



A Nonparametric Method for the Inference of Halo Occupation Distributions

Jacob Kennedy¹ , Eric Gawiser^{1,2} , Kartheik G. Iyer^{3,4} , and L. Y. Aaron Yung⁵

¹ Department of Physics and Astronomy, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA; jnk97@physics.rutgers.edu

² School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540, USA

³ Columbia Astrophysics Laboratory, Columbia University, 550 West 120th Street, New York, NY 10027, USA

⁴ Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA

⁵ Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA

Received 2025 August 18; revised 2026 March 24; accepted 2026 March 27; published 2026 May 29

Abstract

The galaxy–halo connection traces processes by which galaxies form and evolve. The halo occupation distribution (HOD) describes the relationship between galaxies and their host dark matter haloes. Measurements of the galaxy two-point correlation function (2PCF) allow us to extract information about the HODs of observed galaxy samples. Several parametric HOD models have been proposed in the literature, but the choice of parameterization restricts the space of possible HODs. To resolve this issue, we introduce a nonparametric HOD fitting method in which we train an emulator to learn the mappings among the galaxy 2PCF, physical properties used to select galaxy samples, and the HOD, all obtained from simulated past light cones constructed with the Santa Cruz semianalytic model. Implementing this emulator within a likelihood analysis framework, we derive constraints on the HOD of a galaxy sample when provided with a measurement of its 2PCF. Using the emulator to accelerate likelihood evaluations, we test the nonparametric HOD approach on a set of 2PCFs for mock galaxy samples drawn from the TNG100-1 simulation and selected above threshold values of stellar mass and star formation rate. Our framework is able to recover TNG100-1 HODs within 0.2 dex. We use the TNG100-1 mocks to tune the reported uncertainties to estimate those expected in the analysis of observations. Comparing to parametric HOD modelling routines applied to the same mock galaxy samples, our approach consistently infers the HOD with comparable or greater precision and accuracy.

Unified Astronomy Thesaurus concepts: [Large-scale structure of the Universe \(902\)](#); [Clustering \(1908\)](#); [Two-point correlation function \(1951\)](#); [Galaxy dark matter halos \(1880\)](#); [Computational methods \(1965\)](#); [Cosmology \(343\)](#)

1. Introduction

The large-scale structure of the Universe describes how galaxies are distributed on distance scales greater than a single galaxy. While the distribution of galaxies appears homogeneous on very large distance scales (e.g., M. Tegmark et al. 2004; D. W. Hogg et al. 2005; J. Yadav et al. 2005), deviations from this homogeneity on smaller scales encode a richness of information regarding the cosmological parameters of the Universe and the physics of galaxy formation and evolution. Measuring the statistical properties of the spatial distribution of galaxies has been a key focus of astronomical observations for the past several decades. Large-scale galaxy surveys (e.g., the Sloan Digital Sky Survey (SDSS): D. G. York et al. 2000; Two Degree Field Galaxy Redshift Survey: M. Colless et al. 2001; Kilo-Degree Survey: J. T. A. de Jong et al. 2012; DEEP2 Galaxy Redshift Survey: J. A. Newman et al. 2013; the Dark Energy Survey (DES): Dark Energy Survey Collaboration et al. 2016) have provided a highly informative test bed for candidate cosmological and galaxy formation theories.

The statistic of merit in these analyses is often the galaxy two-point correlation function (2PCF; P. J. E. Peebles 1980), which measures the degree of galaxy clustering as a function of scale. Cosmological inference from galaxy clustering is based on the assumption that galaxies trace the underlying

dark matter distribution, which can be specified from initial conditions and simulated reliably and efficiently over large volumes for a given cosmological model (K. N. Abazajian et al. 2009). The prevailing cosmological paradigm predicts that all galaxies form and evolve within dark matter haloes. In this picture, haloes are defined as roughly spherical, virialized regions with a density ≈ 200 times the average background density of the Universe (e.g., S. Cole & C. Lacey 1996; V. R. Eke et al. 1996; J. F. Navarro et al. 1996).

One approach to modeling the galaxy–halo connection that has gained considerable traction is the halo occupation distribution (HOD; e.g., C.-P. Ma & J. N. Fry 2000; J. A. Peacock & R. E. Smith 2000; U. Seljak 2000; R. Scoccimarro et al. 2001; A. A. Berlind & D. H. Weinberg 2002; A. A. Berlind et al. 2003; Z. Zheng et al. 2005). The HOD framework circumvents the messy physics of baryons by establishing an empirical relation between galaxies and haloes that can be constrained from measurements of galaxy clustering. Further, HODs provide a computationally inexpensive method of “painting” galaxies onto N -body simulations compared to the “full physics” approach of hydrodynamic simulations (B. Hadzhiyska et al. 2021). This becomes especially relevant when forecasting the large cosmological volumes of current and next-generation galaxy surveys conducted with the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al. 2016), Euclid (Euclid Collaboration et al. 2025), Vera C. Rubin Observatory (Rubin; Ž. Ivezić et al. 2019), and the Nancy Grace Roman Space Telescope (Roman; R. Akeson et al. 2019).



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

In the standard HOD framework, galaxies are subcategorized into central and satellite populations based on their spatial locations within haloes. A central galaxy is defined as the one residing at the center of mass of the halo, while the satellite galaxies trace the outer gravitational potential. Several parametric HOD models have been proposed in the literature, many of which assume a smoothed step function for the central HOD and a power law for the satellite HOD (e.g., A. A. Berlind & D. H. Weinberg 2002; Z. Zheng et al. 2005; Z. Zheng & D. H. Weinberg 2007). The five-parameter model proposed in Z. Zheng et al. (2005) has demonstrated considerable success in reproducing the occupation statistics of galaxies selected above a minimum stellar mass (M_*) in hydrodynamic and semianalytic simulations. On the contrary, galaxy samples selected on the basis of the minimum star formation rate (SFR) exhibit distinctive features that are not well described by the Z. Zheng et al. (2005) parameterization. The central HODs of SFR-selected galaxy samples are best described by a Gaussian at low masses and a constant (usually) fractional occupation at higher masses, while the satellite HOD follows a smoothed step function at low masses and a power-law scaling at higher masses (J. E. Geach et al. 2012). An in-depth comparison of these two parameterizations is conducted in Appendix A.

The parametric approach to HOD modeling provides a straightforward path to empirically constraining the HOD from measurements of the 2PCF. This involves assuming a functional form for the HOD to construct a forward model of the 2PCF. The best-fit parameters of the HOD are then determined by iteratively exploring the HOD parameter space using a sampling algorithm (e.g., Markov Chain Monte Carlo (MCMC)), forward modeling each set of HOD parameters into a “model” 2PCF, and comparing to the observed 2PCF. This method of HOD estimation has been applied extensively in the literature to derive constraints on the HODs of different galaxy samples over a range of redshifts (e.g., M. White et al. 2007, 2011; Z. Zheng et al. 2007; J. Coupon et al. 2012; J. E. Geach et al. 2012; A. Durkalec et al. 2015; H. Guo et al. 2015; R. K. Cochrane et al. 2017).

While the parametric approach to HOD modeling allows for an empirical determination of the HOD from measurements of the galaxy 2PCF, degeneracies that arise among the HOD model parameters can lead to an underestimation of the HOD uncertainties if the full covariance matrix is not carefully propagated. Further, the choice of parameterization may bias the form of the HOD. This problem hints at a deeper issue of parametric rigidity and implies that existing parameterizations cannot represent the complete space of physically plausible galaxy samples. This arises in part because the shape of the HOD is highly dependent on the physical properties of the selected galaxy sample. This becomes especially relevant when selecting galaxies in a multidimensional physical property space (e.g., a simultaneous cut in M_* and SFR), as can be necessary to study specific types of galaxies. For example, Ly α emitters (M. Ouchi et al. 2020) are selected using Ly α equivalent width and SFR thresholds, and H α emitters are often selected using slitless spectroscopy (D. Sobral et al. 2010), which is indirectly influenced by thresholds in M_* and SFR. Performing a traditional parametric HOD inference on these galaxy samples can bias the form of the inferred HOD.

In this paper, we propose a nonparametric approach to HOD modeling. We do not assume an explicit parametric form for the HOD. Instead, we measure HODs as a function of the physical galaxy property limits that are used to select simulated galaxy samples from cosmological simulations. We then use a Bayesian inference framework to derive constraints on the HOD. In this initial proof-of-principle work, we make the optimistic assumption that the simulated galaxy samples match those that could be observed in the real Universe. Because we are no longer restricted to any particular functional form for the HOD, this approach has the potential to infer unbiased, accurate constraints on the HOD of an arbitrary galaxy sample. However, this nonparametric HOD approach is complicated by the significant computational burden involved in computing the galaxy 2PCF, which has complexity $\propto O(N \log N)$ where N is the number of galaxies in the sample for typical state-of-the-art algorithms, at each point in physical galaxy property space. For this reason, we train a machine learning–based algorithm to predict (i.e., emulate) the 2PCF and HOD as a function of the physical galaxy property limits that are used to select the galaxy sample. Incorporating a trained emulator of the 2PCF and HOD into an MCMC, we can dramatically speed up each likelihood evaluation, and ensure the computational feasibility of the nonparametric HOD approach. Although machine learning–based approaches to HOD modeling have been explored in the literature (e.g., Z. Zhai et al. 2019; D. Wadekar et al. 2020; A. M. Delgado et al. 2022; T. Dumerchat & J. Bautista 2024; A. Jana & L. Samushia 2024; K. Storey-Fisher et al. 2024), all of these analyses adopt an underlying HOD parameterization. Our approach differs from these in that we do not require any assumption on the shape of the HOD a priori.

An alternative empirically motivated approach to studying the galaxy–halo connection is subhalo abundance matching (SHAM; A. V. Kravtsov et al. 2004; A. Vale & J. P. Ostriker 2004). The traditional SHAM model employs a similar wisdom to our proposed nonparametric HOD model in that it does not assume an explicit functional form for the central and satellite HODs. SHAM instead adopts an injective and monotonic relation between galaxy properties (often M_*) and subhalo properties (e.g., the subhalo circular velocity; C. Conroy et al. 2006; A. R. Zentner et al. 2014). Associating host haloes with central galaxies and subhaloes with satellite galaxies naturally determines the halo occupation statistics of a given threshold sample without needing to model baryonic physics (an analogous feature of HOD models). While traditional SHAM modeling can efficiently predict galaxy clustering, it requires careful extension to account for correlations between stellar mass and any other galaxy properties (e.g., A. P. Hearin & D. F. Watson 2013; J. Chaves-Montero et al. 2016).

In what follows, we will demonstrate how the HOD can be inferred from measurements of the galaxy 2PCF without assuming a parametric form for the HOD. We recognize that our method is not a fully nonparametric HOD model in terms of being able to fit absolutely any HOD shape, rather it uses two physical properties as parameters instead of HOD parameters. In Section 2.1 we outline how the 2PCF is estimated from (simulated) galaxy survey data. In Section 2.2 we introduce the standard HOD paradigm, and derive an expression for the galaxy 2PCF in terms of the HOD. In Section 3 we describe the suite of cosmological simulations used to train, validate, and test our emulator, and introduce a

unique approach to efficiently estimating the galaxy 2PCF. In Section 4 we detail the emulation algorithm, training procedure, and introduce the emulator-assisted inference framework used to derive constraints on the HOD from a measurement of the galaxy 2PCF. We validate the HOD recovery performance over a collection of mock observations drawn from the Santa Cruz (SC) semianalytic model (SAM)-based simulation used to train the emulator in Section 5. In Section 6 we test the HOD recovery performance using a set of mock observations drawn from the TNG100-1 simulation. In Section 7 we discuss the methodological limitations and caveats of our nonparametric HOD framework and the direction of future work. In Section 8 we summarize the implications of our results.

