). The so-called redshift space correlation function, $\xi^S(s_\perp, s_\parallel)$, is a combination of real space clustering, $\xi^R(r)$, and the probability of finding a pair of galaxies with a given relative velocity along the line of sight, also denoted as the pairwise velocity distribution, $P(v_\parallel|r_\parallel, r_\perp)$. This is summarised in the following equation for the streaming model of redshift space distortions (Fisher 1995; Scoccimarro 2004)

$$\xi^S(s_\perp, s_\parallel) = \int dr_\parallel \left(1 + \xi^R(r)\right) P(s_\parallel - r_\parallel|r_\parallel, s_\perp). \tag{5}$$

In Cuesta-Lazaro et al. (2020), we show how the mapping between the real and redshift space can be accurately described by an analytical expression for $P(v_\parallel|r_\parallel, r_\perp)$, where $v_\parallel$ is the line of sight velocity in units of conformal $H(a)$. In this series of papers, we will model the cosmological dependence of the different ingredients of the streaming model: i) the two-point correlation function in real space, shown in this paper, and ii) the lowest four-order moments of the velocity field, needed to perform the real-to-redshift space mapping, as shown in Cuesta-Lazaro et al. (2020).

3 THE DARK QUEST SIMULATION SUITE

Here, we briefly describe Dark Quest, a suite of cosmological N-body simulations used to build emulators. A detailed description can be found in Nishimichi et al. (2019).

3.1 N-body simulations

The Dark Quest simulations were performed with 2048^3 dark matter particles in $1\,h^{-1}$ Gpc (hereafter high-resolution runs, denoted HR) or $2\,h^{-1}$ Gpc (low-resolution runs, labelled LR) side-length boxes, using the Gadget2 N-body solver (Springel 2005).

Table 1. Comparison of the characteristics of the DarkQuest suite of simulations and those used to train clustering emulators in the literature. The mass of dark matter particles $M_{\rm part}$ has units of $(\Omega_{\rm m}/0.3)h^{-1}M_\odot$

Simulation Suite	Code	$L_{\rm box}$ $[h^{-1}{\rm Gpc}]$	$N_{\rm part}$	$M_{\rm part}$	Halo Finder	Reference
DarkQuest HR	GADGET2	1	2048^3	1.02×10^{10}	Rockstar	Nishimichi et al. (2019)
DarkQuest LR	GADGET2	2	2048^3	8.158×10^{10}	Rockstar	Nishimichi et al. (2019)
AbacusSummit Base	ABACUS	2	6912^3	2.1×10^{9}	CompaSO	Maksimova et al. (2021)
Aemulus	GADGET2	1.05	1400^3	3.51×10^{10}	Rockstar	DeRose et al. (2019)

The mass resolutions of the HR and LR runs are 1.02×10^{10} and $8.16\times10^{10}(\Omega_{\rm m}/0.3)\,h^{-1}\,M_\odot$, respectively. In Table 1, we show a comparison of the specifications of Dark Quest with those of other simulation suites that have been used to train clustering emulators in the literature (Zhai et al. 2019; Lange et al. 2019; Kobayashi et al. 2020b; Miyatake et al. 2021). Dark Quest, used int his work, has a higher resolution and a larger box size than Aemulus, but a lower resolution than AbacusSummit. In the future, it will be important to demonstrate the impact of differences in N-body codes (e.g. Grove et al. 2021), halo finders (e.g. Gómez et al. 2022), and resolution on the cosmological parameters inferred using simulation-based methods.

The initial conditions were generated using second-order Lagrangian perturbation theory (2LPT, Crocce et al. (2006)) and the redshift at which to generate the initial conditions was chosen depending on the cosmology and resolution (Nishimichi et al. 2019), with $z_{\rm init}\approx 59$ and 29 adopted for the fiducial HR and LR simulations respectively. Each simulation used different random number seeds to generate the initial conditions.

The cosmologies used in the simulations cover 101 flat geometry wCDM models, as shown in Fig. 1. In wCDM, the equation of state (EoS) for dark energy is parameterised through the value of w, also known as the EoS parameter of dark energy, $p_{\rm de}=w\rho_{\rm de}$, whose value is $w=-1$ in ΛCDM. Here, w is assumed to be constant.

The set of cosmological parameters is defined using optimal maximin distance sliced Latin hypercube designs (Ba et al. 2015), which enable efficient sampling from the six-dimensional parameter space,

$$C=\left\{\omega_{\rm b},\omega_{\rm c},\Omega_{\rm de},\ln\left(10^{10}A_{\rm s}\right),n_{\rm s},w\right\}, \tag{6}$$

where $\omega_{\rm b}\equiv\Omega_{\rm b}h^2$ and $\omega_{\rm c}\equiv\Omega_{\rm c}h^2$ are the physical density parameters of baryons and cold dark matter, respectively. The total matter density is the summation of the contributions from baryons, cold dark matter, and non-relativistic neutrinos:

$$\Omega_{\rm m}=\Omega_{\rm b}+\Omega_{\rm c}+\Omega_\nu, \tag{7}$$

where the physical density of neutrinos is fixed in the Dark Quest simulations as $\omega_\nu\equiv\Omega_\nu h^2\equiv0.00064$, corresponding to 0.06 eV for the total mass of the three mass eigenstates. For given values of $\omega_{\rm b},\omega_{\rm c}$ and the density parameter for dark energy $\Omega_{\rm de}$, the Hubble constant is derived from spatial flatness, that is,

$$\Omega_{\rm m}h^2=\omega_{\rm b}+\omega_{\rm c}+\omega_\nu, \tag{8}$$

$$\Omega_{\rm m}+\Omega_{\rm de}=1. \tag{9}$$

$A_{\rm s}$ and $n_{\rm s}$ are the amplitude and slope of the primordial curvature power spectrum normalised at $0.05\,{\rm Mpc}^{-1}$. The range of parameters explored is

$$\begin{aligned}
&0.0211375<\omega_{\rm b}<0.0233625,\\
&0.10782<\omega_{\rm c}<0.13178,\\
&0.54752<\Omega_{\rm de}<0.82128,\\
&2.4752<\ln\left(10^{10}A_{\rm s}\right)<3.7128,\\
&0.916275<n_{\rm s}<1.012725,\\
&-1.2<w<-0.8,
\end{aligned} \tag{10}$$

which is centred on the fiducial best fitting ΛCDM model to the Planck 2015 data alone (Planck Collaboration et al. 2016): $\omega_{\rm b}=0.02225,\omega_{\rm c}=0.1198,\Omega_{\rm de}=0.6844,\ln(10^{10}A_{\rm s})=3.094,n_{\rm s}=0.9645$ and $w=-1$. Fig. 1 shows a two-dimensional representation of the parameter space.

These parameter ranges correspond to the ranges of $(\pm5\%,\pm10\%,\pm20\%,\pm20\%,\pm5\%)$ for the parameters $(\omega_{\rm b},\omega_{\rm c},\Omega_{\rm de},\ln\left(10^{10}A_{\rm s}\right),n_{\rm s})$, respectively. These ranges were chosen to cover a parameter space that extends well beyond the constraints from the 2015 Planck data for a flat-ΛCDM model, for which the corresponding 68% intervals are $(0.72\%,1.25\%,1.33\%,1.10\%,0.51\%)$. Therefore, the Dark Quest simulations cover roughly up to a $\sim10\,\sigma$ range around the central best-fitting model to the Planck 2015 data. However, for the dark energy EoS parameter, w, a different approach was taken. Since Planck data alone cannot place a stringent constraint on w, and also, assuming that wCDM significantly loosens the constraints on the other parameters, we chose a strategy that is not strictly consistent for the six parameters. Instead, we used the Planck data combined with other external data sets only in the case of w (ie, $w=-1.019^{+0.075}_{-0.08}$ at 95% CL), and tried to cover a much wider range.

The simulation outputs were stored at 21 redshifts: 1.48, 1.35, 1.23, 1.12, 1.02, 0.932, 0.846, 0.765, 0.689, 0.617, 0.549, 0.484, 0.422, 0.363, 0.306, 0.251, 0.198, 0.147, 0.0967, 0.0478, and 0. These redshifts are evenly spaced in the linear growth factor for the fiducial Planck cosmology.

3.2 Halo catalogues

The identification of halos is of crucial importance, since the central premise of our method is to emulate dark matter halo properties, which can be robustly measured from N-body simulations. Appendix E of the Dark Quest paper (Nishimichi et al. 2019) provides comprehensive convergence tests of halo properties such as halo mass, the halo mass functions, and halo autocorrelation functions, with respect to the choice of halo finder, halo substructure separation, central/satellite split criterion, etc. In this section, we briefly review the main definitions that will be used in this paper.

The halo catalogues used here were identified using Rockstar (Behroozi et al. 2013), a friends-of-friends (FOF) halo finder that

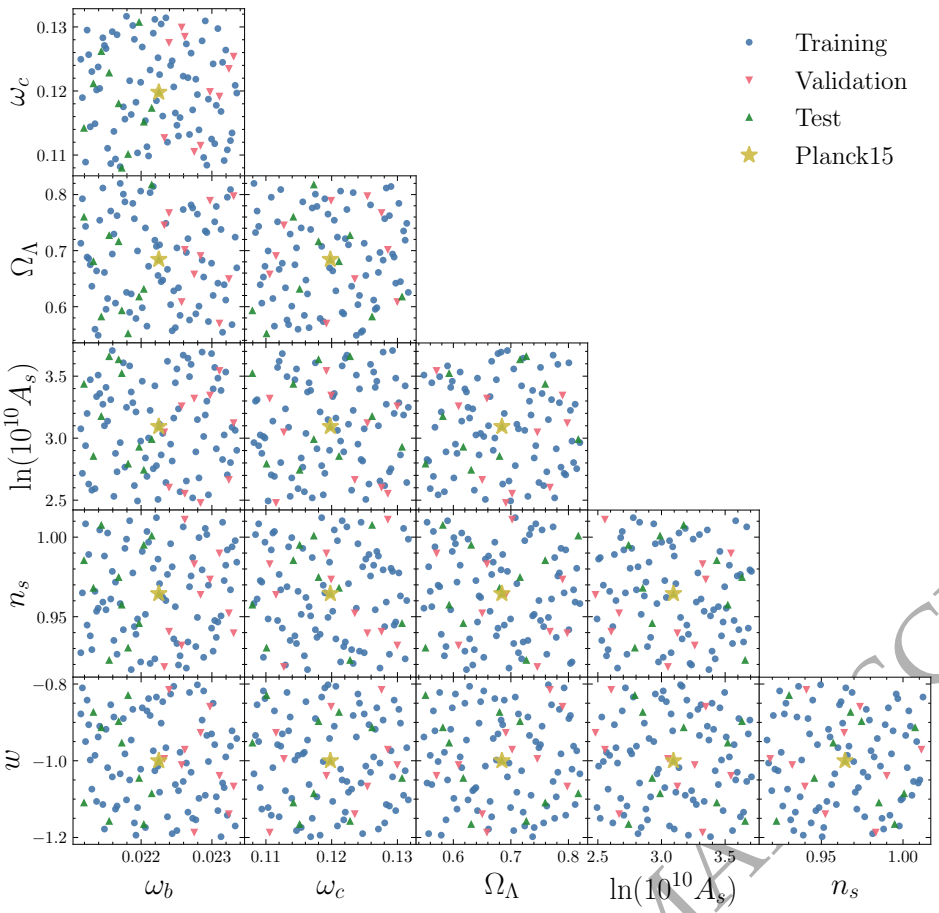

Figure 1. Corner plot representation of the 101 wCDM cosmologies covered by the Dark Quest simulation suite. We show the cosmologies chosen as training, test and validation sets, together with the best fitting fiducial cosmology to the 2015 Planck data, using different symbols, as indicated by the key.

operates in six-dimensional phase space. The halo centre is defined as the centre of mass position of the "core particles", a subset of member particles in the inner part of the halo. M_{200m} is adopted as the halo mass definition in Dark Quest, which is the mass enclosed within R_{200m}, the radius within which the average density is 200 times the mean mass density $\bar{\rho}_{m0}$. This definition of halo mass includes all simulation particles within a radius of R_{200m} from the halo centre, including gravitationally unbound ones. When the separation between the centres of different halos is within R_{200m} of any other halo, the most massive halo is marked as a central halo and the other halo(s) as a satellite halo(s). Only central halos with mass $M_{200m} \geq 10^{12}\,h^{-1}M_\odot$ are used in our analysis.

4 FROM DARK MATTER HALOS TO GALAXIES

As in Nishimichi et al. (2019), Miyatake et al. (2020) and Kobayashi et al. (2020b) we use the halo model to express the galaxy two-point correlation function in terms of dark matter halo properties. This allows us to make theoretical predictions for different galaxy samples, including cross-correlations of two different tracers, such as the ones that would be used in a multitracer analysis (McDonald & Seljak 2009), or the cross-correlation between clusters and galaxies. Moreover, a halo model implementation allows us to model the halo-galaxy connection analytically, which means that the accuracy of the results will not be worsened by emulator inaccuracies. As a downside, complex models of the halo-galaxy connection such as environment-based assembly bias may be harder to implement.

The halo model assumes that galaxies occupy dark matter halos, and therefore that the two-point galaxy correlation function can be split into contributions from galaxy pairs that inhabit the same dark matter halo, and pairs in which each member occupies a different dark mater halo (these terms will be referred to as the one and two halo terms, respectively):

$$\xi_{\rm gg}(r) = \xi^{\rm 1h}_{\rm gg}(r) + \xi^{\rm 2h}_{\rm gg}(r). \tag{11}$$

The one and two halo terms can be further split into correlations between two types of galaxies: centrals and satellites. Central galaxies are positioned at the minimum of the potential well of the dark matter halo and move with the halo's centre of mass velocity. Satellite galaxies orbit within the dark matter halo with virialised velocities. We assume that the distribution of satellite galaxies is given by an NFW profile, $u_{\rm NFW}(r|c(M))$ (Navarro et al. 1997). This approximation has been tested against hydrodynamical simulations, finding it valid for galaxies selected by number density (Bose et al. 2019). The NFW profile is defined by one parameter: the concentration of the halo, c, which varies with halo mass, redshift, and cosmological parameters (Ludlow et al. 2016; Diemer & Joyce 2019). Here, we use the median concentration-mass relation $c(M)$ from Diemer & Joyce (2019).

Regarding the galaxy-halo connection, we use the halo occupation distribution (HOD) (Zheng et al. 2005) to model the number

of galaxies in a given halo as a function of halo mass. The occupation of central galaxies is parameterized as a Bernoulli distribution, whereas that of satellites is assumed to be Poisson distributed. Both distributions are described by their mean parameters

$$\left\langle N_{\rm g}\right\rangle(M) = \langle N_{\rm c}\rangle(M) + \langle N_{\rm s}\rangle(M). \tag{12}$$

We parameterize the mean galaxy numbers as in Zheng et al. (2005) by introducing the following HOD parameters

$$\mathcal{G} = \{M_{\rm min}, \sigma_{\log M}, M_1, \kappa, \alpha\}, \tag{13}$$

where $M_{\rm min}, \sigma_{\log M}$, and M_1, κ, α define the occupation of the centrals and satellites, respectively.

We describe the mean number of central galaxies for a given halo as

$$\langle N_{\rm c}\rangle(M|\mathcal{G}) = \frac{1}{2}\left(1 + {\rm erf}\left(\frac{\log M - \log M_{\rm min}}{\sigma_{\log M}}\right)\right), \tag{14}$$

where ${\rm erf}(x)$ is the error function. The mean occupation number of satellite galaxies is defined as

$$\begin{aligned}\langle N_{\rm s}\rangle(M|\mathcal{G}) &= \langle N_{\rm c}\rangle(M|\mathcal{G})\,\lambda_{\rm s}(M|\mathcal{G}) \\ &= \langle N_{\rm c}\rangle(M)\left(\frac{M - \kappa M_{\rm min}}{M_1}\right)^{\alpha}.\end{aligned} \tag{15}$$

The empirical HOD model that we use is extremely simple. One of the simplifying assumptions is that galaxy occupation depends solely on the mass of the dark matter halo. Although dark matter halo mass correlates strongly with clustering, we know that dark matter halos experience different assembly histories even at a fixed halo mass, which can affect their clustering (Gao et al. 2005; Gao & White 2007). These different assembly histories influence secondary properties of halos, and this might, in turn, affect the formation of galaxies and hence the galactic content of halos of a given mass. These effects together – the variations in halo clustering and galactic content with halo mass and a second halo property – are known as galaxy assembly bias (see Wechsler & Tinker 2018 for a recent review on the galaxy-halo connection and assembly bias). The question we will address in Section 6.3, is whether a simplified version of the galaxy-halo connection is flexible enough to recover unbiased constraints on the cosmological parameters.

Given these assumptions, we can express the two-point galaxy correlation function in terms of dark matter halo properties. To simplify the calculations, we further split the one and two halo terms into correlations of central and satellite galaxies,

$$\xi_{\rm gg}(r) = \xi^{\rm 1h}_{\rm ss}(r) + 2\xi^{\rm 1h}_{\rm cs}(r) + \xi^{\rm 2h}_{\rm cc}(r) + 2\xi^{\rm 2h}_{\rm cs}(r) + \xi^{\rm 2h}_{\rm ss}(r). \tag{16}$$

In the equations below, we highlight the emulated quantities in blue, such as the halo mass functions, ${\rm d}n/{\rm d}M$, and halo auto correlation functions, $\xi_{\rm hh}(r)$, following the convention used in Miyatake et al. (2020). Note that terms involving both centrals and satellites lead to the convolution of the halo profiles and the halo two-point correlation function. It is therefore simpler to compute these terms in Fourier space, where convolutions in coordinate space become simple products, and then apply an inverse Fourier transform to the result. Therefore, we compute

$$P^{\rm 1h}_{\rm ss}(k) = \frac{1}{\bar{n}_{\rm g}^2}\int {\rm d}M \frac{{\rm d}n}{{\rm d}M}(M)\,\langle N_{\rm c}\rangle(M)\lambda^2_{\rm s}(M)u_{\rm NFW}(k|M,c(M))^2, \tag{17}$$

where $u_{\rm NFW}(k|M,c(M))$ is the Fourier transform of the truncated NFW profile (see Eq. (81) in Cooray & Sheth 2002).

The cross-correlation between centrals and satellites that occupy the same halo is given by

$$P^{\rm 1h}_{\rm cs}(k) = \frac{1}{\bar{n}_{\rm g}^2}\int {\rm d}M \frac{{\rm d}n}{{\rm d}M}(M)\,\langle N_{\rm c}\rangle(M)\lambda_{\rm s}(M)u_{\rm NFW}(k|M,c(M)), \tag{18}$$

where ${\rm d}n/{\rm d}M(M)$ is the halo mass function defined as the comoving number density of halos for a given halo mass, and $\bar{n}_{\rm g}$ is the galaxy number density that we obtain by integrating the halo mass function weighted by the halo occupation

$$\bar{n}_{\rm g} = \int {\rm d}M \frac{{\rm d}n}{{\rm d}M}\left(\langle N_{\rm c}\rangle(M) + \langle N_{\rm s}\rangle(M)\right). \tag{19}$$

Meanwhile, the different two-halo terms will result in weighted averages of the dark matter halo two point correlation function and convolutions with NFW profiles when satellite correlators are involved

$$\begin{aligned}P^{\rm 2h}_{\rm cs}(k) =& \frac{1}{\bar{n}_{\rm g}^2}\int {\rm d}M \frac{{\rm d}n}{{\rm d}M}(M)\,\langle N_{\rm c}\rangle(M) \\ &\int {\rm d}M' \frac{{\rm d}n}{{\rm d}M}(M')\,\langle N_{\rm c}\rangle(M')\lambda_{\rm s}(M') \\ &P_{\rm hh}(k|M,M')u_{\rm NFW}(k|c(M'))),\end{aligned} \tag{20}$$

$$\begin{aligned}P^{\rm 2h}_{\rm ss}(k) =& \frac{1}{\bar{n}_{\rm g}^2}\int {\rm d}M \frac{{\rm d}n}{{\rm d}M}(M)\,\langle N_{\rm c}\rangle(M)\lambda_{\rm s}(M) \\ &\int {\rm d}M' \frac{{\rm d}n}{{\rm d}M}(M')\,\langle N_{\rm c}\rangle(M')\lambda_{\rm s}(M') \\ &P_{\rm hh}(k|M,M')u_{\rm NFW}(k|c(M'))u_{\rm NFW}(k|c(M))).\end{aligned} \tag{21}$$

We avoid the Fourier transform when computing central-central terms

$$\begin{aligned}\xi^{\rm 2h}_{\rm cc}(r) =& \frac{1}{\bar{n}_{\rm g}^2}\int {\rm d}M \frac{{\rm d}n}{{\rm d}M}(M)\,\langle N_{\rm c}\rangle(M) \\ &\int {\rm d}M' \frac{{\rm d}n}{{\rm d}M}(M')\,\langle N_{\rm c}\rangle(M')\xi_{\rm hh}(r|M,M').\end{aligned} \tag{22}$$

In the next section, we show how we can use neural networks to emulate the two statistics shown in blue that vary with cosmological parameters: ${\rm d}n/{\rm d}M$ and $\xi_{\rm hh}$.

4.1 The best of both universes: combining simulations of different resolutions

Although the high-resolution (HR) simulations can resolve halos of lower masses than their low-resolution (LR) counterparts, their smaller box size results in a larger sample-variance noise than in the LR boxes, in particular on large scales.

The halo model approach outlined above allows us to calibrate the halo autocorrelation function using the LR simulations, to reduce sample variance when using measurements from one realisation, while calibrating the halo mass function with the HR simulations. The mass resolution of the HR simulations is sufficient to accurately estimate the halo mass function down to a minimum halo mass of $10^{12}h^{-1}\,M_\odot$, smaller than the typical mass of a host halo for CMASS or LOWZ galaxies. On the other hand, precise estimates of halo auto-correlations on large scales require bigger box simulations to decrease the shot noise. In this section, we examine the impact of

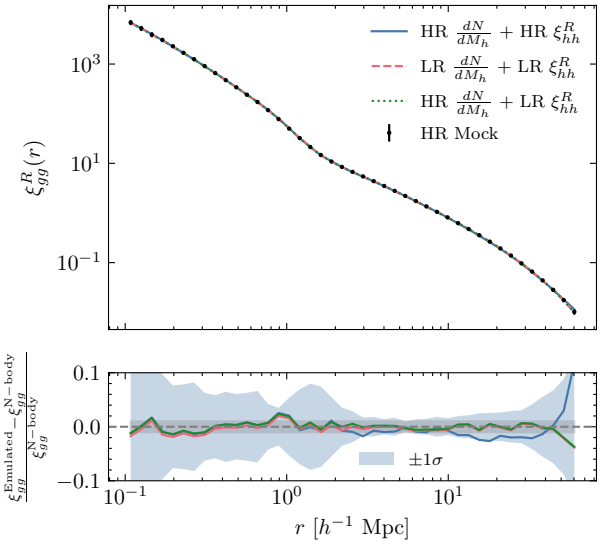

Figure 2. We show $\xi^R_{\rm gg}$ obtained by populating the 25 realizations of the fiducial cosmology on the HR simulations with mock LOWZ galaxies, compared to the result of Eq. 16 when either: i) both $\mathrm{d}N/\mathrm{d}M_{\rm h}$ and $\xi^R_{\rm hh}$ are measured on the HR simulations (in blue), ii) both $\mathrm{d}n/\mathrm{d}M_{\rm h}$ and $\xi^R_{\rm hh}$ are measured on the LR simulations (in red) and iii) $\mathrm{d}n/\mathrm{d}M_{\rm h}$ is obtained from the HR simulations and $\xi^R_{\rm hh}$ from the larger boxsize LR ones (in green). The fractional difference plot in the lower panel shows that the sample variance in the blue line based on the correlation function measured from one HR box is greatly reduced by replacing it with LR simulations without introducing bias. Blue shaded denote the standard deviation of the 25 realizations of the HR simulations. the gray shaded regions denotes 1% errors.

combining the halo mass function of HR simulations with the halo correlation function measured in LR simulations.[1]

In Fig. 2, we show a comparison of a mock LOWZ-like catalogue obtained from the 25 realisations of the fiducial cosmology for the HR simulations, to the result of Eq. 16 when i) we combine the halo mass function from HR simulations, with the halo two-point correlation function estimated from one of the HR boxes (solid blue line), ii) estimate both the halo mass function and halo two-point correlation function from the LR simulations (dashed red), and iii) measure the halo mass function in the HR simulation, and the halo auto-correlation from the LR simulation. Fig. 2 shows that combining clustering measurements from low-resolution simulations with a halo mass function measured in the HR simulation does not introduce any biases and reduces the sample-variance noise.

5 NEURAL NETWORK EMULATORS FOR DARK MATTER HALO PROPERTIES

Nishimichi et al. (2019) fitted both the halo mass function and the halo autocorrelation function measured from the N-body simulations using a combination of principal component analysis (PCA), to reduce the dimensionality of the data vector, and Gaussian processes (GP), to fit the dependence of the principal component coefficients on cosmology. Here, we show how dimensionality reduction can be avoided by using neural network emulators, leading to increased accuracy in the prediction of halo properties.

[1] Note we could also have extended the mass resolution of the LR halo catalogues, using a scheme like the introduced by Armijo et al. (2022).

Fully connected neural networks approximate a function f such that

$$\mathbf{y} = f(\mathbf{x}|\boldsymbol{\theta}), \tag{23}$$

where $\mathbf{x}$ represents the features of the data set, $\mathbf{y}$ the desired outputs, and $\boldsymbol{\theta}$ the network-free parameters, also called trainable parameters. The optimal function f is defined by the set of values θ that minimise the loss function (the form of which is discussed below). The loss function provides a measure of the model's performance when evaluated on the data set.

ReLU (Rectified Linear Unit; Agarap 2018) is the most commonly used activation function in current neural networks used to add non-linearities in the mapping between inputs and outputs, and is defined as

$$\mathrm{ReLU}(x) = \max(0, x), \tag{24}$$

where x is the output of the previous layer of the neural network. Note that ReLU activations are not differentiable at zero. Here, however, we are interested in functions that are differentiable with respect to their inputs and, in particular, with respect to the cosmological parameters (since these derivatives could be used to accelerate parameter inference through Hamiltonian Monte Carlo techniques, e.g. Duane et al. 1987, or to accelerate Fisher forecasts). Therefore, throughout, we use Gaussian error linear units (GELUs) as activation functions instead (Hendrycks & Gimpel 2016):

$$\mathrm{GELU}(x) = 0.5x\left(1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right). \tag{25}$$

To find the optimal parameters, θ, which reproduce the statistics measured from the N-body simulations, we minimise the L1 norm loss function

$$\mathcal{L} = \frac{1}{N}\sum_{i=0}^{N} |y^i_{\rm true} - y^i_{\rm predicted}|, \tag{26}$$

using the Adam optimiser (Kingma & Ba 2014). Note that L1 reduces the importance given to outlier errors compared to the use of the mean squared error (also known as the L2 norm). We will refer to the value of Eq. 26 evaluated in the training and validation dataset as training and validation loss, respectively.