2. Background

2.1. Galaxy Two-point Correlation Function

The spatial clustering of galaxies is often characterized using the isotropic real-space galaxy 2PCF $\xi(r)$. $\xi(r)$ quantifies the excess-over-random probability of finding two galaxies within volume elements dV_1 and dV_2 at separation r :

$$dP_{12}(r) = \bar{n}^2 [1 + \xi(r)] dV_1 dV_2, \quad (1)$$

where $\bar{n}$ is the mean number density of the galaxy sample (F. Bernardeau et al. 2002). In principle, $\xi(r)$ is only measurable when the line of sight distance to each galaxy is available. This is true for galaxy redshift surveys, where individual galaxies are labeled by their sky coordinates $(\alpha_i, \delta_i, z_i)$. Here, α_i is the R.A., δ_i is the decl., and z_i is the redshift. In this work, we treat the redshift z_i as purely cosmological in origin (i.e., the apparent Doppler shift arises solely from the expansion of the Universe), and reserve consideration of peculiar-velocity-induced redshift-space distortions for future work. To compute each separation r , the coordinates $(\alpha_i, \delta_i, z_i)$ are transformed into Cartesian form (X_i, Y_i, Z_i) and the Euclidean separation between galaxy pairs is calculated. This involves the intermediate step of measuring the comoving distance $\chi(z)$ to each galaxy:

$$\chi(z) = c \int_0^z \frac{dz}{H(z)}, \quad (2)$$

where c is the speed of light and $H(z)$ is the Hubble parameter (D. W. Hogg 2000). In this work, we adopt a flat Λ CDM Universe with the cosmological parameter values reported by the Planck Collaboration in 2015 (P. A. R. Ade et al. 2016).

In surveys where the line-of-sight distance to each galaxy is not measured (i.e., photometric galaxy surveys), the galaxy catalog consists only of coordinates (α_i, δ_i) . This results in a measurement of the angular 2PCF $w(\theta)$, related to $\xi(r)$ via a line-of-sight projection integral. $w(\theta)$ can be defined in a similar manner to $\xi(r)$, representing the excess-over-random probability of finding two galaxies in the solid angle elements $d\Omega_1$ and $d\Omega_2$ with a mean surface number density $\bar{\nu}$ and at an angular separation θ (P. J. E. Peebles 1980). The angular separation θ (in degrees) is related to the real-space separation r as

$$\theta = \frac{180}{\pi} \left[2 \arcsin \left(\frac{r}{2\chi(z)} \right) \right]. \quad (3)$$

In this work, we use $w(\theta)$ as the basis for both the emulator discussed in Section 4.1 and the observable space to which we fit. This is because $w(\theta)$ is the nominal observable in photometric galaxy clustering studies over which we plan to validate our methodology in future work.

In practice, $w(\theta)$ is estimated from a data catalog (D) by counting pairs of galaxies as a function of separation distance and comparing with the distribution expected from a random catalog (R) that covers the same geometric footprint as the survey. Various estimators of the 2PCF have been presented in the literature (e.g., M. Davis & P. J. E. Peebles 1983; A. J. S. Hamilton 1993; S. D. Landy & A. S. Szalay 1993), the simplest being the naive estimator originally proposed by M. G. Hauser & P. J. E. Peebles (1973):

$$w(\theta) = \frac{DD(\theta)}{RR(\theta)} \frac{N_R(N_R - 1)}{N_D(N_D - 1)} - 1, \quad (4)$$

where $DD(\theta)$ and $RR(\theta)$ are the number of data–data pair counts and random–random pair counts with pair separations in the range $\theta \pm \Delta\theta/2$, respectively, N_D is the total number of galaxies in the data catalog, and N_R is the total number of points in the random catalog (in this work, we specify $N_R = 10N_D$).

Because N_D and N_R are often large (i.e., $>10^5$ for large-scale galaxy surveys) the 2PCF is generally an expensive statistic to compute. As a result, various algorithms have been employed to speed up the computation of the 2PCF by making approximations on large scales. In this work, we use the Python package `treecorr`⁶ (M. Jarvis et al. 2004) to compute the $DD(\theta)$ and $RR(\theta)$ histograms. `treecorr` speeds up the computation of histograms by implementing a ball-tree-based routine that partitions data points into smaller subsets to achieve an overall complexity scaling of $O(N \log N)$, where N is the number of galaxies. While this represents a considerable improvement, estimating 2PCFs for a large number of galaxy samples, as is necessary for the inference we will conduct, remains a computationally demanding task. This motivates the use of the emulator that is introduced in Section 4.

2.2. Halo Occupation Distributions

The standard HOD framework predicts the average number of galaxies (with a specific set of properties) that reside within a dark matter halo using the conditional probability distribution $P(N|M_h)$, that a halo of virial mass M_h contains N galaxies (A. A. Berlind & D. H. Weinberg 2002). The mean number of galaxies occupying a halo as a function of virial mass (i.e., the first moment of $P(N|M_h)$) is then

$$\langle N(M_h) \rangle = \sum_{N=1}^{\infty} NP(N|M_h). \quad (5)$$

A standard assumption of the HOD framework stipulates that all haloes with an occupation number $N > 0$ contain one central galaxy and $N - 1$ satellite galaxies. As a result, the mean occupation number of galaxies can be decomposed into mean central and satellite occupations:

$$\langle N(M_h) \rangle = \langle N_{\text{cen}}(M_h) \rangle + \langle N_{\text{sat}}(M_h) \rangle \quad (6)$$

⁶ <https://github.com/rmjjarvis/TreeCorr>, v5.0

where $\langle N_{\text{cen}}(M_h) \rangle$ and $\langle N_{\text{sat}}(M_h) \rangle$ will herein be referred to as the central and satellite HOD, respectively.

2.3. Galaxy Clustering and Halo Occupation Distributions

The halo model provides a convenient analytic formalism to describe the clustering of galaxies, operating under the principle that all matter is partitioned over haloes that have well-understood properties (A. V. Kravtsov et al. 2004). This enables the 2PCF to be specified entirely by five components:

1. the halo mass function $\frac{dn_h}{dM_h}$ defined such that $\frac{dn_h}{dM_h} dM_h$ is the number of haloes in the mass range $[M_h, M_h + dM_h]$ per unit comoving volume;
2. the linear halo bias function $b_h(M_h)$;
3. the isotropic normalized, radial number density profile of satellite galaxies in haloes $u_s(r|M_h)$;
4. the linear 2PCF of dark matter $\xi_{\text{DM}}^{\text{lin}}(r)$; and
5. the conditional probability distribution $P(N|M_h)$ provided by the HOD formalism.

Following a similar derivation to A. Cooray & R. Sheth (2002)—and operating in the $\xi(r)$ basis for consistency with the literature—the dependence of the 2PCF on each of the five components listed above can be derived analytically by considering the galaxy overdensity field

$$\delta_g(\mathbf{x}) = \frac{n_g(\mathbf{x})}{\bar{n}_g} - 1, \quad (7)$$

where $n_g(\mathbf{x})$ is the number density of galaxies in real space and $\bar{n}_g$ is its mean (i.e., $\langle n_g(\mathbf{x}) \rangle = \bar{n}_g$). The autocorrelation of this field is the galaxy 2PCF

$$\xi(r) = \frac{1}{\bar{n}_g^2} \langle n_g(\mathbf{x}) n_g(\mathbf{x} + \mathbf{r}) \rangle - 1. \quad (8)$$

Expressing $n_g(\mathbf{x})$ as a summation over the central and satellite galaxy contributions of the i th halo with mass $M_{h,i}$ and a center of mass located at $\mathbf{x}_i$:

$$n_g(\mathbf{x}) = \sum_i N_{\text{cen},i} \delta^3(\mathbf{x} - \mathbf{x}_i) + N_{\text{sat},i} u_s(|\mathbf{x} - \mathbf{x}_i||M_{h,i}), \quad (9)$$

Equation (8) becomes

$$\begin{aligned} \xi(r) = & \frac{1}{\bar{n}_g^2} \sum_{i,j} [\langle N_{\text{cen},i} N_{\text{cen},j} \delta^3(\mathbf{x} - \mathbf{x}_i) \delta^3(\mathbf{x} + \mathbf{r} - \mathbf{x}_j) \rangle + \\ & \langle N_{\text{cen},i} N_{\text{sat},j} \delta^3(\mathbf{x} - \mathbf{x}_i) u_s(|\mathbf{x} + \mathbf{r} - \mathbf{x}_j||M_{h,j}) \rangle + \\ & \langle N_{\text{sat},i} N_{\text{cen},j} u_s(|\mathbf{x} - \mathbf{x}_i||M_{h,i}) \delta^3(\mathbf{x} + \mathbf{r} - \mathbf{x}_j) \rangle + \\ & \langle N_{\text{sat},i} N_{\text{sat},j} u_s(|\mathbf{x} - \mathbf{x}_i||M_{h,i}) u_s(|\mathbf{x} + \mathbf{r} - \mathbf{x}_j||M_{h,j}) \rangle] - 1. \end{aligned} \quad (10)$$

Following convention, we split the summation into $i = j$ (one-halo) and $i \neq j$ (two-halo) terms to separate the contribution from pairs of galaxies located in the same halo from those located in separate haloes, such that

$$\xi(r) = \xi^{\text{1h}}(r) + \xi^{\text{2h}}(r). \quad (11)$$

The one-halo term of Equation (10) then becomes

$$\begin{aligned} \xi^{\text{1h}}(r) = & \frac{1}{\bar{n}_g^2} \sum_i [\langle N_{\text{cen},i}^2 \delta^3(\mathbf{x} - \mathbf{x}_i) \delta^3(\mathbf{x} + \mathbf{r} - \mathbf{x}_i) \rangle + \\ & \langle N_{\text{cen},i} N_{\text{sat},i} \delta^3(\mathbf{x} - \mathbf{x}_i) u_s(|\mathbf{x} + \mathbf{r} - \mathbf{x}_i||M_{h,i}) \rangle + \\ & \langle N_{\text{sat},i} N_{\text{cen},i} u_s(|\mathbf{x} - \mathbf{x}_i||M_{h,i}) \delta^3(\mathbf{x} + \mathbf{r} - \mathbf{x}_i) \rangle + \\ & \langle N_{\text{sat},i}^2 u_s(|\mathbf{x} - \mathbf{x}_i||M_{h,i}) u_s(|\mathbf{x} + \mathbf{r} - \mathbf{x}_i||M_{h,i}) \rangle] \end{aligned} \quad (12)$$

where we have implicitly shifted the -1 in Equation (10) to the two-halo term. Expressing the ensemble average and summation as a double integral over mass and space, the one-halo term then becomes