Moreover, we avoid fine-tuning the value of the learning rate by using a learning rate scheduler that reduces the learning rate by a factor of 10 every time the validation loss does not improve after 20 epochs. We also stop training the model when the validation loss does not improve after 100 epochs. This iterative reduction of the learning rate allows the model to quickly learn the broad characteristics of the data and then reduce the errors by adopting a smaller learning rate. The initial learning rate is always set to 0.015.

In the following subsections, we demonstrate the precision of fully connected networks in reproducing the real-space correlation function and the halo mass function obtained from the DARK QUEST simulations.

5.1 Real space correlation function

5.1.1 Measurement

The details of the halo correlation function measurements are introduced in Nishimichi et al. (2019). Here, we present only a summary of the most important aspects.

First, noisy measurements of $\xi(r|M, M')$ are avoided by instead measuring ξ as a function of halo number density, n, and switching from differential to cumulative mass limits. We then use the halo mass

function to translate predictions as a function of number density into predictions as a function of differential mass through the relation

$$\xi(r|n(m),n(m')) = \frac{\int_m^\infty \mathrm{d}M \int_{m'}^\infty \mathrm{d}M' \xi(r|M,M') \frac{\mathrm{d}n}{\mathrm{d}M}(M) \frac{\mathrm{d}n}{\mathrm{d}M}(M')}{\int_m^\infty \mathrm{d}M \int_{m'}^\infty \mathrm{d}M' \frac{\mathrm{d}n}{\mathrm{d}M}(M) \frac{\mathrm{d}n}{\mathrm{d}M}(M')} = \frac{\int_m^\infty \mathrm{d}M \int_{m'}^\infty \mathrm{d}M' \xi(r|M,M') \frac{\mathrm{d}n}{\mathrm{d}M}(M) \frac{\mathrm{d}n}{\mathrm{d}M}(M')}{n(M)n(M')}, \tag{27}$$

which can be inverted to obtain

$$\xi(r|M,M') = \frac{\frac{\partial^2}{\partial m \partial m'} \left[n(m)n(m')\xi(r|n(m),n(m'))\right]}{\frac{\mathrm{d}n}{\mathrm{d}M}(M) \frac{\mathrm{d}n}{\mathrm{d}M}(M')} = \frac{\partial^2}{\partial n \partial n'} \left[n(m)n(m')\xi(r|n(m)n(m'))\right]. \tag{28}$$

Measurements are made in 8 logarithmically spaced bins in number density over the range $n_\mathrm{h} = \left[10^{-6}, 10^{-2.5}\right] \left(h^{-1}\mathrm{Mpc}\right)^{-3}$. Note that there are 36 independent combinations for two halo samples with different number densities. The pair separation r is split into 40 logarithmically spaced bins from 0.01 to 5 h^{-1} Mpc and 75 linear bins from 5 to 150 h^{-1} Mpc, and over the 21 simulation snapshots spanning from $z = 1.48$ to $z = 0$.

In total, the data set is made up of 80 cosmologies in the training set, 10 in the validation set and 10 in the test set, each with its corresponding 21 snapshots and 36 number density bins.

On large scales, we can reduce cosmic variance by using the propagator-based prescription of Crocce & Scoccimarro (2006). For Gaussian initial conditions, the propagator can be expressed as the ratio of the cross-power spectrum between the density field at the initial conditions and the nonlinear field at the redshift of interest, to the linear power spectrum. This calculation was originally performed for the matter density, but can be extended to the halo density field. The propagator quantifies how much of the memory of the initial conditions is preserved in the final nonlinear density field. The propagator describes the smearing of BAO feature due to large-scale bulk flows. One can straightforwardly generalize this approach to any tracer. This function also describes the linear bias factor in the large-scale limit. The advantage of using the propagator is that a large fraction of sample-variance error is cancelled when the ratio between the two spectra is taken. In addition, it is known that the k dependence of the propagator is simple. A Gaussian-like parameterized function is sufficient to model this accurately (see Nishimichi et al. 2019 for more details).

We have slightly updated the implementation of this idea here. In Nishimichi et al. (2019), to evaluate the correlation function, both the directly emulated correlation function (for small separations) and the propagator-based model (for large separations), in which the propagator is also emulated, are computed and then stitched together to cover a wide range of separations. This requires us to build two separate emulators and both of them must be used when evaluating the correlation function. Here, instead, we now work at the data level: for each simulation box, we construct a data vector that combines the two methods. We refined the stitching scheme to yield a smoother transition between the two regimes (Nishimichi et al. in prep.). Now, our neural-network emulator learns this new datavector, to which the propagator trick has already been applied.

5.1.2 Emulation

We train a fully connected neural network, f, to perform the following mapping

$$\log_{10}\left(\xi^\mathrm{R}_\mathrm{hh}(r)\right) = f(C, \log_{10}(n_1), \log_{10}(n_2), z), \tag{29}$$

where n_1 and n_2 denote the number densities of each halo sample, z is the redshift and C represents the set of cosmological parameters in Eq. 6.

Note that the input to the neural network has been standardised to facilitate training (such that its mean is 0 and standard deviation is 1). The output of the neural network is the logarithm of the correlation function $\log_{10}(\xi_\mathrm{hh})$, which is also standardised:

$$\log_{10}\left(\xi^\mathrm{R}_\mathrm{hh}(r)\right) \to \frac{\log_{10}\left(\xi^\mathrm{R}_\mathrm{hh}(r)\right) - \left\langle \log_{10}\left(\xi^\mathrm{R}_\mathrm{hh}(r)\right)\right\rangle}{\sqrt{\mathrm{Var}\left(\log_{10}\left(\xi^\mathrm{R}_\mathrm{hh}(r)\right)\right)}}, \tag{30}$$

where $\left\langle \log_{10}\left(\xi^\mathrm{R}_\mathrm{gg}(r)\right)\right\rangle$ and $\mathrm{Var}\left(\log_{10}\left(\xi^\mathrm{R}_\mathrm{gg}(r)\right)\right)$ are the mean and variance of all correlation functions, estimated from the training set.

The output of the neural network is all the values of the correlation function evaluated for the pair-separation vector, r. Interestingly, when fitting the neural network with r as input, the model tends to overfit the data and converges to a less accurate overall model, while combining all pair separations shares the weights of the neural network across the values of r and reduces the level of overfitting.

We summarise the best-fitting hyperparameters of the neural network in Table 2.

In Fig. 3, we show the performance of the neural network as a function of pair separation compared to that found in Nishimichi et al. (2019). Fig. 3 shows the absolute errors estimated in the test set, as a function of pair separation r. Number densities and redshifts have been averaged.

The median absolute errors are lower than 2% throughout the entire scale range, a factor of 4 smaller than the upper limit of Nishimichi et al. (2019), while 68% had errors smaller than 6%, which is a factor of 5 smaller. We further compare the variance of the emulator errors (68th percentile fractional residuals) to the variance in the simulations themselves (grey solid background). This comparison shows that the emulator is already performing at a level similar to the variance in the simulations over the full-scale range. Note also that we cannot accurately estimate the model accuracy below the level of sample variance in the simulations, given that we only compare the accuracy of the model against one N-body realisation for each cosmology in the test set.

5.2 Halo mass function

5.2.1 Measurement

As explained earlier, we used the HR simulations to model the halo mass function. To do this, we first create a histogram of the number of halos in 80 logarithmically spaced bins in halo mass over the range of 10^{12} to $10^{16}\, h^{-1}\, M_\odot$. Following Nishimichi et al. (2019), we apply a correction to individual halo masses to account for systematics due to the finite number of particles. The corrected mass is given by (e.g. Warren et al. 2006):

$$\tilde{M} = (1 + N_\mathrm{p}^{-0.55})M, \tag{31}$$

where N_p is the number of simulation particles contained in the halo. The raw histogram is rather noisy, especially at the high-mass tail due to the small number of halos per bin. To produce a smooth mass

Table 2. The summary of the best performing set of hyperparameters for the neural network emulators used to predict halo properties. The last column indicates the simulation resolution from which the quantity listed in the first column is measured.

Statistic	Batch size	Activation	$N_{\rm hidden}$	Resolution
$\xi_{\rm hh}$	5000	GELU	1024, 512, 512	LR
$\frac{{\rm d}n}{{\rm d}M}$	5000	GELU	1024, 512, 512	HR

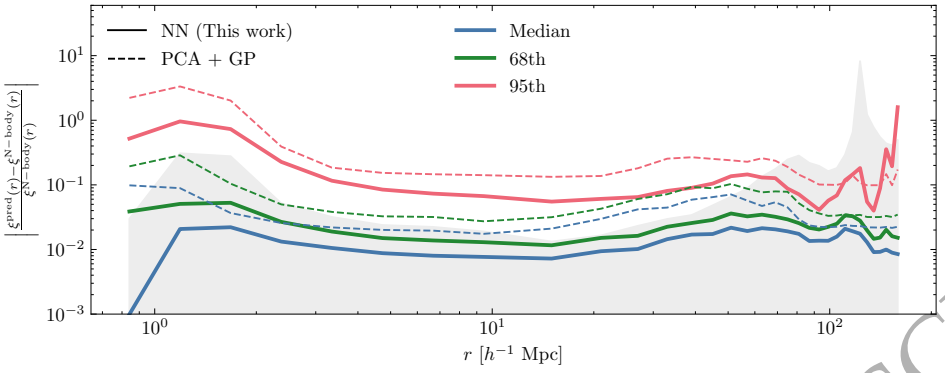

Figure 3. Comparison of the absolute fractional errors of the neural network emulator for the halo real space two point correlation function, with the Gaussian process + PCA approach presented in Nishimichi et al. (2019). Note that we only include test set data, but for all redshifts and halo number densities. The grey shading shows the variance estimated from the simulations using the 15 realisations of the fiducial Planck cosmology, $\sigma_{\xi_{\rm fiducial}}/\xi_{\rm fiducial}$.

function, we fit the data points using the functional form employed in Tinker et al. (2008). In doing so, we fix the parameter "b" in the formula, which controls the low mass behaviour, to the original value in Tinker et al. (2008) and allow the other three parameters to vary freely. We weight the bins according to the Poisson noise, which is more important at high masses, and the mass-determination accuracy, which is sensitive to the number of particles in the halo

$$\frac{\Delta N_{\rm h}}{N_{\rm h}} = \frac{1}{\sqrt{N_{\rm h}}} + \frac{1}{N_{\rm p}}. \tag{32}$$

The uncertanties in the fitted parameters are propagated to the smooth model prediction to obtain the expectation value, as well as the uncertanties of the estimated halo number counts in each mass bin.

5.2.2 *Emulation*

As in the case of the halo two-point correlation function, we train the model on the logarithm of the halo mass function to reduce the dynamic range of the observable. In this case, the mapping we obtain is

$$\log_{10}\left(\frac{{\rm d}n}{{\rm d}M}\right) = f(C, z). \tag{33}$$

As before, we standardise inputs and outputs before training the model.

In Fig. 4, we compare the N-body measurements from the 10 test cosmologies with the emulator predictions at $z = 0$. The emulator achieves subpercent accuracy for halo masses smaller than 10^{14} $h^{-1}\,M_\odot$, with the error increasing for larger halo masses. Estimating the error is, however, challenging for halo masses larger than 10^{14} $h^{-1}\,M_\odot$ due to the large Poisson noise that affects the measurements caused by the small number of cluster-size halos in the simulations.

In Fig. 5, we evaluate the overall accuracy of the halo mass function

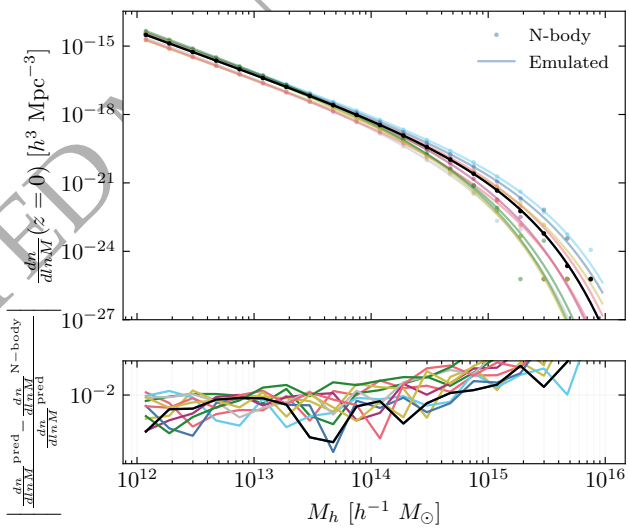

Figure 4. N-body measurements (points) and emulator predictions (lines) for the halo mass function at $z = 0$ in the 10 test set cosmologies. The lower panel shows the absolute fractional errors as a function of halo mass. The fiducial Planck cosmology is shown in black.

emulator at all redshifts (left panel) and as a function of the redshift (right panel). We find that the median emulator error for all redshifts is below 1 per cent for halo masses smaller than $10^{13.5}$ $h^{-1}\,M_\odot$, and increases rapidly to values larger than 10 per cent for the most massive halos ($M_{\rm h} > 10^{15}$ $h^{-1}\,M_\odot$). The right panel of Fig. 5 shows that the accuracy of the emulator degrades slightly at the highest redshifts considered ($z = 1.48$).

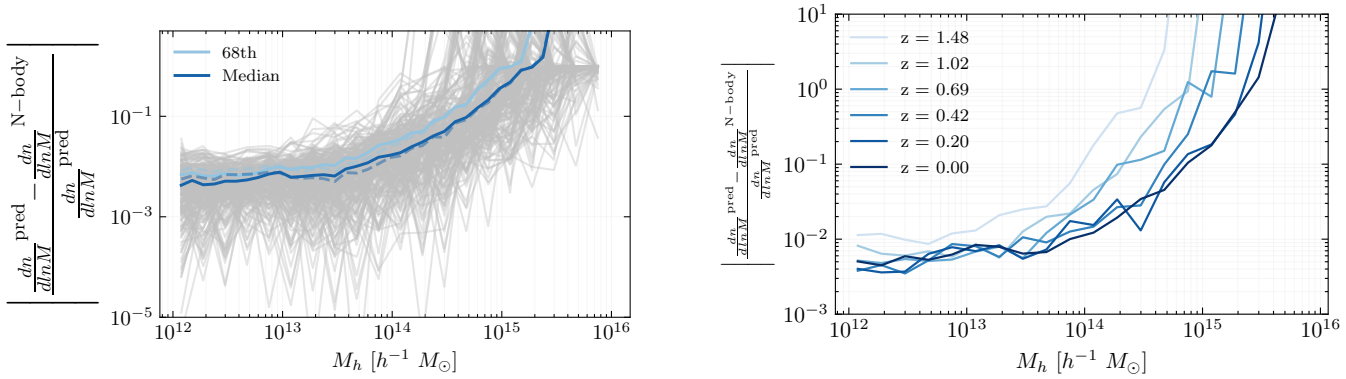

Figure 5. Absolute fractional errors on the halo mass function emulator predictions as a function of halo mass. The left panel shows the result for each test set sample (the 10 set cosmologies evaluated at the 21 different redshifts) as a gray line, along with the median (dark blue line) and 68th percentile range (light blue line) of the absolute fractional errors. The errors from the Gaussian process + PCA approach presented in Nishimichi et al. (2019) are shown in dashed lines for comparison. We find similar levels of accuracy. The right panel shows the median absolute error as a function of halo mass, with different lines showing different redshifts, as indicated by the legend.

5.3 Galaxy clustering

We now assess the impact that inaccuracies in halo emulators have on galaxy clustering predictions. To do so, we populate the 10 test and 10 validation LR simulations with mock galaxies. We populate each cosmology at four different snapshots (z=0.1,0.25,0.5 and 0.75) and 5 different galaxy number densities, logarithmically spaced between $\log\left(\bar{n}_{\rm gal}/(h^{-1}{\rm Mpc})^{-3}\right) = -3.7$ and $\log\left(\bar{n}_{\rm gal}/(h^{-1}{\rm Mpc})^{-3}\right) = -4.3$. Note that halo property emulators cannot estimate galaxy clustering for arbitrary number densities, given that the lowest halo mass resolved by the Dark Quest simulations is $10^{12}\ h^{-1}\ M_\odot$.

For each combination of cosmology, redshift, and number density, we randomly sampled the HOD parameters from the ranges

$$\begin{aligned}\sigma_{\log M} &\in [0.1, 0.8]\\ \alpha_{\rm sat} &\in [0.5, 1.]\\ \kappa &\in [0.1, 0.8]\\ \log M_1 &\in [13.5, 14.5]\ .\end{aligned}$$

The remaining HOD parameter, $\log M_{\rm min}$, is fixed by the given galaxy number density. In total, we built a diverse sample of 400 HOD mocks with varying cosmology, HOD parameters, and redshift, to test the performance of the emulator.

Fig. 6 shows the emulator predictions for 20 HOD mocks at fixed redshift ($z = 0.25$), each of the curves is generated from a different set of cosmological parameters in the test and validation sets. Comparing the mock HOD catalogues with the emulator predictions, we find that the median error of the emulator is below 3 per cent on scales smaller than 50 h^{-1} Mpc, as shown in Fig. 7. Furthermore, the 68th percentile interval of the error increases only by 1 per cent point with respect to the median. There is a small increase ($\approx$ 1 per cent point) in the error in the transition from one-to-two-halo term that occurs between 1 and 2 h^{-1} Mpc. On large scales, the variance of the measurements is large, making it difficult to accurately determine the error of the emulator.

Fig. A2 shows the performance of the emulator as a function of the galaxy number density and redshift. In both cases, the emulator shows similar levels of performance and therefore does not show any bias.

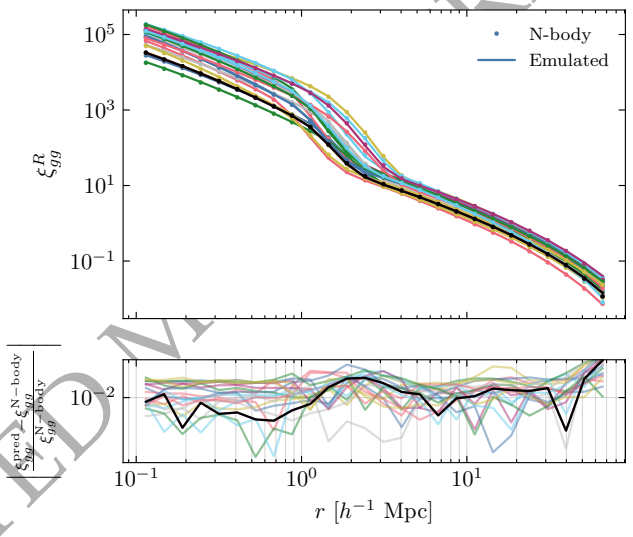

Figure 6. Emulator predictions for a subset of the 400 HOD mocks generated to test the accuracy of galaxy clustering. We show only those at $z = 0.25$. Planck cosmology is shown in black. The top panel shows all measurements from the 20 HOD catalogues and the corresponding emulator prediction. On the bottom pannel, we show the absolute error of the emulator as a function of scale.

6 SOLVING THE INVERSE PROBLEM: FROM CORRELATIONS TO COSMOLOGY

Here, we show how the galaxy two-point correlation function emulator is able to recover the cosmological parameters from mock simulated galaxies, first using the same HOD prescription as the one implemented in our theoretical model within the 68% credible interval for all parameters.

It should be emphasised that we focus on the three-dimensional two-point correlation of galaxies in real space, which is not directly observable in galaxy surveys. What we observe is the redshift space two-point correlation function of galaxies, which will be the subject of the second paper in this series. On large scales, it is also important to account for the additional anisotropy in the correlation function of galaxies introduced by the conversion between observed angular separations and redshifts into comoving coordinates, the so-called

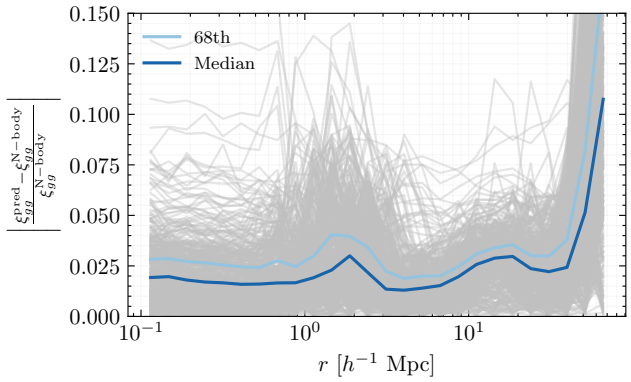

Figure 7. We show the absolute error of the emulator as a function of scale for each of the 400 HOD mocks generated to test the accuracy of galaxy clustering predictions for different cosmologies, redshifts, and galaxy number densities. The light and dark blue lines show the 68th credible interval and the median of the absolute errors.

Alock-Paczynski (AP) effect (Alcock & Paczynski 1979). However, it is important to show that the emulator is capable of recovering the parameters of interest for a mock dataset and to study the potential biases that might arise from adopting a too simplistic HOD model. We will also examine the scale dependence of the cosmological information content, which will, in turn, be important in determining the information content in redshift space.

We generated mock galaxy catalogues for LOWZ SDSS-like galaxies based on the fiducial Planck cosmology of the Dark Quest HR simulations, following Kobayashi et al. (2020a). See Table 3 for the characterisation of the mock sample.

We use nested sampling, in particular the implementation of pymultinest (Buchner et al. 2014), to obtain samples from the posterior distribution. The posterior is defined as

$$p(\theta|\mathcal{D}) \propto \mathcal{L}(\mathcal{D}|\theta)p(\theta), \tag{34}$$

where θ are the parameters to be estimated, $p(\theta|\mathcal{D})$ is the posterior distribution of the parameters given the data, $\mathcal{L}(\mathcal{D}|\theta)$ describes the likelihood of the data given the parameters, and $p(\theta)$ is the prior distribution of the model parameters.

We used a combination of the real space two-point correlation function and galaxy number density as our data vector and assumed that the likelihood follows a Gaussian distribution. Therefore, we compute the log-likelihood (up to a normalisation factor) as follows

$$\begin{aligned}\mathcal{L}(\mathcal{D}|\theta) = -\frac{1}{2}\sum_{r_i,r_j}\left[\xi^s(r_i)-\xi^s(r_i|\theta)\right]\times\mathcal{C}^{-1}(\xi^s(r_i),\xi^s(r_j)) \\ \times\left[\xi^s(r_j)-\xi^s(r_j|\theta)\right]+\frac{(n_{\rm g}^s-n_{\rm g}^s(\theta))^2}{\sigma_{n_{\rm g}}^2},\end{aligned} \tag{35}$$

where $\xi^s(r_i)$ denotes the two-point correlation function of the data for sample s, and $\xi^s(r_i|\theta)$ is the prediction of the theoretical model where θ denotes the model parameters, i.e. cosmological and HOD ($\mathcal{C}+\mathcal{G}$), C is the data covariance matrix for a volume of 0.67 $(h^{-1}\,\mathrm{Gpc})^3$, $n_{\rm g}^s$ is the galaxy number density estimated from the data, $n_{\rm g}^s(\theta)$ the theoretical prediction, and $\sigma_{n_{\rm g}}$ the estimated error of the data that we fix to a nominal value of 5 per cent. Note that the galaxy number density depends both on the HOD parameters and on cosmology, as seen in Eq. (19). See Appendix B for a description of how the covariance matrix is estimated from N-body simulations. Note that we have ignored the errors of the emulator in the covariance matrix used in this analysis since these are sub-dominant for the fiducial cosmology being analysed. For a real data analysis, these would have to be estimated and added to the error budget through the covariance matrix.

Unless otherwise stated we will use the entire range of scales on which the emulator was trained, $0.1\,h^{-1}\,\mathrm{Mpc} \le r \le 150\,h^{-1}\,\mathrm{Mpc}$, to perform inference. Furthermore, although we vary the cosmological parameters $C = \{\Omega_\Lambda, \ln A_{\rm s}, \omega_{\rm c}\}$, we show constraints on the derived parameters most commonly used $C = \{\Omega_{\rm m}, \sigma_8, h\}$. The priors on the cosmological parameters are chosen to be uniform within the range of the sampled latin hyper-cube (Eq. 10); the priors on the HOD parameters are also chosen to be uniform with the ranges shown in Table 3.

6.1 Fiducial constraints

Here, we show that the emulator is capable of recovering the fiducial parameters of the mock catalogue within the 68% confidence interval for all parameters. The resulting 2-D posterior distributions are shown in blue in Fig. 8.

In the same figure, we also show the resulting constraints when the HOD parameters are fixed to their fiducial values (green) and the constraints on the HOD parameters when the cosmological parameters are fixed to their fiducial values (red).

Although taking either of these two steps in a real analysis would underestimate the error on the estimated parameter values, and most likely bias them, this is a useful exercise to determine how much more one could learn by combining the two-point correlation function with other statistics that can constrain the HOD parameters more accurately. For example, Hahn & Villaescusa-Navarro (2021) demonstrated how using the bispectrum could help us to improve constraints on both the cosmological and HOD parameters, by breaking degeneracies between them. Other probes, such as galaxy-galaxy weak lensing (More et al. 2015) can also be used to infer the HOD parameters. Fig. 8 shows that the constraints on $\Omega_{\rm m}$ and σ_8 could be significantly improved by breaking the degeneracies with the HOD parameters.

On the other hand, it is mostly the mass scales $M_{\rm min}$ and M_1 that are better constrained by galaxy clustering when fixing the cosmological parameters. The remaining satellite parameters α and κ do not improve significantly by fixing cosmology. This is probably due to the fact that LOWZ galaxies have a low fraction of satellites, compared with other galaxy selections, and therefore their galaxy two-point correlation function is not very sensitive to these two satellite occupation parameters.

Fig. C1 shows the effect of removing the number density constraint from the likelihood. As previously found in Miyatake et al. (2020), the constraints on cosmological parameters are not strongly affected by the number density term. However, the HOD parameters are sensitive to this change, with the parameters that influence the number of centrals becoming much more poorly constrained when the number density is not used.