$$\begin{aligned} \xi^{\text{1h}}(r) = & \frac{1}{\bar{n}_g^2} \left[2 \int dM_h \frac{dn_h}{dM_h} \langle N_{\text{cen}}(M_h) \rangle \langle N_{\text{sat}}(M_h) \rangle u_s(r|M_h) \right. \\ & + \int dM_h \frac{dn_h}{dM_h} \langle N_{\text{sat}}(M_h) (N_{\text{sat}}(M_h) - 1) \rangle \\ & \left. \times \int d\mathbf{y} u_s(|\mathbf{x} - \mathbf{y}||M_h) u_s(|\mathbf{x} + \mathbf{r} - \mathbf{y}||M_h) \right], \end{aligned} \quad (13)$$

where we have assumed that the occupation statistics of central and satellite galaxies are independent to separate $\langle N_{\text{cen}} N_{\text{sat}} \rangle = \langle N_{\text{cen}} \rangle \langle N_{\text{sat}} \rangle$ and ignored the contribution from central–central self-pairs, which introduce artificial power at $r = 0$ via $\delta^3(r)$ (M. Asgari et al. 2023). We can further simplify the second term of Equation (13) if we assume the satellite occupation to follow Poisson statistics:

$$\begin{aligned} \xi^{\text{1h}}(r) = & \frac{1}{\bar{n}_g^2} \left[2 \int dM_h \frac{dn_h}{dM_h} \langle N_{\text{cen}}(M_h) \rangle \langle N_{\text{sat}}(M_h) \rangle u_s(r|M_h) \right. \\ & \left. + \int dM_h \frac{dn_h}{dM_h} \langle N_{\text{sat}}(M_h) \rangle^2 (u_s * u_s)(r|M_h) \right], \end{aligned} \quad (14)$$

where $(u_s * u_s)(r|M_h)$ is the auto-convolution of the radial number density profile of satellites. The two-halo term can be obtained similarly by considering the $i \neq j$ case of Equation (10), although with two additional considerations. Because dark matter haloes are themselves clustered, the halo–halo correlation function $\xi_{\text{hh}}(r|M_{h,1}, M_{h,2})$ needs to be accounted for. Typically, haloes are expected to follow the clustering of the dark matter density field, albeit with a mass-dependent enhancement called the linear halo bias function $b_h(M_h)$ (H. J. Mo et al. 1996; R. K. Sheth & G. Tormen 1999). Under this approximation $\xi_{\text{hh}}(r|M_{h,1}, M_{h,2}) \approx b_h(M_{h,1}) b_h(M_{h,2}) \xi_{\text{DM}}^{\text{lin}}(r)$. A further simplification can be made by assuming $u_s(r|M)$ to be sharply peaked relative to the separations considered in the two-halo regime such that $u_s(\mathbf{x} - \mathbf{y}|M_h) \approx \delta^3(\mathbf{x} - \mathbf{y})$, yielding

$$\xi^{\text{2h}}(r) \approx \xi_{\text{DM}}^{\text{lin}}(r) \left[\frac{1}{\bar{n}_g} \int dM_h \frac{dn_h}{dM_h} \langle N(M_h) \rangle b_h(M_h) \right]^2. \quad (15)$$

The term in brackets is often referred to as the large-scale galaxy bias b_g , such that $\xi^{\text{2h}}(r) = b_g^2 \xi_{\text{DM}}^{\text{lin}}(r)$ (N. Kaiser 1984). The two-halo term in angular space can be obtained by directly substituting the angular linear 2PCF of dark matter $w_{\text{DM}}^{\text{lin}}(\theta)$ for $\xi_{\text{DM}}^{\text{lin}}(r)$, yielding $w^{\text{2h}}(\theta) = b_g^2 w_{\text{DM}}^{\text{lin}}(\theta)$.

The halo model derivation of the one-halo term (which dominates on small scales) and two-halo term (which dominates on large scales) emphasizes the dependence of clustering on the underlying halo occupation statistics. More specifically, the one-halo term depends on the second moment of $P(N|M_h)$, $\langle N(N-1) \rangle = \sum_{N=1}^{\infty} N(N-1)P(N|M_h)$ and the two-halo term depends on the first moment of $P(N|M_h)$, $\langle N(M_h) \rangle$.

3. Simulations

3.1. SC Semianalytic Model

To train and validate our emulator we use the suite of simulated light cones presented in L. Y. A. Yung et al. (2023). These simulations include predicted physical galaxy properties and photometry for the deep-field extragalactic surveys that will be carried out by a variety of space- and ground-based observatories, including the James Webb Space Telescope (JWST; J. P. Gardner et al. 2023), Roman, Euclid, and Rubin in the near future. Each of the five $2 \deg^2$ light cones has a redshift range of $z \sim 0$ to 10. In the initial investigation presented in this paper, we utilize slices of the light cones spanning $0.7407 < z < 0.7793$, serving the dual purpose of enabling eventual validation against the SDSS Baryon Oscillation Spectroscopic Survey (D. J. Eisenstein et al. 2011) that covers a similar redshift range and ensuring the line-of-sight width aligns with the other simulation we consider in Section 3.2. We additionally validate our methodology over alternative redshift ranges ($0.1 < z < 0.2$ and $1.3 < z < 1.4$), and find that the results of our analysis do not vary significantly. The underlying dark matter haloes in the light cones are extracted from the dark-matter-only N -body SMDPL simulation in the MultiDark suite (A. Klypin et al. 2016), with dark matter halo merger histories generated on the fly with the extended Press–Schechter (ePS) formalism (e.g., R. S. Somerville & T. S. Kolatt 1999). The galaxy populations within these dark matter haloes are simulated with the versatile SC SAM for galaxy formation (see R. S. Somerville & J. R. Primack 1999; R. S. Somerville et al. 2008, 2012, 2015, for an in-depth description of the various components of this model; and see R. S. Somerville et al. 2021; L. Y. A. Yung et al. 2022, for details regarding simulated light cones). To model galaxy formation, the SC SAM uses a combination of physical processes that are either described analytically from theory or calibrated from observations and hydrodynamic simulations (L. Y. A. Yung et al. 2019, 2021). The SC SAM provides a wide range of predictions for the physical properties of galaxies. This results in an extensive list of galaxy properties available as criteria to select simulated galaxy samples. The SC SAM has been shown to well reproduce a wide range of observational constraints, including one-point distribution functions of M_{UV} , M_* , and SFR and two-point autocorrelation functions from $0 \lesssim z \lesssim 7.5$. In this preliminary analysis, we select galaxy samples using lower limits on stellar mass (M_*) and on the instantaneous SFR. Future work can expand the space of physical galaxy properties considered.

For a given simulated galaxy sample, we compute three histograms: (1) the normalized galaxy pair counts within haloes (i.e., the normalized one-halo pair count distribution) as a function of angular separation— $dd^{1h}(\theta)$, (2) the number density of central galaxies as a function of halo mass— $n_{\text{cen}}(M_h)$, and (3) the number density of satellite galaxies as a function of

halo mass— $n_{\text{sat}}(M_h)$. These histograms were chosen because they can be combined, along with a few additional quantities, to efficiently estimate both the 2PCF and HOD of a galaxy sample. The HODs are computed as

$$\langle N_{\text{cen}}(M_h) \rangle = \frac{n_{\text{cen}}(M_h) V_{\text{sim}}}{N_h(M_h)}, \quad (16)$$

$$\langle N_{\text{sat}}(M_h) \rangle = \frac{n_{\text{sat}}(M_h) V_{\text{sim}}}{N_h(M_h)}, \quad (17)$$

where $N_h(M_h)$ is the halo mass distribution of the simulation and V_{sim} is the simulation volume in comoving units. To compute the 2PCF, we first consider decomposing $DD(\theta)$ from Equation (4) into one- and two-halo contributions as $DD(\theta) = DD^{1h}(\theta) + DD^{2h}(\theta)$, where $DD^{1h}(\theta)$ is the unnormalized version of $dd^{1h}(\theta)$ defined above, and $DD^{2h}(\theta)$ is the pair count distribution of galaxies located in separate haloes. Therefore Equation (4) can be rewritten as

$$w(\theta) = \frac{dd^{1h}(\theta) N_R(N_R - 1)}{RR(\theta) 2} + \left[\frac{DD^{2h}(\theta) N_R(N_R - 1)}{RR(\theta) N_D(N_D - 1)} - 1 \right], \quad (18)$$

where we identify the term in brackets as the two-halo term $w^{2h}(\theta)$. Given we have already derived an analytical approximation for $w^{2h}(\theta)$, we can conveniently avoid having to numerically estimate $DD^{2h}(\theta)$ through pair counting. Substituting the angular analog of Equation (15) into Equation (18) produces a computationally efficient equation for estimating $w(\theta)$ that can be obtained from the three aforementioned histograms:

$$w(\theta) \approx \frac{dd^{1h}(\theta) N_R(N_R - 1)}{RR(\theta) 2} + w_{\text{DM}}^{\text{lin}}(\theta) \left[\frac{1}{n_D} \int dM_h \frac{dn_h}{dM_h} \langle N(M_h) \rangle b_h(M_h) \right]^2. \quad (19)$$

In the above equation, the number density of galaxies in the sample is

$$\bar{n}_D = \sum_i n_{\text{cen}}(M_{h,i}) + n_{\text{sat}}(M_{h,i}). \quad (20)$$

We use the Core Cosmology Library Python package `pyccl`⁷ (N. E. Chisari et al. 2019) to compute $b_h(M_h)$ and adopt the J. L. Tinker et al. (2010) halo bias model. To determine $w_{\text{DM}}^{\text{lin}}(r)$ we use the heuristic detailed in Section 7.

Equation (19) is particularly useful because it naturally isolates the one-halo term from the two-halo term, allowing one to study the behavior of each component separately. In Figure 1, we show the forms of $\frac{dn_h}{dM_h}$, $b_h(M_h)$, and $w_{\text{DM}}^{\text{lin}}(\theta)$ used in this work. We also show the HOD and 2PCF (with its one- and two-halo decomposition) for a galaxy sample with $(\log_{10}(M_*/M_\odot), \log_{10}(\text{SFR}/M_\odot \text{ yr}^{-1})) \geq (8.45, -2.45)$. The halo mass function, HOD, and 2PCF curves shown in Figure 1 are the means across the five light cone realizations considered.