6.2 The complementary role of small scales

Here, we study how the constraints vary as a function of the minimum scale included in the likelihood evaluation. This is a test of the performance of our model and its accuracy on small scales, and serves to illustrate the usefulness of small scales in reducing the errors on the recovered parameters. We show the results of this test in Fig. 9.

The small-scale information mainly constrains the fluctuation amplitude, σ_8, as shown in the upper panel of Fig. 9. From $r_{\rm min}$ =

Table 3. The fiducial values and priors of the parameters for mock galaxy surveys that resemble the LOWZ galaxy sample.

	$\bar{z}$	$\bar{n}_{\rm g}\ [(h^{-1}{\rm Mpc})^{-3}]$	$\log M_{\rm min}\ [h^{-1}M_\odot]$	$\sigma_{\log M}$	$\log M_1\ [h^{-1}M_\odot]$	κ	$\alpha_{\rm sat}$
Fiducial	0.251	2.174×10^{-4}	13.62	0.6915	14.42	0.51	0.9168
Min prior	-	-	12	0.1	12	0.01	0.5
Max prior	-	-	14.5	1	16	3	3

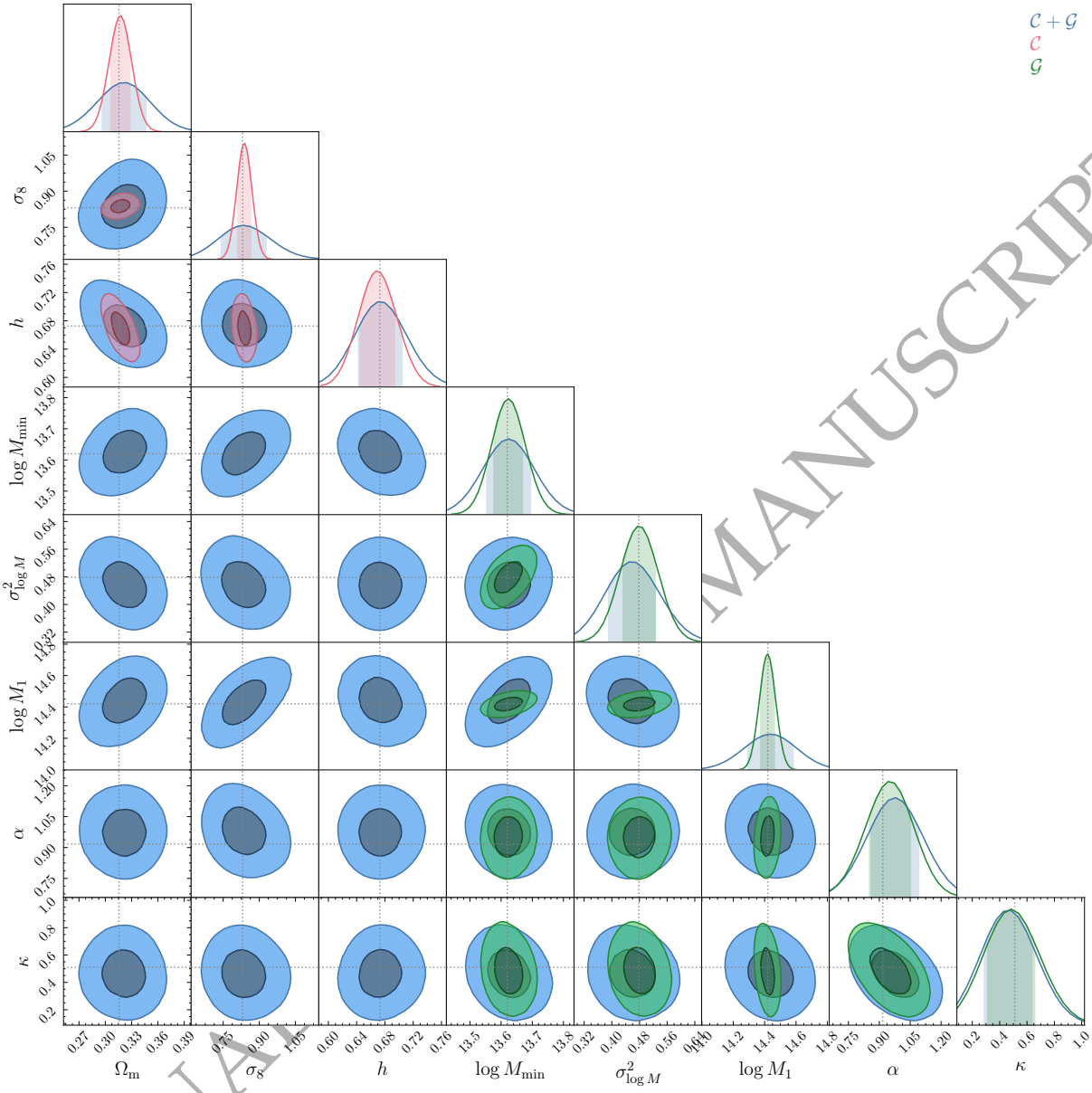

Figure 8. This plot shows that the emulator can recover the true cosmological and HOD parameters within the confidence intervals. We show the posteriors which result when varying both cosmology and HOD parameters ($\mathcal{C}$ and $\mathcal{G}$) (blue, labelled "$\mathcal{C}$ + $\mathcal{G}$") and the cosmological constraints found when the HOD parameters ($\mathcal{C}$) are set to their fiducial values (red, labelled "$\mathcal{C}$"). The constraints on the HOD parameters ($\mathcal{G}$) obtained by fixing the cosmological parameters to their fiducial values are shown in green (labelled "$\mathcal{G}$"). The true values that generated the simulated data are shown by the dotted gray lines.

$1\,h^{-1}$ Mpc to $r_{\rm min} = 5\,h^{-1}$ Mpc, the errorbars on σ_8 increase by a factor of 2.

In the same figure, we also show how the constraints on cosmological parameters would change if we fixed the HOD parameters. Interestingly, the $\Omega_{\rm m}$ constraints would also be improved by including small-scale information by about a factor of 2 if there were no degeneracies with the HOD parameters. The constraints on h are dominated by the BAO scale and therefore do not change noticeably when smaller scales are included or the HOD parameters are fixed.

In the bottom panel of Fig. 9, we show the opposite effect, that of excluding large-scale information. The BAO scale has a very small effect on the recovered value of σ_8, whereas it dominates the constraints on the cosmological parameters $\Omega_{\rm m}$ and h, after marginalising over the HOD parameters. Note that most emulators (Zhai et al. 2019; Yuan et al. 2022) focus on scales smaller than $30\,h^{-1}$ Mpc, and therefore lose constraining power on Ω_m and h.

6.3 The consequences of ignoring assembly bias

We now test whether the halo-connection model used here is flexible enough to obtain unbiased cosmological constraints when modelling the clustering of a sample known to contain assembly bias. Although dark matter halo mass correlates strongly with galaxy clustering, we know that dark matter halos experience different assembly histories even at fixed halo mass, and can display different clustering. These different assembly histories influence secondary properties of halos, and this, in turn, might also affect the formation of galaxies, and hence result in different galactic contents for halos of the same mass.

These effects are known as *halo* and *galaxy assembly bias*. Note that although these two effects share the word bias, they refer to different effects

- *Halo assembly bias* refers to differences in the clustering of dark matter halos at a fixed halo mass. These differences depend on the choice of secondary halo properties, which usually correlate with the formation history of the halo, such as halo concentration or substructure fraction.
- *Galaxy assembly bias* refers to differences in the number of galaxies within dark matter halos at a fixed halo mass, which in turn may depend on secondary halo properties.

Galaxy clustering is shaped by both of these effects. On one hand, halo assembly bias implies that, at fixed halo mass, grouping dark matter halos by a secondary property results in a different clustering signal. On the other hand, the way galaxies occupy dark matter halos might depend on properties other than mass. The combination of both effects determines how strongly galaxy clustering depends on secondary dark-matter halo properties, and therefore how important it is to model this dependency in order to obtain unbiased cosmological constraints.

Here, we want to test how assembly bias affects our constraints when we include effects similar to those observed in hydrodynamical simulations (Hadzhiyska et al. 2021) and semi-analytical models of galaxy formation (Zehavi et al. 2018; Xu et al. 2021; Jiménez et al. 2021) in our mock galaxy catalogues. In this way, we can assess whether the halo model is flexible enough to recover unbiased constraints from realistic galaxy mocks when including small-scale information.

In particular, we implement the assembly bias model based on environment introduced in Xu et al. (2021). The authors showed that the smoothed matter density can account for most of the assembly bias signal observed in a semi-analytic galaxy formation model. This is in agreement with other studies using hydrodynamical simulations (Hadzhiyska et al. 2021).

To create mock galaxy catalogues with an environment-based assembly bias signal, we first determine the local density around each halo. We compute the dark matter density field smoothed with a Gaussian filter over a scale of $2.5\,h^{-1}$Mpc, by first measuring the counts-in-cell dark matter particle density on a 512^3 grid and then multiplying with a Gaussian kernel in Fourier space. The matter overdensity value at the position of each halo is found by interpolating over the 3D grid. Finally, we rank the overdensity values of the halos at fixed halo mass and normalise them to be between 0 and 1. Note that we have computed the ranks inside 50 logarithmically spaced halo mass bins in the range $12 < \log_{10}\left[M_{\rm h}/(h^{-1}M_\odot)\right] < 16$. These ranks, $\delta^{\rm rank}_{2.5}$, are then normalised between 0 and 1 in each halo mass bin.

Once we have determined the ranked environment density around each halo, we assign galaxies to dark matter halos through equations Eq. (14) and Eq. (15), modifying the values of $\log M_{\rm min}$ and $\log M_1$ with the rank of the halo's overdensity value

$$\log_{10} M_{\rm min}(\delta^{\rm rank}_{2.5}) = \log_{10} M^0_{\rm min} + B_{\rm cen} \times \left(\delta^{\rm rank}_{2.5} - 0.5\right), \tag{36}$$

$$\log_{10} M_1(\delta^{\rm rank}_{2.5}) = \log_{10} M^0_1 + B_{\rm sat} \times \left(\delta^{\rm rank}_{2.5} - 0.5\right), \tag{37}$$

where $B_{\rm cen}$ and $B_{\rm sat}$ are the central and satellite assembly bias parameters that control the strength of the effect. Since more galaxies will form in overdense regions, the values of $B_{\rm cen}$ and $B_{\rm sat}$ will be negative.

To explore the possible biases that ignoring assembly bias may introduce in the estimated cosmological parameters, we study two scenarios: i) a weak assembly bias effect with values $B_{\rm cen} = -0.1$ and $B_{\rm sat} = -0.2$, and ii) a strong one with values $B_{\rm cen} = -0.2$ and $B_{\rm sat} = -0.4$. The weak assembly bias parameters have been chosen to mimic the level of assembly bias signal found in Xu et al. (2021) for a sample with a galaxy number density of $n_{\rm gal} = 0.01\left(h^{-1}{\rm Mpc}\right)^{-3}$. In Fig. D1, we show that the weak scenario produces changes in the two-point correlation function of up to 10 per cent compared with the case with no assmebly bias, while the strong case increases the clustering by up to 20 per cent.

Fig. 10 shows the constraints obtained using our model (which ignores assembly bias) to fit the clustering measured from the mock galaxy samples described above, with weak and strong assembly bias. In both the weak and strong assembly bias scenarios, we can robustly recover the cosmological parameters $\Omega_{\rm m}$ and h since they are mostly determined by the BAO scale. However, σ_8 is biased towards smaller values in both scenarios. In the strong assembly bias case, this shift is more than $1-\sigma$ away from its true value. However, we note that the strong assembly bias scenario is unrealistic for a LOWZ-like sample of galaxies (Yuan et al. 2022).

Fig. D2 shows the full 2D posterior, including the HOD parameters that have shifted in the expected direction. Intuitively, the environmt assembly bias effect leads to more galaxies forming in overdense regions (thus, the assembly bias parameters are negative). The left hand side of Fig. D1 shows that higher number densities in the assembly bias mocks correspond to a higher mean number of galaxies, that could be effectively reproduced by lowering $M_{\rm min}$.

Fig. 11 shows how the constraints on σ_8 change as we vary the minimum scale included in the determination of the likelihood. If we restrict the analysis to scales larger than $10\ h^{-1}$ Mpc, the halo model recovers unbiased cosmological constraints by biasing the HOD parameters. However, on scales smaller than $10\,h^{-1}$ Mpc, when

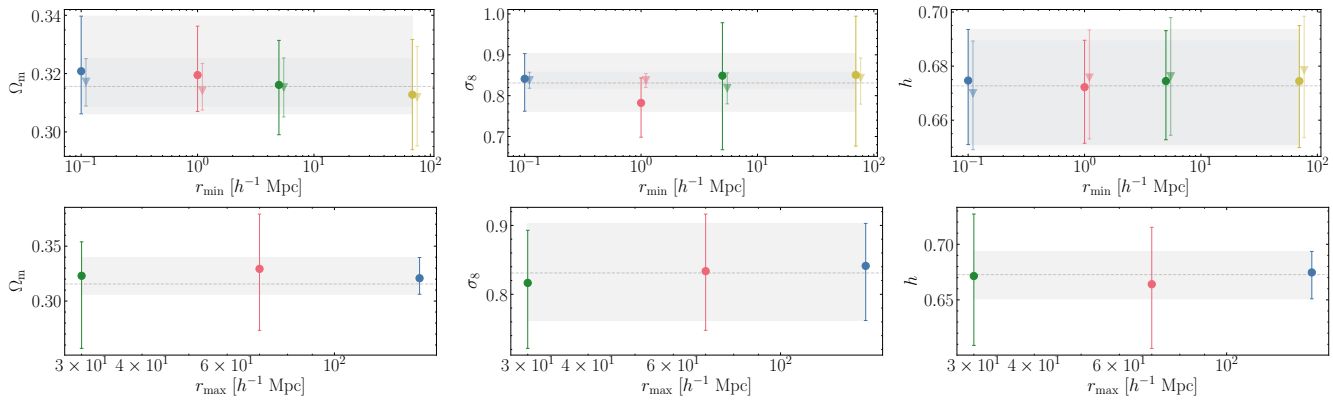

Figure 9. We show the estimated maximum likelihood parameters, together with their estimated uncertainties, for varying minimum and maximum pair separation scales used in the analysis. In the top panel we show both the cosmological constraints obtained when marginalizing over the HOD parameters (circles) and when fixing the HOD parameters to their fiducial values (triangles). This shows that the constraints on the cosmological parameters improve as more non-linear scales are included for all parameters but h, whose constraints are dominated by the BAO information.

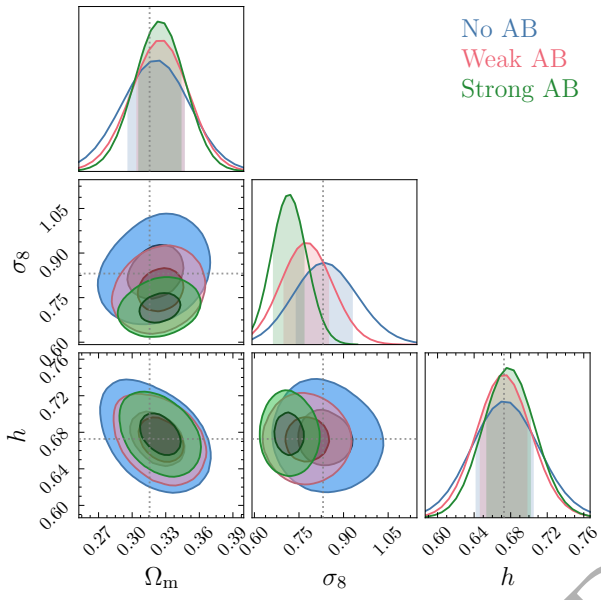

Figure 10. Constraints obtained when fitting mock catalogues that include the environment-based assembly bias model presented in Xu et al. (2021) with our halo model emulator, which ignores the effect of assembly bias. The cosmological parameters Ω_m and h can still be recovered within the estimated confidence intervals, since they are mainly constrained by the BAO peak, whereas σ_8 shows a small bias towards smaller values in both the weak and strong assembly bias scenarios.

the constraining power on σ_8 doubles, lowering the mass of halos that host a central cannot mimic the effects shown in Fig.D1, and σ_8 needs to be lowered to describe the changes around the one to two halo term transition.

We can monitor the evidence of the model to detect whether the halo-galaxy connection model has been mispecified. The evidence is defined as

$$P(\mathcal{D}) = \int \mathrm{d}\theta\, \mathrm{P}(\mathcal{D}|\theta)\mathrm{P}(\theta), \tag{38}$$

and can be interpreted as the likelihood of the data given the model. The values of the evidence estimated by nested sampling are 20.87 for mocks without assembly bias, 18.34 for those with a weak assembly bias signal, and 16.37 for those with a strong assembly bias effect.

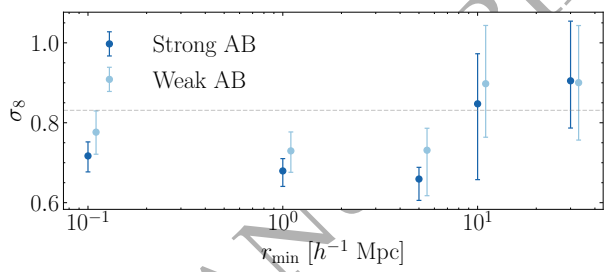

Figure 11. Inferred values of σ_8 and their estimated uncertainties as a function of the minimum scale, r_{min}, used in the likelihood analysis. This plot shows the systematic introduced by assembly bias can only be removed by excluding the small scale information.

Given the importance of unbiased constraints on σ_8 to resolve the σ_8–S_8 tension, we will work on adding environment-based assembly bias to our emulator for its application to DESI Y1 data.

6.4 Comparison with Lagrangian Perturbation Theory

In this section, we compare the emulator constraints with those obtained by 1-loop Lagrangian perturbation theory (Chen et al. 2020, 2021) using the publicly available code VELOCILEPTORS[2]. We fit the bias parameters b_1, b_2, and b_s, together with the cosmological parameters. We find that a LOWZ-like sample in real space cannot constrain the one-loop effective field theory counter-terms and therefore we set them to zero.

In Fig. 12, we show how the emulator can obtain constraints similar to LPT when analysed over the same scale range, even after marginalising the halo-galaxy connection parameters, which are in total 6 free parameters (compared to only 3 for LPT). Note that the LPT predictions are slightly biased in σ_8, this is due to the strong degeneracy between b_1 and σ_8 that is accentuated in real space. In such a situation, the 1-D marginalized posterior for σ_8 can depend strongly on the prior or the parameterisation of the nuisance parameters, potentially leading to a biased estimate (Sugiyama et al. 2020). The biased estimate of σ_8 tends to be alleviated by including more

[2] https://github.com/sfschen/velocileptors

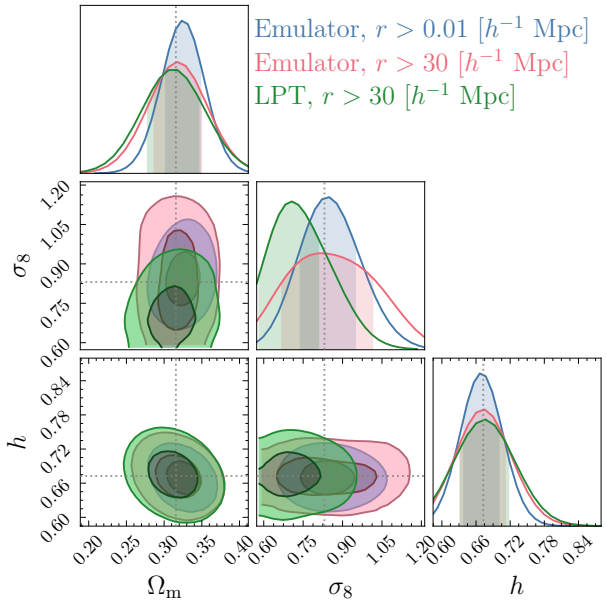

Figure 12. Comparison of the constraints obtained by the emulator based model to the 1 loop Perturbation Theory model presented in Chen et al. (2020, 2021).

information, e.g., redshift space distortions. As shown in Fig. 12, including small scale information does allow the emulator to constrain the parameters more accurately.

7 DISCUSSION AND CONCLUSION

We show that after marginalizing over uncertainties in the galaxy-halo connection parameters, an emulator of the real space correlation function based on the halo model can obtain tighter constraints on the cosmological parameters than Lagrangian Perturbation Theory (LPT) given that the latter cannot extract the additional information contained in small scale galaxy clustering.

The treatment of galaxy bias in both approaches is very different. On the one hand, the bias treatment of LPT is based on expanding the galaxy number density perturbation, $\delta_g(\boldsymbol{x})$, in terms of a series of local operators (Desjacques et al. 2018), which are meant to capture the effect of the large-scale environment on the formation and evolution of galaxies. Each operator is associated with a free coefficient, called bias parameter, which depends on the selected population of galaxies and needs to be fitted to the data. The number and type of operators up to a given order in perturbation theory can be fully determined by symmetry considerations (McDonald & Roy 2009; Chan et al. 2012; Assassi et al. 2014; Senatore 2015; Mirbabayi et al. 2015; Desjacques et al. 2018; Eggemeier et al. 2019), which guarantees that within its regime of validity the perturbative bias expansion can model any galaxy-matter connections (including assembly bias). On the other hand, the HOD approach implemented in this paper has the advantage that it can be extended further into the non-linear regime compared to the perturbative expansion, but is restricted by the assumption one makes about the halo properties that determine halo clustering and galaxy occupations. More work is needed to determine the robustness of both approaches against uncertainties in the model connecting halos to galaxies as well as their constraining power on cosmological parameters. In the future, we plan to compare the constraints obtained with both models using large hydrodynamic simulations or semi-analytic models of galaxy formation.

Regarding the emulation approach, we have combined an emulator trained in halo properties with an analytical prescription of how galaxies populate halos, as already done by Nishimichi et al. (2019). Most other emulators, however, are trained on HOD catalogues built on N-body simulations (Zhai et al. 2019; Yuan et al. 2022). Our approach has advantages and disadvantages. In particular, the halo model allows us to reduce emulator errors through an analytical galaxy-halo connection, which also simplifies the task for the emulator that only needs to learn the dependency of halo clustering on cosmological parameters. Moreover, the analytical model allows us to compute different observables, such as the galaxy-cluster cross-correlation function or a multitracer two-point correlation function. Obtaining cosmological information from small scales through these observables will be the subject of future work. It also allows us to combine emulators trained on simulations with different resolutions to reduce cosmic variance on large scales and perform an analysis using the full-shape of the correlation function.

Regarding the disadvantages of our approach, extending the halo model approach to arbitrary statistics could potentially be difficult. The emulation of statistics such as the bispectrum, would be simplified if one were to follow the procedure outlined in Zhai et al. (2019); Yuan et al. (2022). Moreover, more work needs to be done in order to go beyond the vanilla HOD model used in this work to introduce effects such as the environment-based assembly bias shown in Section 6.3. In the future, we plan to introduce a correction based on binning the halo two-point correlation function in terms of halo environment.

We have shown that including environment-based assembly bias in the model is important to avoid biased constraints on σ_8. This is especially relevant given the $f\sigma_8$ tension. Previously, Kobayashi et al. (2022) and Miyatake et al. (2020) had performed tests similar to the one presented in Sec. 6.3 to emulators based also on the halo model. Kobayashi et al. (2022) studied the effect that ignoring concentration-based assembly bias would have on the cosmological parameters inferred when emulating the redshift space power spectrum through the halo model. They found that although the mock galaxies show 10 – 20 per cent higher amplitudes than the mocks without assembly bias, they can still recover unbiased cosmological constraints through a change in the HOD parameters. In contrast, Miyatake et al. (2020) found that the same effects of assembly bias would introduce biases in Ω_m and σ_8 when the data vector is a combination of the projected two-point correlation function of galaxies and galaxy-galaxy lensing. In this case, the fact that one can use galaxy-galaxy lensing to accurately determine the scaling of halo bias with halo mass restricts the flexibility of the HOD model, which is not able to adapt the parameters in such a way that unbiased constraints can be recovered.

We have here explored an assembly bias model inspired by semi-analytic methods of galaxy formation and hydrodynamical simulations. In fact, these studies find that the magnitude of concentration-based assembly bias is small. Ignoring environment-based assembly bias in the theory model, we find that the halo model is not flexible enough to obtain unbiased cosmological constraints already when the effect of assembly bias only impacts clustering by about 10%. Moreover, we find that including the BAO scale allows us to obtain robust constraints on Ω_m.

To summarise, we have

- Presented a neural network which models the full-shape galaxy clustering in real space based on the halo model, which is more accurate and faster than previously published Gaussian process emu-

lators Nishimichi et al. (2019), when trained on the same dataset. The method presented here can produce a galaxy correlation function in less than 300 ms on a single core.

- Shown that small scale galaxy clustering ($r < 5\ h^{-1}$ Mpc) in real space improves the constraints on σ_8 by a factor of 2, whereas marginalising over the HOD parameters erases the information contained on small scales for $\Omega_{\rm m}$.
- Shown that a halo model that ignores effects of environment-based assembly bias similar to those observed in hydrodynamic simulations and semianalytic models of galaxy formation could introduce bias in the inferred σ_8, while the BAO peak ensures that we can recover $\Omega_{\rm m}$ and h robustly.
- Found that the above-mentioned bias in the value of inferred σ_8 disappears when analysing scales larger than 10 h^{-1} Mpc.

In the second paper of this series, we will present analogous neural network emulators of the pairwise velocity moments that will be used to i) perform the real to redshift space mapping to predict the cosmological dependence of redshift-space galaxy clustering, and ii) constrain observations of the peculiar velocity field.

In the future, we also plan to use the neural network emulators on DESI Y1 data to constrain the cosmological parameters. This requires that the models be trained on simulations with lower particle mass so that they can reach the high galaxy number densities that DESI will measure. For this, a new simulation campaign, Dark Quest II., is currently ongoing to cover a wider mass range (down to a few $10^{11}\ h^{-1} M_\odot$) in an extended cosmological model space including massive neutrinos, time-varying dark energy equation-of-state parameter and spatial curvature using a newly developed fast N-body code (Nishimichi *et al.* in prep.).