⁷ <https://github.com/LSSTDESC/CCL>, v3.2

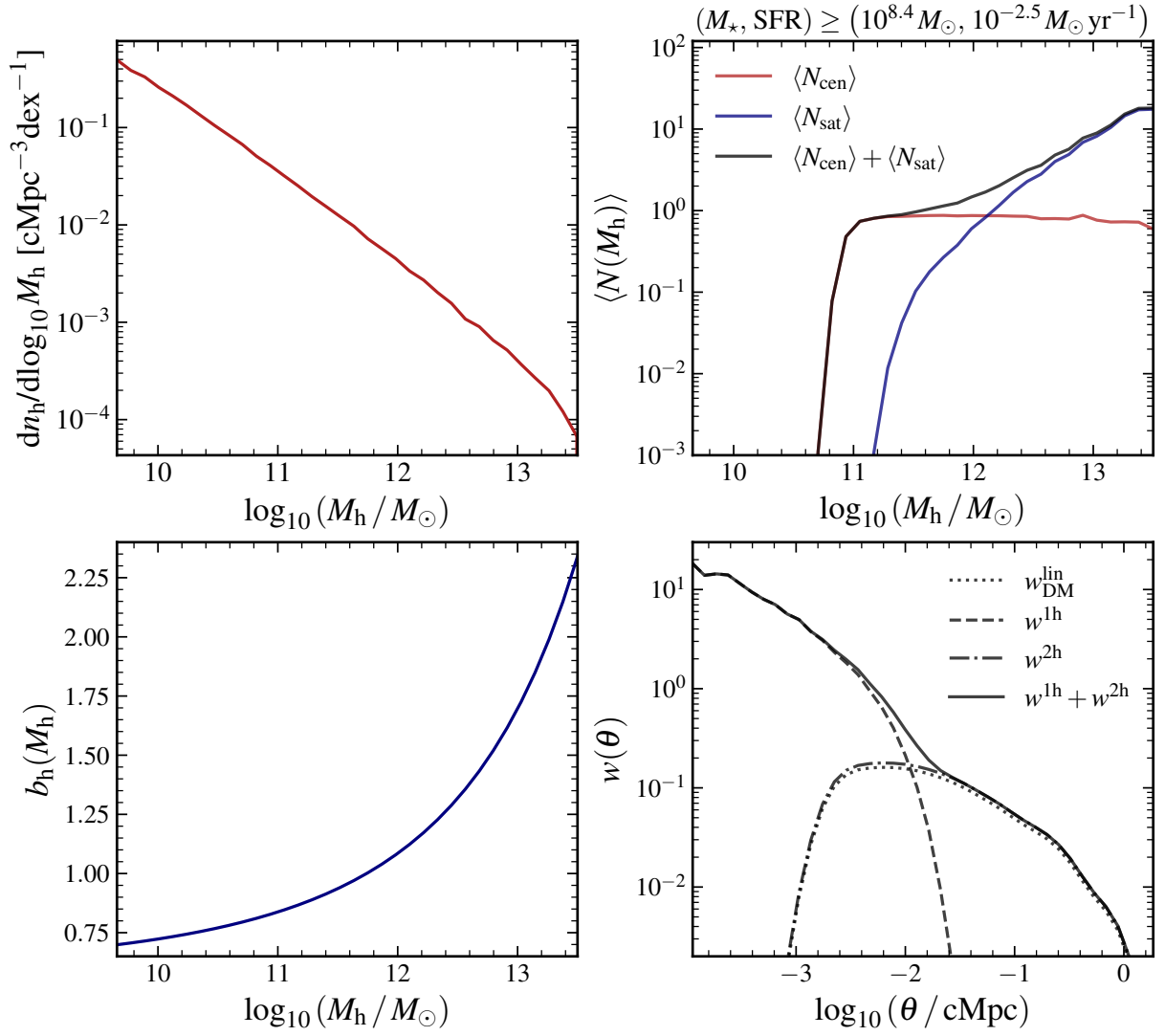


Figure 1. Upper left panel: halo mass function for haloes found in the light cones at $0.7407 < z < 0.7793$. Lower left panel: the J. L. Tinker et al. (2010) linear halo bias at $z \sim 0.76$. Upper right panel: the central, satellite, and total HODs for a galaxy sample selected with $(M_*, \text{SFR}) \geq (10^{8.45} M_\odot, 10^{-2.45} M_\odot \text{yr}^{-1})$ from the five 2 deg^2 light cones. Lower right panel: the 2PCF for the same galaxy sample. The one-halo term (dashed line) is computed numerically using the first term in Equation (19) while the two-halo term (dashed-dotted line) is estimated using the approximation in the second term of Equation (19). The two-halo term is a rescaling of the linear dark matter 2PCF (dotted). The linear dark matter correlation function is truncated on angular scales $\lesssim 0.001$ after imposing the scale-dependent additive correction detailed in Section 7. We emphasize that at small scales the dark matter correlation function represents a highly subdominant contribution (i.e., $\lesssim$ a few percent) to the total galaxy 2PCF.

3.2. IllustrisTNG

To test our methodology, we use the IllustrisTNG project’s TNG100-1 simulation (D. Nelson et al. 2019). TNG100-1 is the main high-resolution IllustrisTNG100 run that utilizes the full TNG physics model. For consistency with the SC SAM light cones, we select the nearest redshift snapshot from the TNG100-1 simulation at $z = 0.76$, which has a cubic volume of 110.7^3 cMpc^3 . In TNG100-1, group (i.e., dark matter halo) fields are derived using the friends-of-friends algorithm (M. Davis et al. 1985) with a linking length of $b = 0.2$ (D. Nelson et al. 2019) and subhalo (i.e., galaxy) fields are derived using the SUBFIND algorithm (V. Springel et al. 2001). To ensure a fair comparison between the SC SAM and TNG100-1, we implement a completeness cut in halo mass, given TNG100-1 has a lower mass resolution and a larger volume than the SC SAM. This restricts the range of halo masses considered in the two simulations to $\log_{10}(M_h/M_\odot) \in [9.55, 13.5]$. As alluded to in Section 3.1, we

select an $\approx 110.7 \text{ cMpc}$ slice of the SC SAM along the line-of-sight axis that is centered on the TNG100-1 snapshot redshift of 0.76. This ensures that comparisons between measurements of $w(\theta)$ in the SC SAM and TNG100-1 employ the same amount of projection along the line of sight. As detailed in Section 3.1 for the SC SAM, we similarly consider TNG100-1 galaxy samples selected above lower bound thresholds in M_* and SFR. A comparison of the two-dimensional M_* and SFR number density distributions of the two simulations is shown in Figure 2. Notably, the SC SAM employs a strict threshold at $\log_{10}(M_*/M_\odot) = 7$, below which galaxies are not resolved (L. Y. A. Yung et al. 2022). Further, TNG100-1 has an effective minimum SFR of $\approx 7 \times 10^{-4} M_\odot \text{yr}^{-1}$ imposed by the baryonic mass resolution ($1.4 \times 10^6 M_\odot$) and the star formation model. Therefore, we impose minimum lower bound thresholds of $\log_{10}(M_*/M_\odot) = 7$ and $\log_{10}(\text{SFR}/M_\odot \text{yr}^{-1}) = -3$ on the galaxy populations considered in this work. Notably, the SC SAM satellite distribution

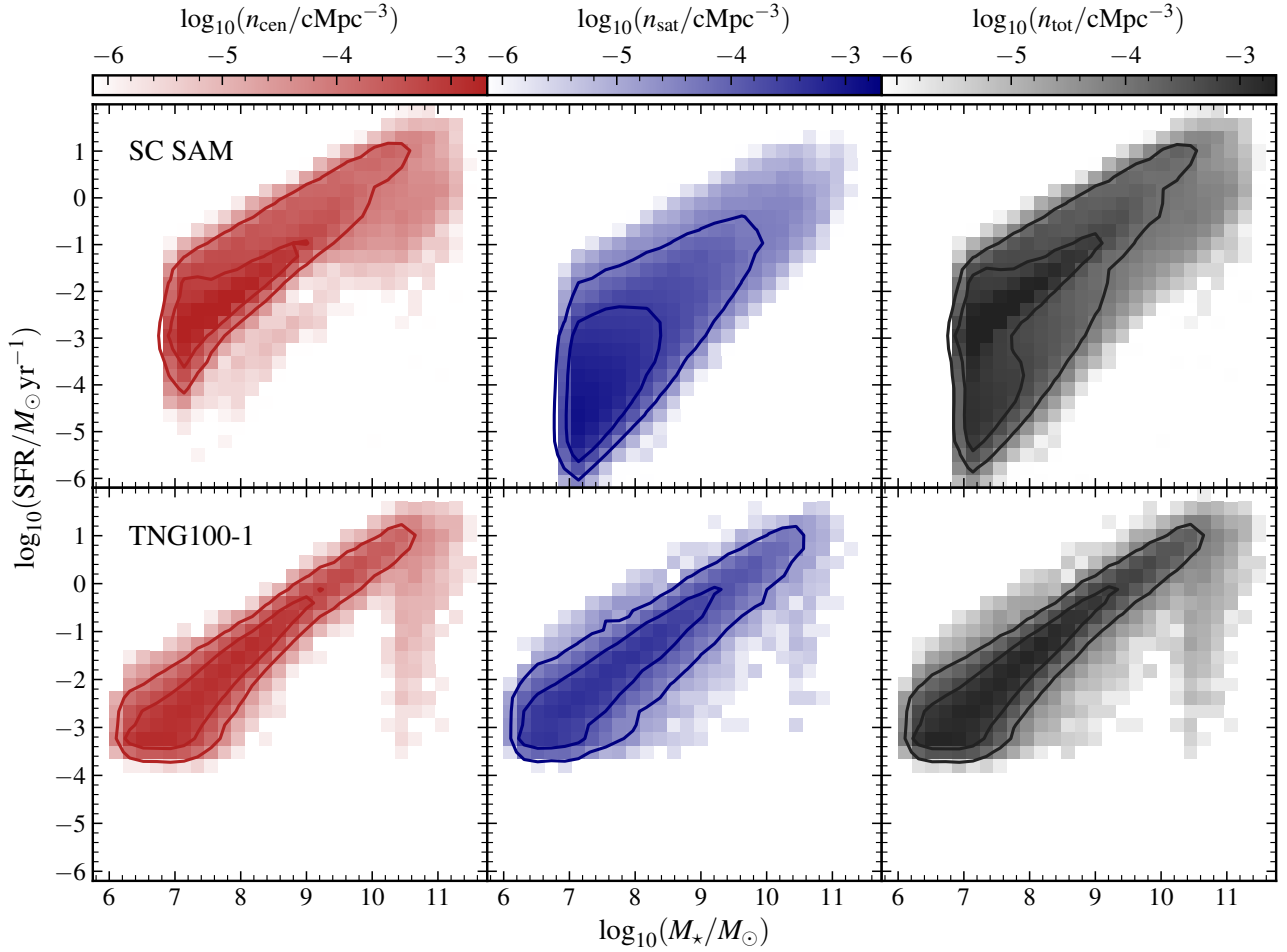


Figure 2. SFR– M_* number density distributions of central galaxies (red), satellite galaxies (blue), and their sum (black) in the SC SAM (lower row) and TNG100-1 snapshot (bottom row). Contours containing 68% (inner) and 95% (outer) of the respective distributions are plotted overtop in each panel. The SC SAM does not resolve galaxies below $\log_{10}(M_*/M_\odot) = 7$, while TNG100-1 has an effective minimum resolved SFR of $\log_{10}(\text{SFR}/M_\odot \text{ yr}^{-1}) \approx -3.5$.

falls off steeply below $\log_{10}(\text{SFR}/M_\odot \text{ yr}^{-1}) \approx -1.5$ due to quenching (see Section 7 for further discussion). We also note that while the definitions of M_* , SFR, and halo mass differ slightly in the two simulations, an effort has been made to compare the most analogous quantities across the SC SAM and TNG100-1, following the discussion in Appendix A of A. Gabrielpillai et al. (2022).

4. Emulation and Inference Framework

4.1. Gaussian-process-based Emulation

To learn the mapping between the 2PCF space, the galaxy physical property space, and the HOD space, we use a Gaussian process (GP)-based emulator. A GP is a collection of random variables, any finite number of which have a joint Gaussian distribution (see C. E. Rasmussen & C. K. I. Williams 2006, for an in-depth discussion of GPs). GPs are well suited to our task given their demonstrated success at efficient, nonlinear interpolation across high-dimensional parameter spaces, while reporting uncertainties.