ACKNOWLEDGEMENTS

CC would like to acknowledge support from the Science Technology Facilities Council through a Centre for Doctoral Training in Data Intensive Science studentship (ST/P006744/1). AE is supported at the AIfA by an Argelander Fellowship. SB is supported by the UK Research and Innovation (UKRI) Future Leaders Fellowship [grant number MR/V023381/1]. This work was also supported by STFC grant ST/T000244/1. This work used the DiRAC@Durham facility managed by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk). The equipment was funded by BEIS capital funding via STFC capital grants ST/K00042X/1, ST/P002293/1, ST/R002371/1 and ST/S002502/1, Durham University and STFC operations grant ST/R000832/1. DiRAC is part of the National eInfrastructure. This work was also supported in part by MEXT/JSPS KAKENHI Grant Number JP19H00677, JP20H05861, JP21H01081 and JP22K03634. We also acknowledge financial support from Japan Science and Technology Agency (JST) AIP Acceleration Research Grant Number JP20317829.

DATA AVAILABILITY

The data shown in this article will be shared on reasonable request to the corresponding author.

REFERENCES

Abbott T. M. C., et al., 2022, Phys. Rev. D, 105, 023520
Abdalla E., et al., 2022, Journal of High Energy Astrophysics, 34, 49
Agarap A. F., 2018, arXiv preprint arXiv:1803.08375
Alcock C., Paczynski B., 1979, Nature, 281, 358
Armijo J., Baugh C. M., Padilla N. D., Norberg P., Arnold C., 2022, MNRAS, 510, 29
Assassi V., Baumann D., Green D., Zaldarriaga M., 2014, J. Cosmology Astropart. Phys., 2014, 056
Ba S., Myers W. R., Brenneman W. A., 2015, Technometrics, 57, 479
Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013, ApJ, 762, 109
Benson A. J., Cole S., Frenk C. S., Baugh C. M., Lacey C. G., 2000, Monthly Notices of the Royal Astronomical Society, 311, 793
Bose S., Eisenstein D. J., Hernquist L., Pillepich A., Nelson D., Marinacci F., Springel V., Vogelsberger M., 2019, MNRAS, 490, 5693
Buchner J., et al., 2014, A&A, 564, A125
Calafut V., et al., 2021, Phys. Rev. D, 104, 043502
Carlson J., White M., Padmanabhan N., 2009, Phys. Rev. D, 80, 043531
Chan K. C., Scoccimarro R., Sheth R. K., 2012, Phys. Rev. D, 85, 083509
Chen S.-F., Vlah Z., White M., 2020, J. Cosmology Astropart. Phys., 2020, 062
Chen S.-F., Vlah Z., Castorina E., White M., 2021, J. Cosmology Astropart. Phys., 2021, 100
Cooray A., Sheth R., 2002, Physics Reports, 372, 1
Crocce M., Scoccimarro R., 2006, Phys. Rev. D, 73, 063520
Crocce M., Pueblas S., Scoccimarro R., 2006, MNRAS, 373, 369
Cuesta-Lazaro C., Li B., Eggemeier A., Zarrouk P., Baugh C. M., Nishimichi T., Takada M., 2020, MNRAS, 498, 1175
DESI Collaboration et al., 2016, arXiv e-prints, p. arXiv:1611.00036
Dawson K. S., et al., 2013, AJ, 145, 10
Dawson K. S., et al., 2016, AJ, 151, 44
DeRose J., et al., 2019, ApJ, 875, 69
Desjacques V., Jeong D., Schmidt F., 2018, Physics Reports, 733, 1
Diemer B., Joyce M., 2019, The Astrophysical Journal, 871, 168
Duane S., Kennedy A. D., Pendleton B. J., Roweth D., 1987, Phys. Lett. B, 195, 216
Dupuy A., Courtois H. M., Kubik B., 2019, MNRAS, 486, 440
Eggemeier A., Scoccimarro R., Smith R. E., 2019, Phys. Rev. D, 99, 123514
Elsner F., Schmidt F., Jasche J., Lavaux G., Nguyen N.-M., 2020, J. Cosmology Astropart. Phys., 2020, 029
Fisher K. B., 1995, ApJ, 448, 494
Gao L., White S. D. M., 2007, MNRAS, 377, L5
Gao L., Springel V., White S. D. M., 2005, MNRAS, 363, L66
Gómez J. S., Padilla N. D., Helly J. C., Lacey C. G., Baugh C. M., Lagos C. D. P., 2022, MNRAS, 510, 5500
Grove C., et al., 2021, arXiv e-prints, p. arXiv:2112.09138
Hadzhiyska B., Liu S., Somerville R. S., Gabrielpillai A., Bose S., Eisenstein D., Hernquist L., 2021, MNRAS, 508, 698
Hahn C., Villaescusa-Navarro F., 2021, Journal of Cosmology and Astroparticle Physics, 2021, 029
Hahn C., Villaescusa-Navarro F., Castorina E., Scoccimarro R., 2020, J. Cosmology Astropart. Phys., 2020, 040
Hendrycks D., Gimpel K., 2016, arXiv e-prints, p. arXiv:1606.08415
Hikage C., et al., 2019, PASJ, 71, 43
Jiménez E., Padilla N., Contreras S., Zehavi I., Baugh C. M., Orsi Á., 2021, MNRAS, 506, 3155
Joudaki S., et al., 2016, MNRAS, 465, 2033
Kim A. G., Linder E. V., 2020, Physical Review D, 101
Kingma D. P., Ba J., 2014, arXiv e-prints, p. arXiv:1412.6980
Kobayashi Y., Nishimichi T., Takada M., Takahashi R., 2020a, Phys. Rev. D, 101, 023510
Kobayashi Y., Nishimichi T., Takada M., Takahashi R., Osato K., 2020b, Phys. Rev. D, 102, 063504
Kobayashi Y., Nishimichi T., Takada M., Miyatake H., 2022, Phys. Rev. D, 105, 083517
Kuhlen M., Vogelsberger M., Angulo R., 2012, Phys. Dark Univ., 1, 50
Lange J. U., van den Bosch F. C., Zentner A. R., Wang K., Hearin A. P., Guo H., 2019, MNRAS, 490, 1870
Lange J. U., Hearin A. P., Leauthaud A., van den Bosch F. C., Guo H., DeRose J., 2021, Monthly Notices of the Royal Astronomical Society, 509, 1779

Laureijs R., et al., 2011, arXiv e-prints, p. arXiv:1110.3193
Leclercq F., Heavens A., 2021, MNRAS, 506, L85
Ludlow A. D., Bose S., Angulo R. E., Wang L., Hellwing W. A., Navarro J. F., Cole S., Frenk C. S., 2016, MNRAS, 460, 1214
Maksimova N. A., Garrison L. H., Eisenstein D. J., Hadzhiyska B., Bose S., Satterthwaite T. P., 2021, MNRAS, 508, 4017
McDonald P., Roy A., 2009, J. Cosmology Astropart. Phys., 2009, 020
McDonald P., Seljak U., 2009, Journal of Cosmology and Astroparticle Physics, 2009, 007
Mirbabayi M., Schmidt F., Zaldarriaga M., 2015, J. Cosmology Astropart. Phys., 2015, 030
Miyatake H., et al., 2020, arXiv e-prints, p. arXiv:2101.00113
Miyatake H., et al., 2021, arXiv e-prints, p. arXiv:2111.02419
More S., Miyatake H., Mandelbaum R., Takada M., Spergel D. N., Brownstein J. R., Schneider D. P., 2015, The Astrophysical Journal, 806, 2
Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493
Nishimichi T., et al., 2019, Astrophys. J., 884, 29
Paillas E., Cai Y.-C., Padilla N., Sánchez A. G., 2021, MNRAS, 505, 5731
Peebles P. J. E., 1980, The large-scale structure of the universe
Percival W. J., Friedrich O., Sellentin E., Heavens A., 2021, Monthly Notices of the Royal Astronomical Society, 510, 3207
Philcox O. H. E., Ivanov M. M., 2022, Phys. Rev. D, 105, 043517
Planck Collaboration et al., 2016, A&A, 594, A13
Rasmussen C. E., Williams C. K. I., 2005, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press
Scoccimarro R., 2004, Phys. Rev. D, 70, 083007
Senatore L., 2015, J. Cosmology Astropart. Phys., 2015, 007
Springel V., 2005, MNRAS, 364, 1105
Sugiyama S., Takada M., Kobayashi Y., Miyatake H., Shirasaki M., Nishimichi T., Park Y., 2020, Phys. Rev. D, 102, 083520
Sunyaev R. A., Zeldovich Y. B., 1980, MNRAS, 190, 413
Takada M., et al., 2014, Publications of the Astronomical Society of Japan, 66
Tinker J., Kravtsov A. V., Klypin A., Abazajian K., Warren M., Yepes G., Gottlöber S., Holz D. E., 2008, ApJ, 688, 709
Valogiannis G., Dvorkin C., 2022, Phys. Rev. D, 105, 103534
Vlah Z., Seljak U. c. v., Baldauf T., 2015, Phys. Rev. D, 91, 023508
Warren M. S., Abazajian K., Holz D. E., Teodoro L., 2006, ApJ, 646, 881
Wechsler R. H., Tinker J. L., 2018, ARA&A, 56, 435
Xu X., Zehavi I., Contreras S., 2021, MNRAS, 502, 3242
Yuan S., Garrison L. H., Eisenstein D. J., Wechsler R. H., 2022, arXiv e-prints, p. arXiv:2203.11963
Zehavi I., Contreras S., Padilla N., Smith N. J., Baugh C. M., Norberg P., 2018, ApJ, 853, 84
Zhai Z., et al., 2019, ApJ, 874, 95
Zhai Z., et al., 2022, arXiv e-prints, p. arXiv:2203.08999
Zheng Z., et al., 2005, ApJ, 633, 791

APPENDIX A: EVALUATION OF THE EMULATORS AS A FUNCTION OF REDSHIFT AND NUMBER DENSITY

In this appendix we show detailled evaluations of the halo auto-correlation emulator (Fig. A1) and the galaxy auto-correlation emulator (Fig. A2).

For halo auto-correlations, we find that the emulator accuracy decreaes for lower number densities, which are more affected by shot noise, whereas it decreases for high redshifts ($z = 1.5$).

For galaxy auto-correlations we do not find any substantial biases for neither redshift or galaxy number density.

APPENDIX B: ESTIMATING THE COVARIANCE MATRIX

In Section 6, we used an estimate of the covariance matrix to obtain the posterior of cosmological parameters given a mock data vector. The covariance matrix was estimated from a set of 1600 N-body simulations part of the AbacusSummit suite (Maksimova et al. 2021). These are high resolution small boxsize simulations ($L_{\rm box} = 500\ h^{-1}$ Mpc).

Given the small boxsize of the simulations, we re-scale the covariance by a factor of $0.5^3/0.67$ to estimate the expected errors for a LOWZ-like sample, whose effective volume is 0.67 $(h^{-1}\ {\rm Gpc})^3$. We also correct the covariance estimated from the mocks with Eq. 56 in Percival et al. (2021).

APPENDIX C: THE EFFECT OF CONSTRAINING GALAXY NUMBER DENSITY IN THE LIKELIHOOD ANALYSIS

In this appendix, we show the effect of removing the galaxy number density term in Eq. 35.

Fig. C1 shows that the number density constrain does not change the constraints on cosmological parameters noticeably, whereas it mainly improves those of the HOD parameters. In particular, it breaks the degeneracy between the central occupation parameters, $\log M_{\rm min}$ and $\sigma_{\log M}$.

APPENDIX D: ASSEMBLY BIAS MOCKS DETAILS

Here, we describe here the occupation variations of the environment-based assembly bias mocks used in Section 6.3.

Fig. D1 shows how the mean number of centrals and satellites change as a function of halo mass and halo environment, δ_s, for both the strong and weak assembly bias mocks. At fixed halo mass, halos residing in denser environments will have a higher mean number of galaxies (both centrals and satellites) than those occupying underdense regions.

On the right hand side of Fig. D1 we also show the ratio of the galaxy two-point correlation function with a strong and weak assembly bias signal to that of the no assembly bias case. The deviations can be as large as 10% for the weak case, and 20% for the strong one.

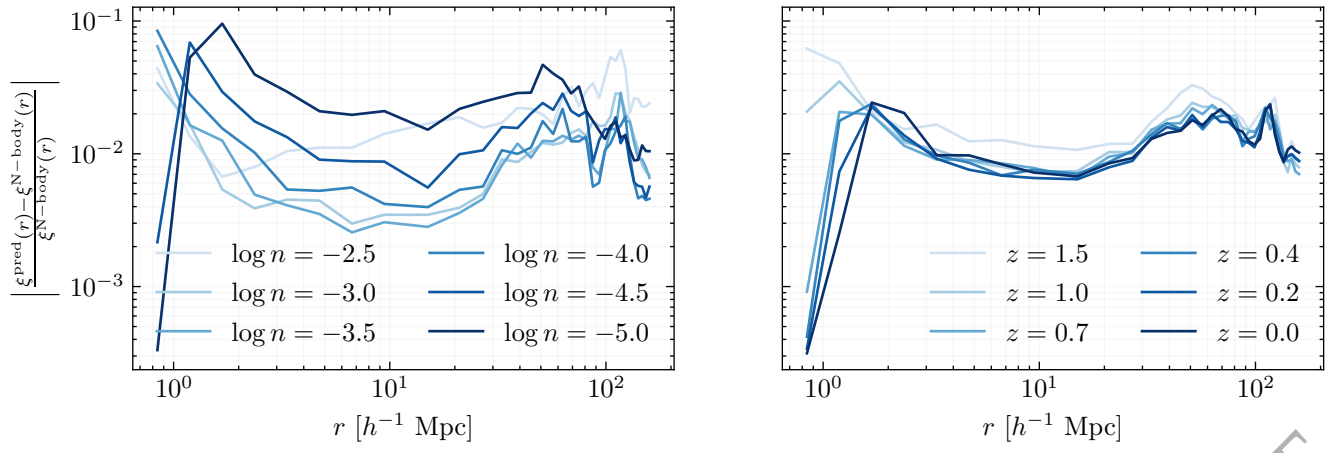

Figure A1. Median absolute errors of the halo two-point correlation function as a function of number density (left), averaged over redshift and test set cosmologies, and as a function of redshift (right), averaged over number density and test set cosmologies.