We construct our emulator using the Python package `GPY`⁸ (GPY 2012). We initialize an independent GP for each spatial separation bin and halo mass bin of the three histograms introduced above, which are shown in Figure 3. We construct a

training set design matrix X for each GP with a shape $N_{\text{input}} \times N_{\text{train}}$, where N_{input} is the number of physical galaxy properties considered (i.e., two) and N_{train} is the size of the training set. We additionally construct a training set target vector $\mathbf{y}$ with length N_{train} , given that our task is designed to predict a scalar output. Assuming that the targets have additive Gaussian distributed noise, each target $y \in \mathbf{y}$ has the form $y = f(\mathbf{x}) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ and f is the function we aim to learn. We assume f to have zero mean and a covariance described by the kernel function k . The prior over $\mathbf{y}$ is specified by the covariance matrix $\text{cov}(\mathbf{y}) = K(X, X) + \sigma_n^2 I$. Here, $K(X, X)$ is a matrix with shape $N_{\text{train}} \times N_{\text{train}}$, with entries $k(\mathbf{x}_i, \mathbf{x}_j)$ describing how strongly correlated pairs of points $\mathbf{x}_i, \mathbf{x}_j \in X$ are based on the kernel k . In this work, we specify k to be the squared exponential kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_k^2 \exp \left[-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{2l^2} \right], \quad (21)$$

where σ_k controls the amplitude of the kernel function, $|\dots|$ is the Euclidean norm, and l is the length scale parameter. σ_k and l will serve as the hyperparameters that we aim to optimize when training the GP. We similarly construct a validation set design matrix X_* with shape $N_{\text{input}} \times N_{\text{val}}$ and a target vector $\mathbf{f}_*$ of length N_{val} , where N_{val} is the size of the validation set. The joint distribution of the training and validation set targets can

⁸ <https://github.com/SheffieldML/GPY>, v1.13

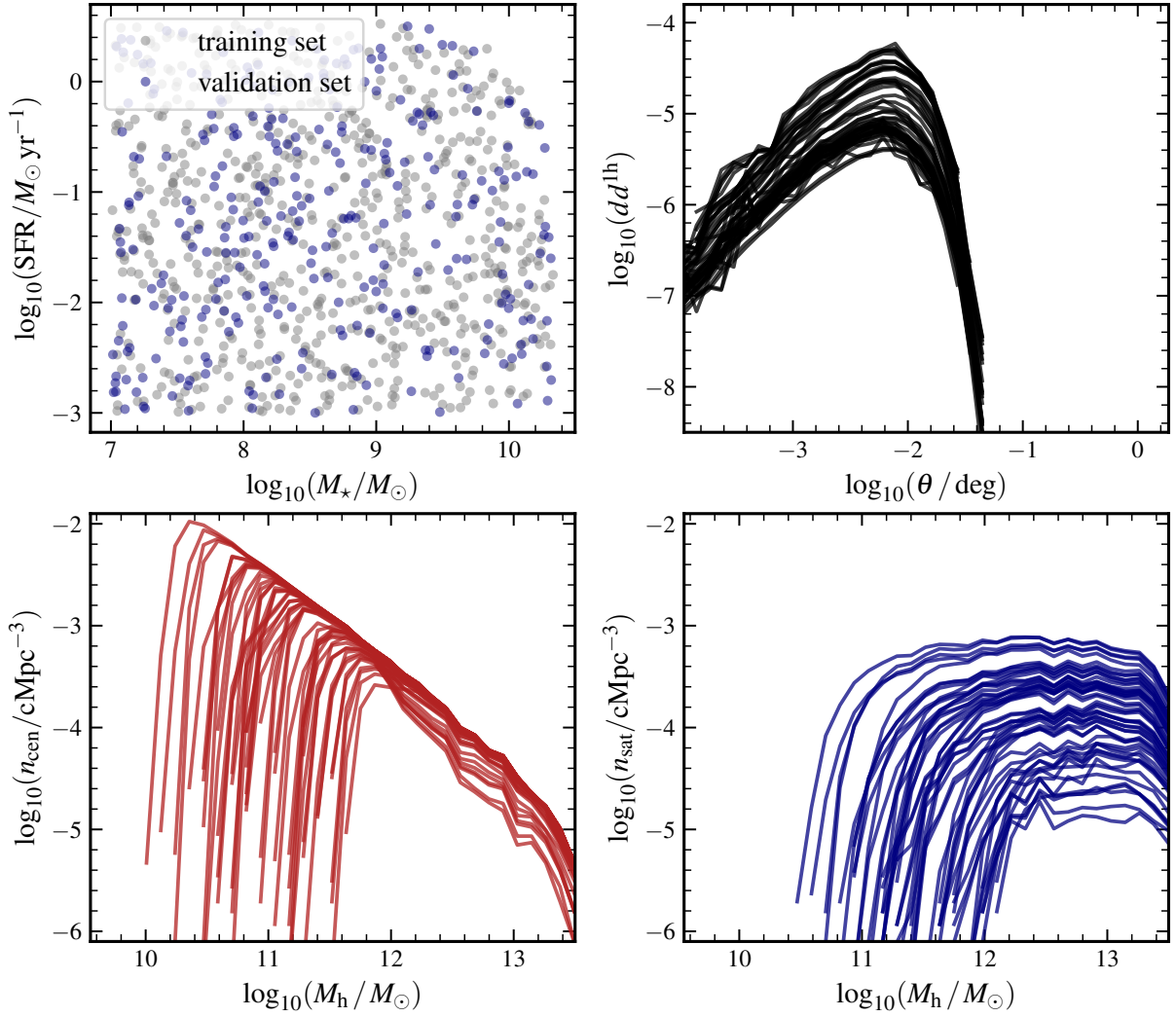


Figure 3. Upper left panel: training set (gray, 700 samples) and validation set (blue, 300 samples) tuples of physical galaxy property lower bound thresholds. The imposed limit on the minimum galaxy sample size manifests as a restriction on the joint maximum lower bound ($\log_{10}(M_*)$, $\log_{10}(\text{SFR})$) thresholds seen in the upper right hand corner of this panel. Upper right panel: the normalized galaxy pair counts within haloes as a function of separation distance for 40 randomly selected physical galaxy property threshold tuples in the training set. Lower left panel: the number density of central galaxies as a function of halo mass for the same 40 randomly selected physical galaxy property threshold tuples. Lower right panel: the number density of satellite galaxies as a function of halo mass for the same 40 randomly selected physical galaxy property threshold tuples.

then be written as

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{\theta}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I} & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right), \quad (22)$$

where $K(\mathbf{X}_*, \mathbf{X})$, $K(\mathbf{X}, \mathbf{X}_*)$, and $K(\mathbf{X}_*, \mathbf{X}_*)$ are matrices defined similarly to $K(\mathbf{X}, \mathbf{X})$ and with shapes dependent on the dimensionality of their respective input design matrices.

The conditional distribution for $\mathbf{f}_*$ is then

$$\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \mathcal{N}(K(\mathbf{X}_*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}]^{-1} \mathbf{y}, \text{cov}(\mathbf{f}_*)), \quad (23)$$

where

$$\begin{aligned} \text{cov}(\mathbf{f}_*) &= K(\mathbf{X}_*, \mathbf{X}_*) \\ &\quad - K(\mathbf{X}_*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}]^{-1} K(\mathbf{X}, \mathbf{X}_*). \end{aligned} \quad (24)$$

Training GPs for each bin of $\text{dd}^{1h}(\theta)$, $n_{\text{cen}}(M)$, and $n_{\text{sat}}(M)$ amounts to finding the set of kernel hyperparameters (σ_k , l)

that maximize the log-marginal likelihood

$$\begin{aligned} \ln(\mathbf{y} | \mathbf{X}, (\sigma_k, l)) &= -\frac{1}{2} \mathbf{y}^T (K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} \\ &\quad - \frac{1}{2} \ln(\det[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}]) - \frac{N_{\text{train}}}{2} \ln(2\pi). \end{aligned} \quad (25)$$

This optimization is performed using the L-BFGS-B algorithm implementation in `SciPy`⁹ (P. Virtanen et al. 2020).

To train and validate the GPs, we generate a near-random sample of 1000 ($\log_{10}(M_*)$, $\log_{10}(\text{SFR})$) tuples using the Latin hypercube sampling (LHS) algorithm (M. D. McKay et al. 1979). LHS enables homogeneous and efficient sampling in high-dimensional parameter spaces, whereby each sample lies on its own unique axis-aligned hyperplane. While a simple random sampling algorithm can easily cover the two-dimensional parameter space we consider here, LHS will become especially useful when scaling to higher-dimensional spaces of galaxy properties. In this work, we consider lower bound thresholds

⁹ <https://docs.scipy.org/doc/scipy/>, v1.12

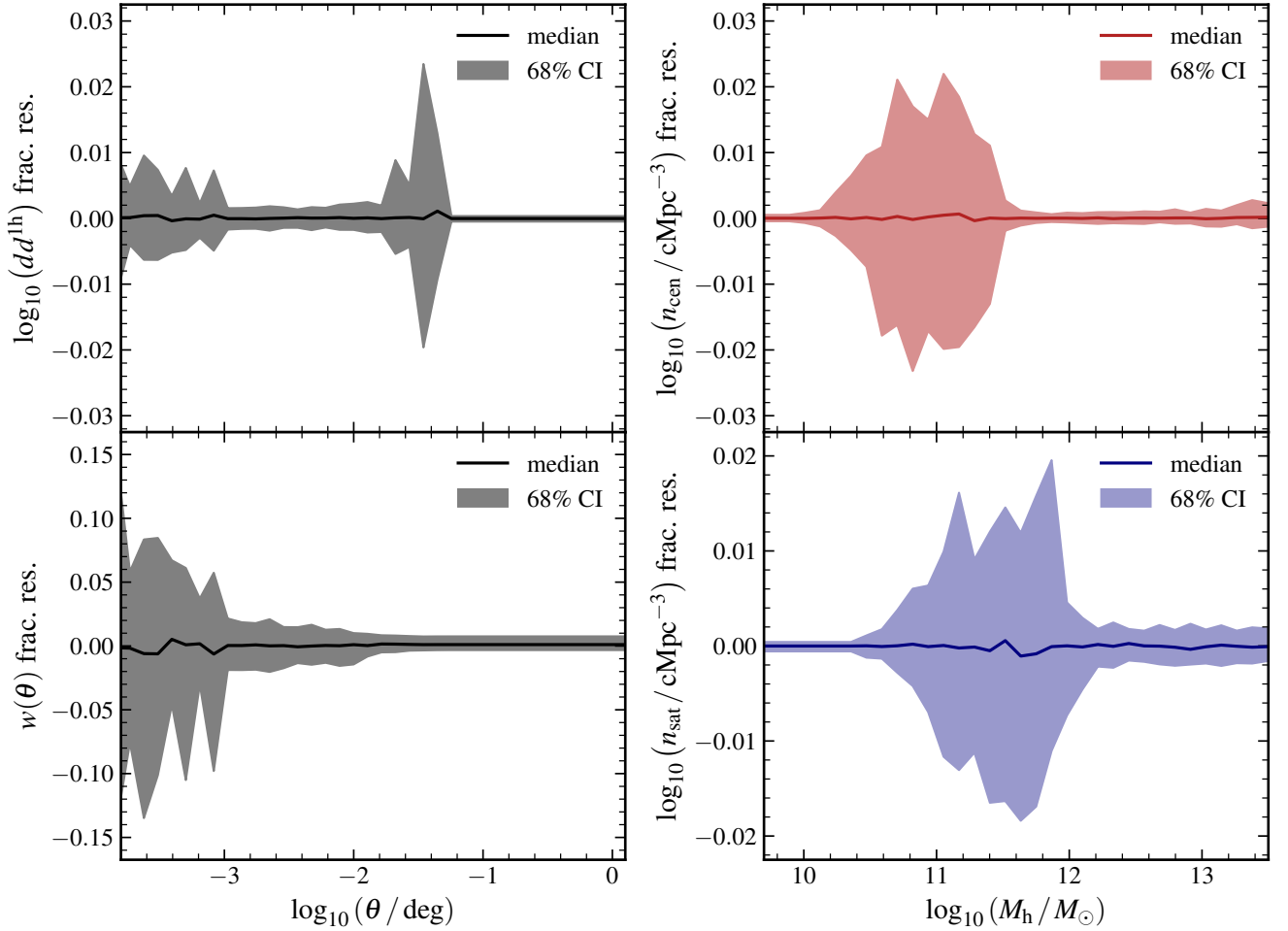


Figure 4. Median fractional residuals (solid curves) and 68% CIs (shaded regions) for $\log_{10}(\text{dd}^{\text{lh}})$ (upper left), $w(\theta)$ (bottom left), $\log_{10}(n_{\text{cen}})$ (upper right), and $\log_{10}(n_{\text{sat}})$ (bottom right) computed using the GP-based emulator over the 300 validation set samples. The fractional error reaches a maximum of $\sim 5\%$ across all angular separation and mass bins for the three one-dimensional histograms, indicating the GP has reliably learned the mapping, even with the limited size of the training set. The fractional error in $w(\theta)$ increases from $<3\%$ on intermediate to large angular scales to $\approx 5\%$ – 10% on angular scales $\lesssim 0.001^\circ$ deg. Note that we have plotted the fractional residuals in $w(\theta)$ and not $\log_{10}(w(\theta))$ as they are more interpretable.

spanning the ranges of $\log_{10}(M_*/M_\odot) \in [7.00, 10.33]$ and $\log_{10}(\text{SFR}/M_\odot \text{ yr}^{-1}) \in [-3.00, 0.52]$. Tuples that would generate samples of fewer than 1000 galaxies are excluded, as can be seen in the top right corner of the upper left panel of Figure 3. This ensures that the galaxy samples we analyze are representative of those typically encountered in real galaxy survey data. We randomly select 700 tuples for the training set and 300 tuples for the validation set (i.e., $N_{\text{train}} = 700$, $N_{\text{val}} = 300$). We emphasize that the training and validation sets are generated using samples drawn from the SC SAM (and not TNG100-1). The training set is used to obtain the optimal set of model weights for the emulator.

The physical galaxy property tuples corresponding to the training and validation sets are shown in the top left panel of Figure 3, while example dd^{lh} , n_{cen} , and n_{sat} distributions from the training set are shown in the remaining three panels. The emulator’s performance evaluated over the validation set is discussed in Section 4.2. We specify dd^{lh} to span an angular range of $\log_{10}(\theta/\text{deg}) \in [-3.95, 0.27]$ and n_{cen} and n_{sat} to span a mass range of $\log_{10}(M_{\text{h}}/M_\odot) \in [9.66, 13.5]$. This minimum angular scale ($\theta \approx 0.040''$) lies above the minimum measurable separation for current-generation (e.g., JWST) space-based imaging, and just below the minimum measurable separation of upcoming ground-based surveys (e.g., Rubin).

We operate in the logarithmic space of both the inputs (physical galaxy property tuples) and the outputs (dd^{lh} , n_{cen} , and n_{sat} histograms) to ensure numerical stability when training and validating the emulator. For this reason, we quantify the emulator performance in logarithmic space for the three histograms in Figure 4.

4.2. Emulator Performance

To characterize the predictive performance of the emulator we compute the fractional residual in the i th bin of each of the emulated quantities (denoted here as x) as

$$(\text{fractional residual})_{x,i} = \frac{x_{\text{truth},i} - x_{\text{emu},i}}{x_{\text{truth},i}}, \quad (26)$$

where $x_{\text{emu},i}$ is the emulator’s prediction and $x_{\text{truth},i}$ is the value in the validation set. The median fractional residuals in each of the emulated quantities [$\log_{10}(\text{dd}^{\text{lh}})$, $\log_{10}(n_{\text{cen}})$, $\log_{10}(n_{\text{sat}})$] and their respective 68% CIs computed over the 300 validation set samples are shown in Figure 4. We also quantify the emulator performance with respect to $w(\theta)$ in Figure 4, given it is the key observable. This is done by forward modeling the three histograms into $w(\theta)$ using Equation (19), and computing the fractional residuals over the same 300 validation set samples.

The median fractional residuals across all angular separation and halo mass bins do not exceed 5% and are typically $<1\%$ for the three one-dimensional histograms in Figure 4. Further, the residuals appear to be symmetrically distributed about zero. Note that the regions where the fractional residuals are ≈ 0 most often correspond to bins where the distributions are either all zero or approach an approximately constant value that is easily learned by the GP (e.g., large real-space separation bins have predictable, slowly varying behavior in the two-halo regime, and the occupation statistics of low-mass haloes decay to zero). The fractional error in $w(\theta)$ increases from $<3\%$ on intermediate to large angular scales to 5%–10% on smaller angular scales. Satisfied with this level of performance, the emulator can be reliably used as a surrogate model for the 2PCF and HOD within an MCMC framework.

4.3. Emulator-accelerated MCMC

In order to constrain the HOD for a given 2PCF, we use an MCMC. Equipped with a set of sufficiently trained GPs, we have an efficient and reliable emulator of the 2PCF and HOD over the space of (M_*, SFR) thresholds. Incorporating this emulator into the MCMC framework, we can dramatically speed up each likelihood evaluation and thus reduce the overall time necessary for convergence. We find that the emulator reduces the time for a single likelihood evaluation by a factor of $\gtrsim 500$ compared to brute-force computing the 2PCF at each evaluation.

To implement the MCMC, we use the Python package `emcee`¹⁰ (D. Foreman-Mackey et al. 2013) and adopt the default affine-invariant ensemble sampling algorithm originally introduced in J. Goodman & J. Weare (2010). We configure the MCMC with a proposal distribution that includes a combination of the “stretch move” and “differential evolution” proposals (see `emcee` documentation for details). We tune the step sizes for each proposal distribution along with their respective frequencies to obtain a mean acceptance fraction in the range of 0.2–0.5.

As discussed in Section 2.1, the observable space in this work is $w(\theta)$. From a Bayesian perspective, the posterior probability distribution over the HOD and the lower bound thresholds $\mathbf{G} = (\log_{10}(M_*), \log_{10}(\text{SFR}))$ given a mock observation of the 2PCF $w^{\text{obs}}(\theta)$ is

$$p(\langle N(M_h) \rangle, \mathbf{G} | w^{\text{obs}}(\theta)) \propto p(w^{\text{obs}}(\theta) | \langle N(M_h) \rangle, \mathbf{G}) p(\langle N(M_h) \rangle). \quad (27)$$

We assume a Gaussian likelihood of the form

$$p(w^{\text{obs}}(\theta) | \langle N(M_h) \rangle, \mathbf{G}) \propto \exp \left[-\frac{1}{2} (w^{\text{mod}}(\theta) - w^{\text{obs}}(\theta))^T \mathbf{C}^{-1} (w^{\text{mod}}(\theta) - w^{\text{obs}}(\theta)) \right] \quad (28)$$

where $w^{\text{mod}}(\theta)$ is the 2PCF computed at $\mathbf{G}$ using the emulator and Equation (19). The covariance matrix $\mathbf{C} = \mathbf{C}^{\text{obs}} + \mathbf{C}^{\text{mod}}$, with $\mathbf{C}^{\text{mod}}$ the sum of the cosmic variance (here, quantified through the sample variance among the five separate light cone realizations of the SC SAM mocks) and the emulator’s predicted variance (obtained by propagating the GP’s variance predictions on the emulated quantities through Equation (19)). $\mathbf{C}^{\text{obs}}$ is the covariance matrix of the mock observation

estimated using the jackknife resampling algorithm with 64 patches in `treecorr` (see P. Norberg et al. 2009, for an in-depth discussion of the jackknife method in the context of galaxy 2PCFs). The number of patches was chosen after testing jackknife resampling with 4, 16, 64, and 128 patches and finding that the estimated covariance matrix converged at 64. This covariance matrix serves as a placeholder for the covariance matrix of observational uncertainties when comparing to real survey data. We specify a uniform prior over $\mathbf{G}$ with bounds set by the parameter ranges considered in the training set. Given we do not wish to bias the inferred HOD by imposing a restrictive prior on its shape, we specify a Gaussian prior on the galaxy number density in its place. This prior also serves to break degeneracies among similar 2PCFs across different regions of parameter space. The logarithm of the posterior the MCMC aims to maximize is then

$$\begin{aligned} \ln p(\langle N(M_h) \rangle, \mathbf{G} | w^{\text{obs}}(\theta)) \propto & -\frac{1}{2} (w^{\text{mod}}(\theta) - w^{\text{obs}}(\theta))^T \mathbf{C}^{-1} (w^{\text{mod}}(\theta) - w^{\text{obs}}(\theta)) \\ & - \frac{(n_{g, \text{mod}} - n_{g, \text{obs}})^2}{2(\sigma_{n_{g, \text{obs}}}^2 + \sigma_{n_{g, \text{mod}}}^2)} - \frac{1}{2} \ln(\det[\mathbf{C}]) \\ & - \sum_{i=1}^{N_{\text{input}}} \ln \Theta(\mathbf{G}_i - \mathbf{G}_{i, \text{min}}) + \ln \Theta(\mathbf{G}_{i, \text{max}} - \mathbf{G}_i). \end{aligned} \quad (29)$$

Here, $n_{g, \text{mod}}$ is the galaxy number density corresponding to w^{mod} , $n_{g, \text{obs}}$ is the galaxy number density corresponding to w^{obs} , $\sigma_{n_{g, \text{mod}}}^2$ is the variance in $n_{g, \text{mod}}$, and $\sigma_{n_{g, \text{obs}}}^2$ is the variance in $n_{g, \text{obs}}$ (taken to be the variance in number density across the five light cone realizations). The final term enforces the uniform top-hat prior on the physical galaxy properties, where $\mathbf{G}_{i, \text{min}}$ and $\mathbf{G}_{i, \text{max}}$ denote the minimum and maximum values of the top-hat prior for the i th component of $\mathbf{G}$, respectively.