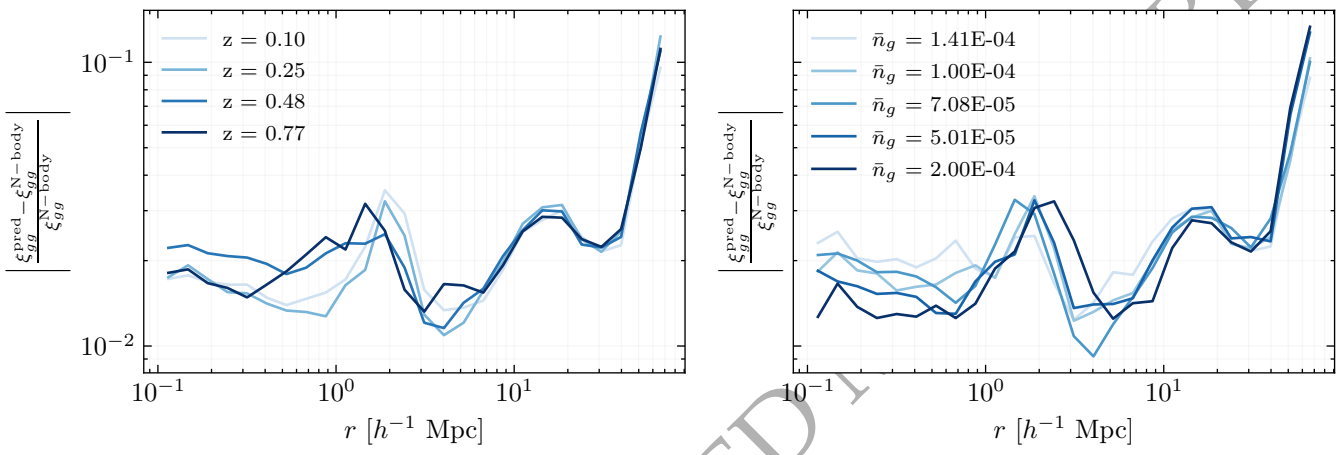

Figure A2. Median absolute errors of the galaxy two-point correlation function as a function of number redshift (left), averaged over galaxy number density and test set cosmologies, and as a function of galaxy number density (right), averaged over number redshift and test set cosmologies. In both cases the emulator accuracy does not show noticeable biases.

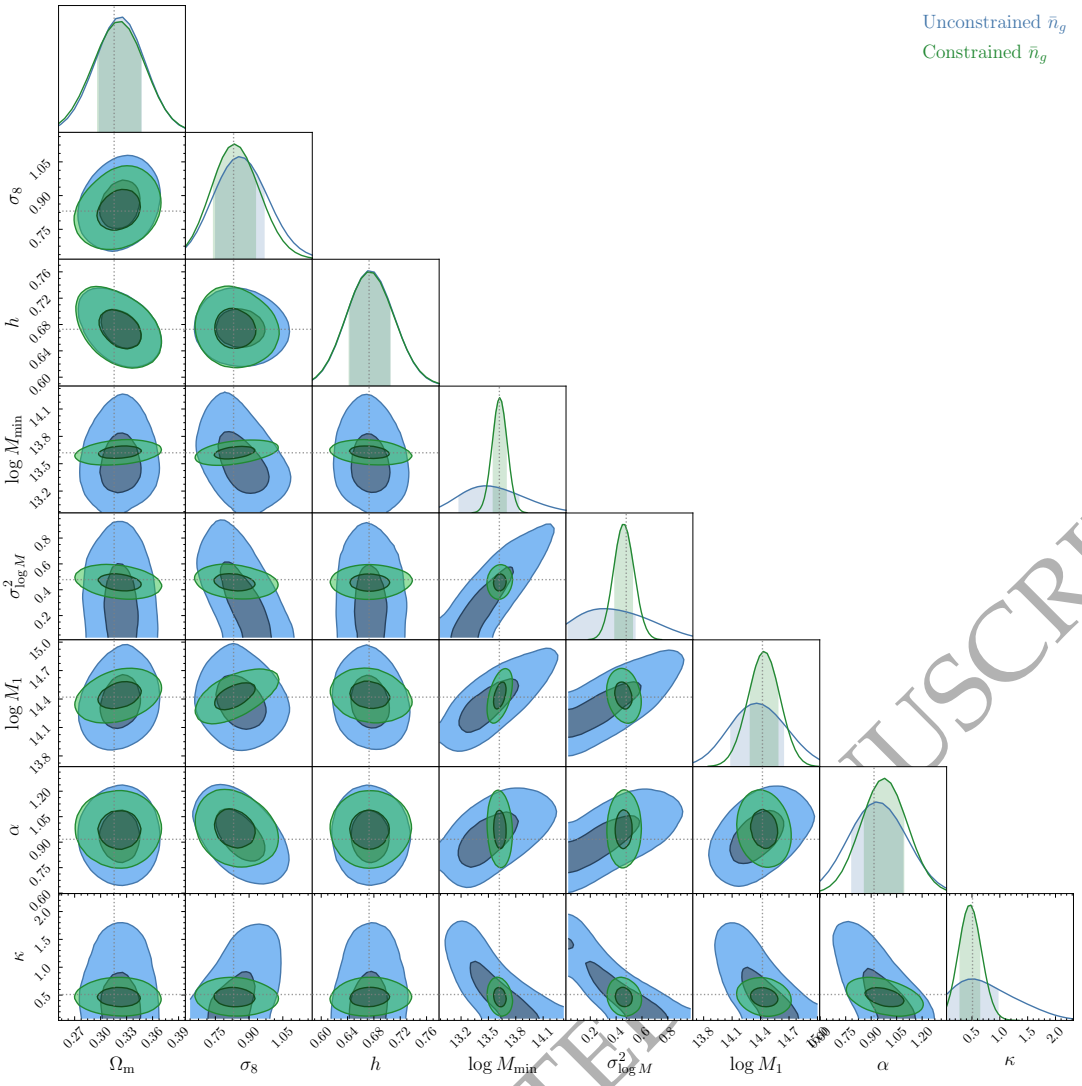

Figure C1. Comparison of constraints on cosmological and HOD parameters when the galaxy number density is included in the likelihood (Constrained $\bar{n}_g$) and when it isn't (Unconstrained $\bar{n}_g$. Including number density constraints only helps determine the HOD parameters with a higher accuracy.

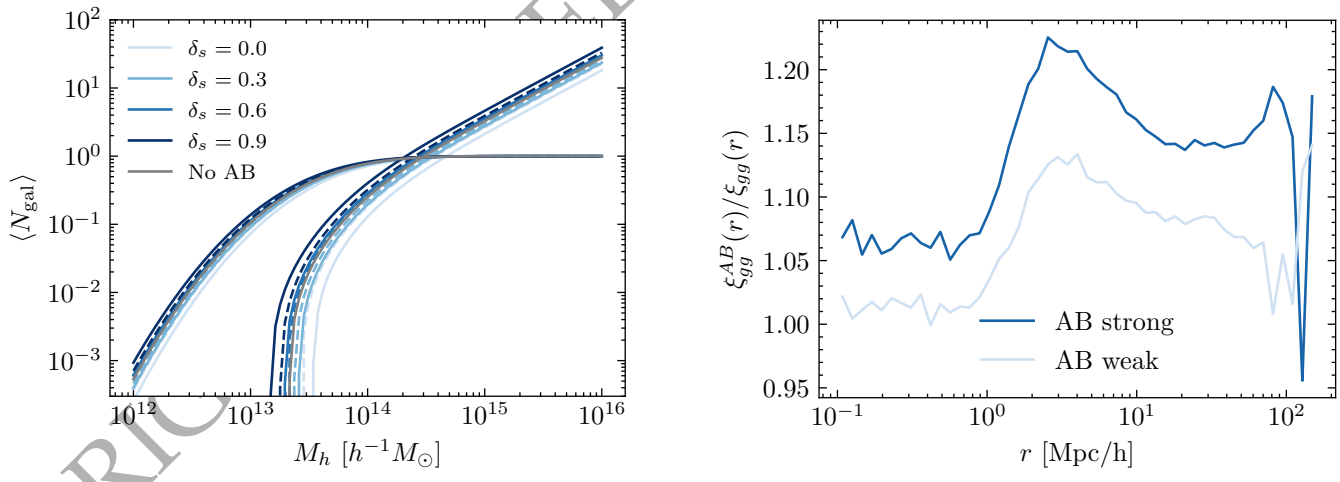

Figure D1. Details of the assembly bias mocks. On the left, we show the mean number of central and satellite galaxies as a function of halos mass and halo environment, δ_s, for both the strong assembly bias model (solid lines) and the weak assembly bias model (dashed lines). On the right, we show the ratios of the galaxy two-point correlation functions for assembly bias models, and their non-assembly bias counterpart.

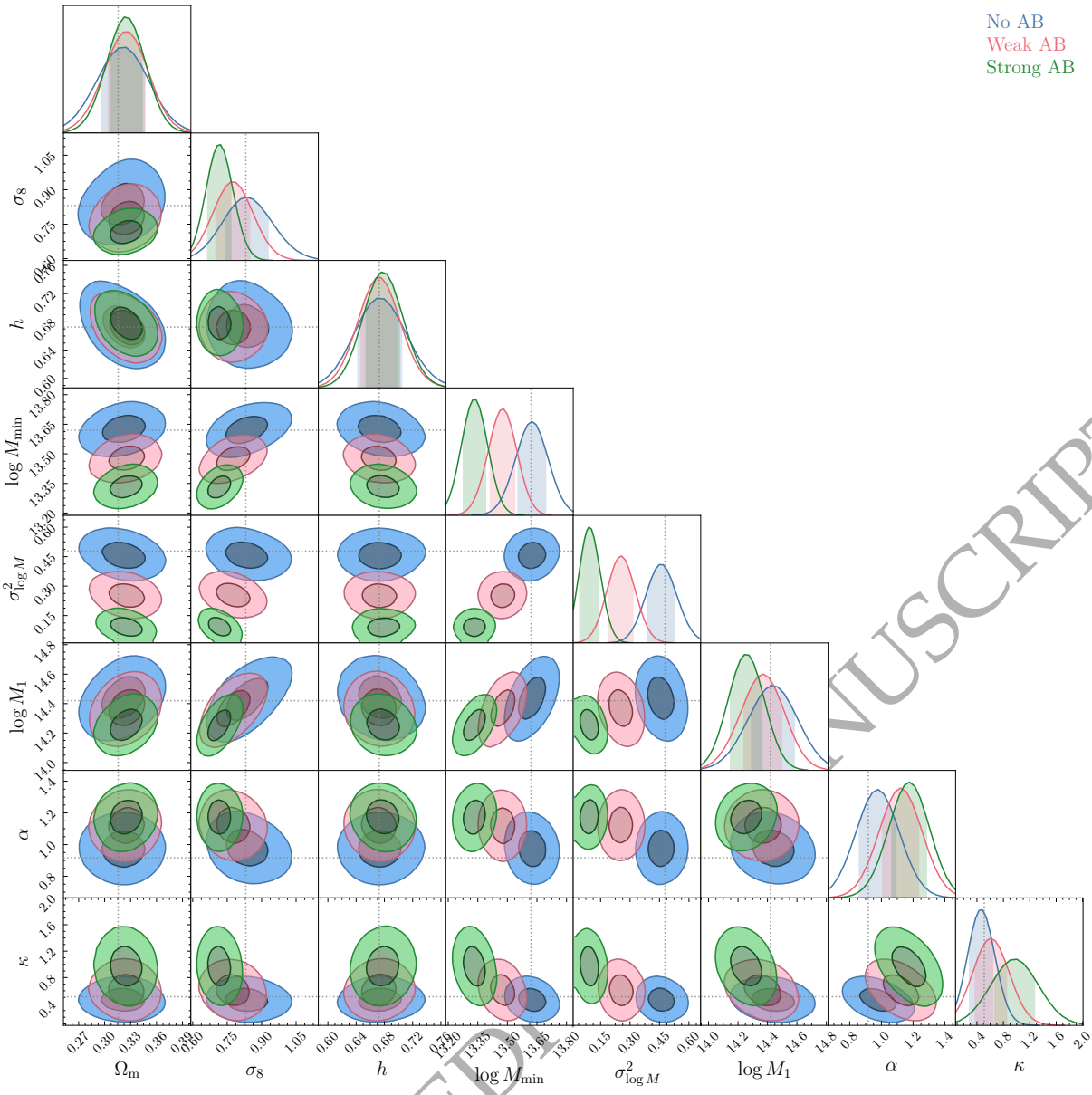

Figure D2. Full 2D posteriors obtained for data with i) no assembly bias effect, ii) a weak assembly bias signal, and iii) a strong assembly bias signal.