We supply this form of the posterior to the `emcee` affine-invariant ensemble sampler to explore the joint posterior of the (M_*, SFR) thresholds. We initialize eight walkers in a random uniform fashion across the range of $\mathbf{G}$ thresholds discussed in Section 4.1, and run the sampler for 1000 steps with 200 steps discarded as burn in. We additionally discard any walkers with an acceptance fraction < 0.1 . To determine the HOD that corresponds to the recovered joint posterior over M_* –SFR space, we randomly draw 1000 samples from the flattened MCMC chains and use the emulator to compute the central, satellite, and total HOD for each sample. For each of these quantities we report the bin-by-bin median and associated 68% CI. We quantify the recovery performance in HOD space for the validation set in Section 5 and the test set in Section 6.

5. Validation

To validate the recovery performance in HOD space we compute the fractional and standardized residuals in the recovered HODs using the 300 SC SAM mock observations in the validation set discussed in Section 4.1. We emphasize that when assessing the MCMC recovery performance, the input $w(\theta)$ observations are computed using `treecorr` (following Equation (4)) and not the approximation in Equation (19). This ensures that our methodology accurately accounts for the differences between empirically derived measurements of $w(\theta)$ from galaxy survey data and those obtained via forward

¹⁰ <https://github.com/dfm/emcee>, v3.1

modeling. The 300 mock observations are divided into three groups, each containing 100 samples: (1) only M_* lower bound thresholds, (2) only SFR lower bound thresholds, and (3) joint (M_* , SFR) lower bound thresholds. The values of the lower bound thresholds are generated in accordance with the LHS routine described in Section 4.1, albeit with an additional modification. To obtain the one-dimensional thresholds from the two-dimensional LHS-generated thresholds, we “zero out” the SFR (M_*) to obtain M_* -only (SFR-only) thresholds. When running the MCMC we therefore restrict the physical galaxy property space to the singular dimension that corresponds to the one-dimensional threshold. The one-dimensional thresholds provide a useful test of our framework’s generalizability but will also serve as a good-faith comparison in Appendix A against the Z. Zheng et al. (2005) and J. E. Geach et al. (2012) parameterizations that are nominally used to model galaxy samples selected in this manner.

To characterize the recovery performance in (M_* , SFR) space, we compute both the arithmetic and standardized residuals in the recovered (M_* , SFR) tuples using the set of 300 mock observations discussed in Section 4.1. The arithmetic residuals are computed as $(x_{\text{truth},i} - x_{50,i})$, where $x_{\text{truth},i}$ is the ground-truth value corresponding to the mock observation and $x_{50,i}$ is the median of the recovered quantity. The standardized residuals are computed as

$$(\text{standardized residual})_{x,i} = \frac{x_{\text{truth},i} - x_{50,i}}{\sigma_{x,i}}, \quad (30)$$

where $\sigma_{x,i}$ is the upper or lower bound on the 68% CI that corresponds to the sign of the residual in the numerator (given the recovered 68% CIs are generally asymmetric). We note that for one-dimensional thresholds we report only the recovery performance with respect to the acting threshold (e.g., for M_* -only thresholds we do not try to recover the lower limit on the SFR).

For a single, randomly selected MCMC recovery, we show the corner plot over M_* –SFR space in Figure 5 and the corresponding recovered HOD in Figure 6. Figures 5 and 6 provide a visual confirmation that our nonparametric HOD approach can simultaneously infer the physical galaxy properties and HODs that correspond to a mock observation of the galaxy 2PCF. In Figure 5, we are able to identify the maximum of the joint M_* –SFR posterior within 1σ in both $\log_{10}(M_*)$ and $\log_{10}(\text{SFR})$ and provide an estimate for its uncertainty. In Figure 6, we recover both the central and satellite HOD to within $\approx 1\sigma$ across nearly all of the halo mass bins.

The arithmetic and standardized residuals for $\log_{10}(M_*)$ and $\log_{10}(\text{SFR})$ computed over all 300 SC SAM mock observations are shown in Figure 7. In Figure 7, we find that our approach successfully recovers the lower bound thresholds on all three types of galaxy samples analyzed. For the M_* -only samples we recover $\log_{10}(M_{*,\text{thresh}})$ with a median arithmetic residual of $0.006^{+0.006}_{-0.043}$ dex and a median standardized residual of $0.162^{+0.149}_{-1.879}$. For the SFR-only samples we recover $\log_{10}(\text{SFR}_{\text{thresh}})$ with a median arithmetic residual of $0.012^{+0.016}_{-0.016}$ dex and a median standardized residual of $0.318^{+0.457}_{-0.394}$. For joint (M_* , SFR) samples we recover $(\log_{10}(M_{*,\text{thresh}}), \log_{10}(\text{SFR}_{\text{thresh}}))$ with median arithmetic residuals of $-0.006^{+0.029}_{-0.204}$ dex in $\log_{10}(M_*)$ and $0.016^{+0.070}_{-0.043}$ dex in $\log_{10}(\text{SFR})$ and median standardized residuals of $-0.128^{+0.506}_{-1.208}$ in $\log_{10}(M_*)$ and $0.259^{+0.616}_{-0.502}$ in $\log_{10}(\text{SFR})$.

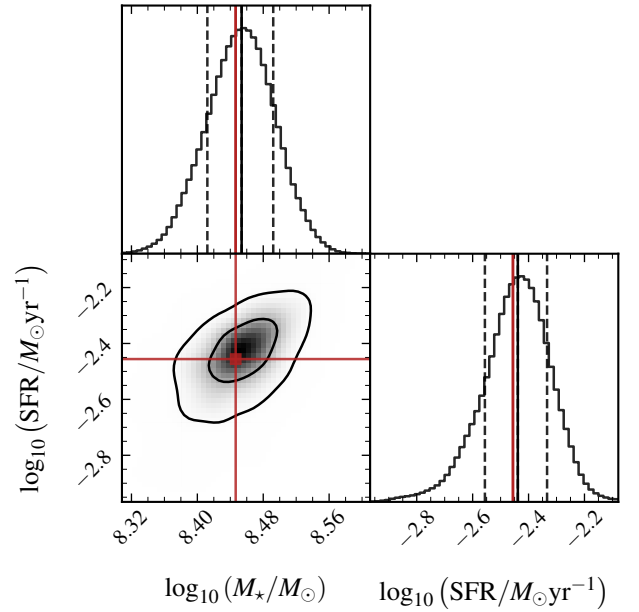


Figure 5. Corner plot from a single MCMC run for an SC SAM mock observation with thresholds $(\log_{10}(M_*/M_\odot), \log_{10}(\text{SFR}/M_\odot \text{ yr}^{-1})) \geq (8.45, -2.45)$ indicated with red solid lines. The black lines in the one-dimensional posteriors indicate the 16th (left dashed), 50th (solid), and 84th (right dashed) percentiles, with recovered values $\log_{10}(M_*/M_\odot) = 8.45^{+0.04}_{-0.04}$ and $\log_{10}(\text{SFR}/M_\odot \text{ yr}^{-1}) = -2.44^{+0.10}_{-0.12}$. The inner and outer contours in the two-dimensional posterior contain 39.3% and 86% of the samples, respectively (i.e., 1σ and 2σ contours for a two-dimensional Gaussian distribution). In this example, we are able to recover the “true” thresholds well within 1σ in both $\log_{10}(M_*)$ and $\log_{10}(\text{SFR})$ and with high precision.

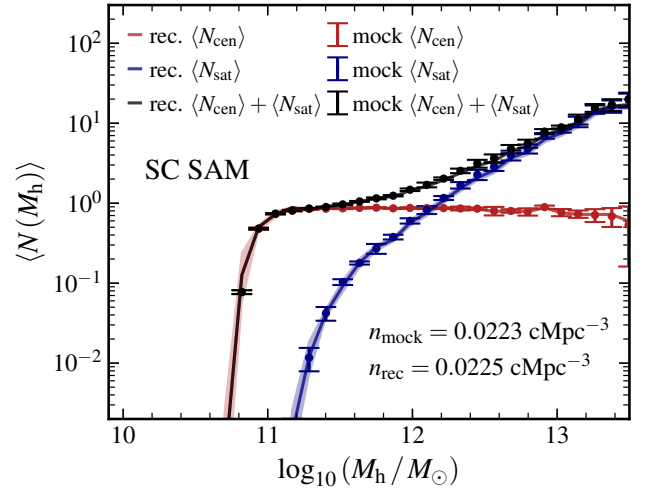


Figure 6. Posterior constraints on the recovered median central HOD (red solid), satellite HOD (blue solid), and total HOD (black solid) determined by sampling the MCMC chains in M_* –SFR space for the same SC SAM mock observation in Figure 5. The respective 68% CIs on the HODs are indicated as lightly shaded regions. The HODs corresponding to the mock observation are shown as data points, with errors bars representing the sample variance across the five SC SAM light cone realizations. The significant overlap between the 68% CIs on the recovered HOD curves and the mock HOD data points indicates that our recovery is both accurate and precise.

The arithmetic and standardized residuals in the recovered HODs for the 300 SC SAM mock observations are presented in Figure 8. Here, we find that our nonparametric HOD approach consistently recovers the central, satellite, and total HODs with a high degree of accuracy and precision. Accuracy is reflected in the median curves, which exhibit typical

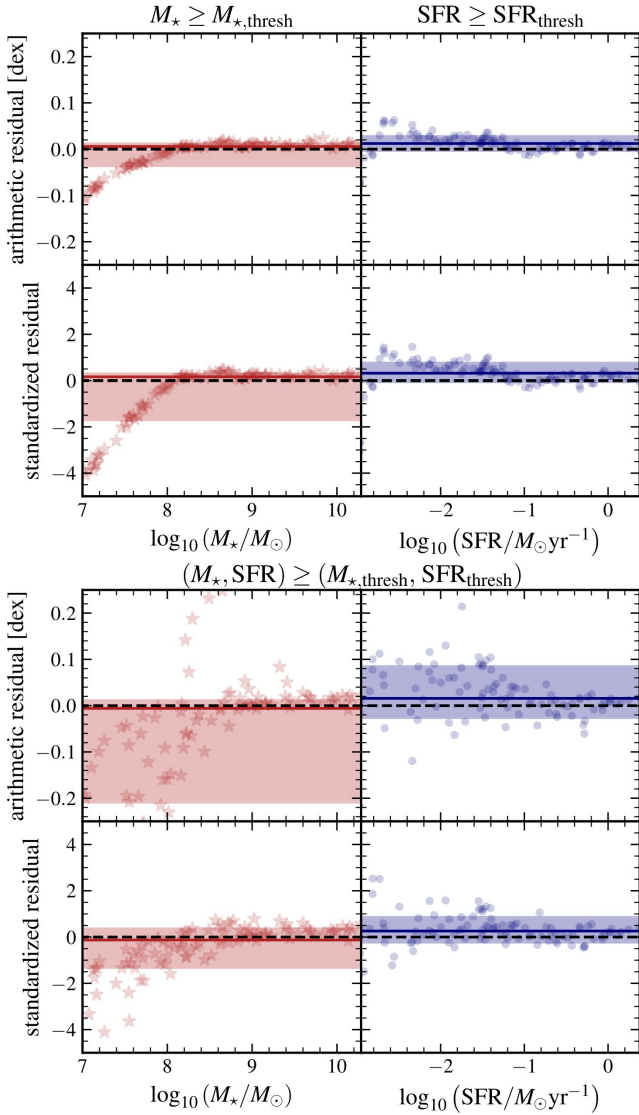


Figure 7. Arithmetic and standardized residuals for the true and MCMC-recovered lower bound thresholds on $\log_{10}(M_*)$ (left column) and $\log_{10}(\text{SFR})$ (right column) as a function of the “true” threshold value of the 300 SC SAM mocks for the three types of samples analyzed: only M_* (upper left panels), only SFR (upper right panels), and joint (M_*, SFR) samples (bottom four panels). Residuals from each of the 100 realizations in the three sets are plotted as lightly shaded markers, while the median and 68% CIs are plotted as solid lines and shaded regions, respectively. We are able to consistently recover both lower bound thresholds within ≈ 0.1 dex or $\approx 1.5\sigma$.

arithmetic residuals within 0.1 dex and standardized residuals within 1σ . Precision is evident from the relatively narrow 68% CIs in both the arithmetic and standardized residual spaces.

While we have shown that we can characterize the recovery performance in physical galaxy property space, we emphasize that the HOD is the primary quantity of interest in this analysis. For this reason, we consider the lower bound thresholds on (M_*, SFR) to be nuisance parameters in the context of the TNG100-1 test set, and therefore do not consider their recovery performance in this work.

6. Results

To evaluate the recovery performance of our methodology, we compute the arithmetic and standardized residuals in the recovered HODs for the test set of 300 TNG100-1 mock

observations in an analogous manner to Section 5. The TNG100-1 mocks are generated using the same selection criteria tuples as the SC SAM validation set mocks in Section 5, and provide a noncircular assessment of the method’s performance (given the emulator is trained on SC SAM mocks). We demonstrate the nonparametric HOD recovery for a single TNG100-1 mock in Figure 9. To generate the standardized residual curves shown in Figure 10, we rescale the 68% CIs of the recovered HOD posteriors by the nominal standardized residual in the HOD, producing median standardized residuals that are bounded between $\pm 1\sigma$. Under the assumption that the level of offset between TNG100-1 and the SC SAM is similar to the offset between the real Universe and the SC SAM, this makes our reported uncertainties a reasonable estimate for analysis of observational data.

In Figure 10, there is a noticeable difference in overall HOD recovery performance for the TNG100-1 mocks relative to the SC SAM mocks in Figure 8. While the central HODs are consistently recovered within ≈ 0.2 dex, our method consistently underestimates the satellite HOD by a significant margin ($\gtrsim 0.3$ dex) across nearly all types of galaxy samples. This behavior is directly observable in Figure 9, where the recovered satellite HOD falls below the mock data points, ultimately producing a worse fit than for the SC SAM mock HOD in Figure 6. We note however, that the halo mass range where the arithmetic residuals in the satellite HOD typically exceed 0.3 dex corresponds to lower mass bins where the satellite contribution to the total HOD is insignificant, as evidenced by the reasonable arithmetic residuals for the total HOD in the same bins.

We understand the biases in HOD recovery to arise in part from the differing treatment of satellite galaxies in the SC SAM and TNG100-1. As is evident in Figure 2, the M_* –SFR distributions of satellites in the two simulations have fairly distinct shapes, which produce different galaxy populations when considering lower bound threshold-selected samples. Further, the radial distribution of satellites in the two simulations—which strongly influences the clustering in the one-halo regime of $w(\theta)$ —differ as well (see Appendix B for details). An additional source of disparity can be attributed to differences in the distribution of satellite multiplicities in the two simulations. The one-halo term of the $w(\theta)$ is sensitive to both $\langle N_{\text{sat}} \rangle$ and $\langle N_{\text{sat}}^2 \rangle$ in Equation (14), both of which are have different, and distinct, characteristics across the SC SAM and TNG100-1. We describe these differences in Appendix C. The cumulative effect of these factors can be understood to modify the correspondence between $w(\theta)$ and the HOD that our method is reliant on, and the nature of which has been learned from the SC SAM. As a result, the recovery performance for satellite HODs in TNG100-1 is noticeably poorer than for satellite HODs in the SC SAM, and especially pronounced for galaxy samples selected using the SFR (see Section 7 for further discussion).

In order to compare the HOD recovery performance of our nonparametric approach relative to the standard parametric routine, we similarly compute the arithmetic and standardized residuals for the parametric HOD recoveries for TNG100-1 mocks following the framework detailed in Appendix A, and adopting the Z. Zheng et al. (2005) and J. E. Geach et al. (2012) parameterizations. Given these parameterizations are often used to model the HODs of M_* - and SFR-selected

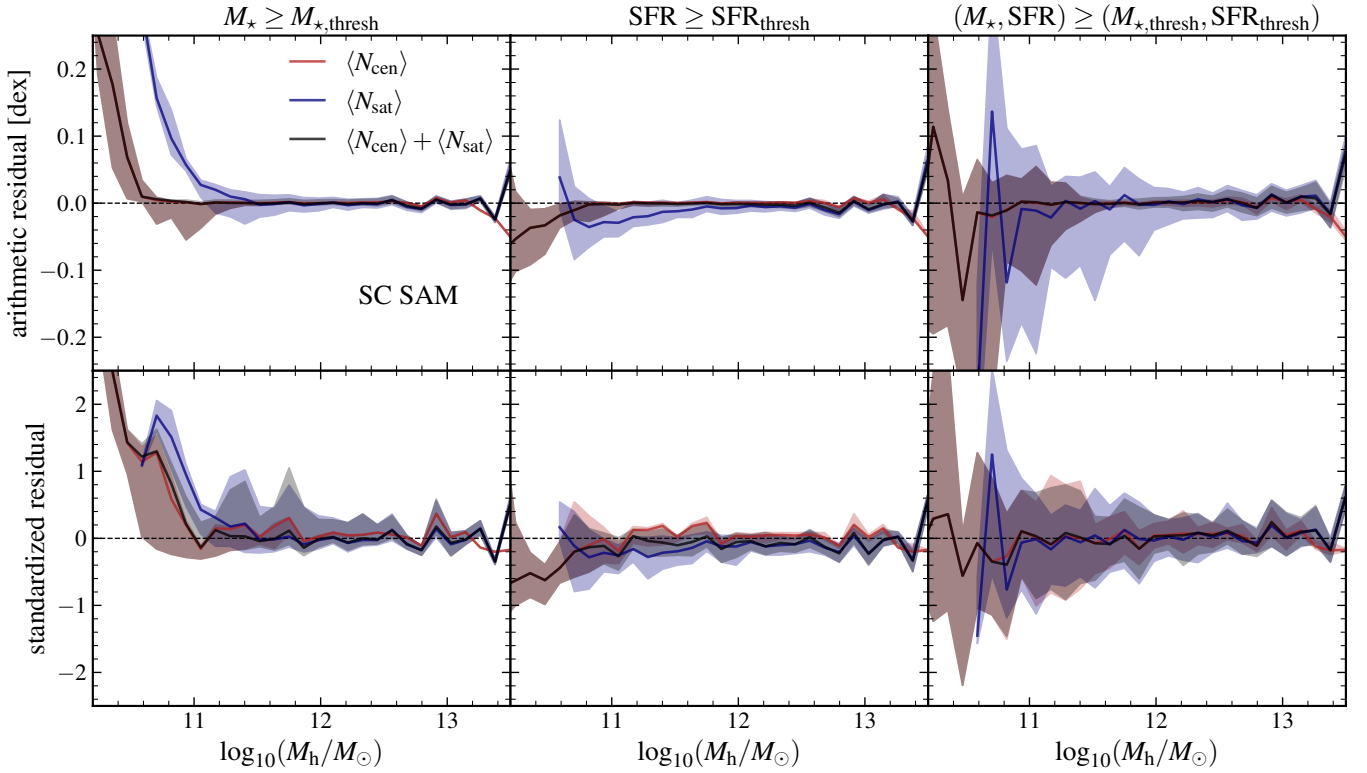


Figure 8. Upper row: arithmetic residuals between the “true” SC SAM validation set HODs and the recovered HODs obtained using our nonparametric approach. Lower row: same as the upper row, although now for standardized residuals computed using Equation (30). The left column displays residuals for galaxy samples selected with only an M_* lower bound threshold, the middle column displays residuals for samples selected with only an SFR lower bound threshold, and the right column displays residuals for galaxy samples selected above joint (M_*, SFR) thresholds. Each of the three columns displays the median arithmetic (upper row) and standardized (lower row) residuals computed across the 100 mock observations that fall into each category as a solid curve, and the corresponding 68% CI as a shaded region. We are able to consistently recover the “true” SC SAM HODs within ~ 0.1 dex or $\sim 1\sigma$.

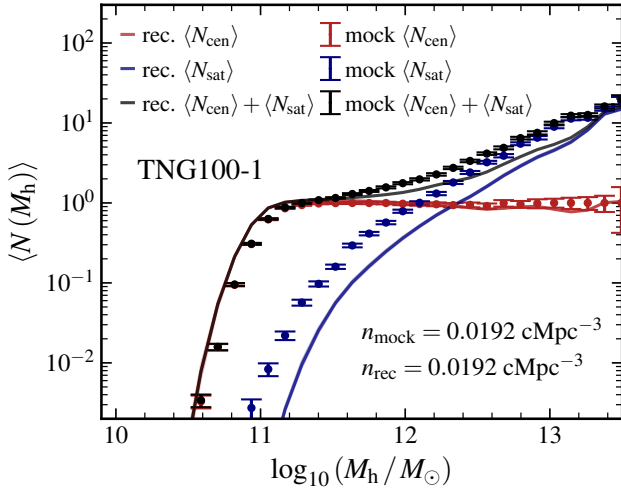


Figure 9. Same as Figure 6, but now for the TNG100-1 mock generated using the same set of (M_*, SFR) lower bound thresholds. The error bars on the mock HOD data points are computed assuming the galaxy counts in each mass bin follow Poisson statistics.

samples, the division of the 300 mocks into the three above-defined groups—only M_* , only SFR, and joint (M_*, SFR) —provides a natural basis for the comparison of HOD recovery performance across the nonparametric and parametric approaches. In Appendix A, we find that in general, our nonparametric HOD model recovers HODs with comparable or better precision and accuracy than parametric models across most halo mass bins.

7. Discussion

7.1. Caveats of the SC Semianalytic Model

An in-depth discussion of the caveats pertaining to the SC SAM light cone used in this paper is presented in L. Y. A. Yung et al. (2022). We draw attention to the SC SAM’s treatment of satellite galaxies and the potential impact on the satellite HOD. In particular, assumptions embedded within the SC SAM may starve satellite galaxies of fresh gas, leading to premature quenching and reduced star formation. Given the SFR is a primary selection criterion considered in this work, this process may reduce the satellite contribution to the HOD above a specified SFR threshold. Furthermore, the SC SAM implements a satellite position reassignment scheme in which the radial profiles of satellites are adjusted to reproduce the dark matter density profile of the host dark matter halo (i.e., a Navarro–Frenk–White (NFW) profile)—as is expected from N -body simulations. This has the potential to influence the one-halo term of the 2PCF, and subsequently influence the inferred satellite HOD. For these reasons, we will explore training our framework on alternative simulations in future work (see Section 7.3). While training our routine on additional simulations should improve the overall generality of the framework, the ability to reconstruct HODs will still be limited by the range of HODs modeled in those simulations.

7.2. Caveats of the Clustering Model

We also draw attention to the methodological limitations of the approach to modeling the one- and two-halo terms

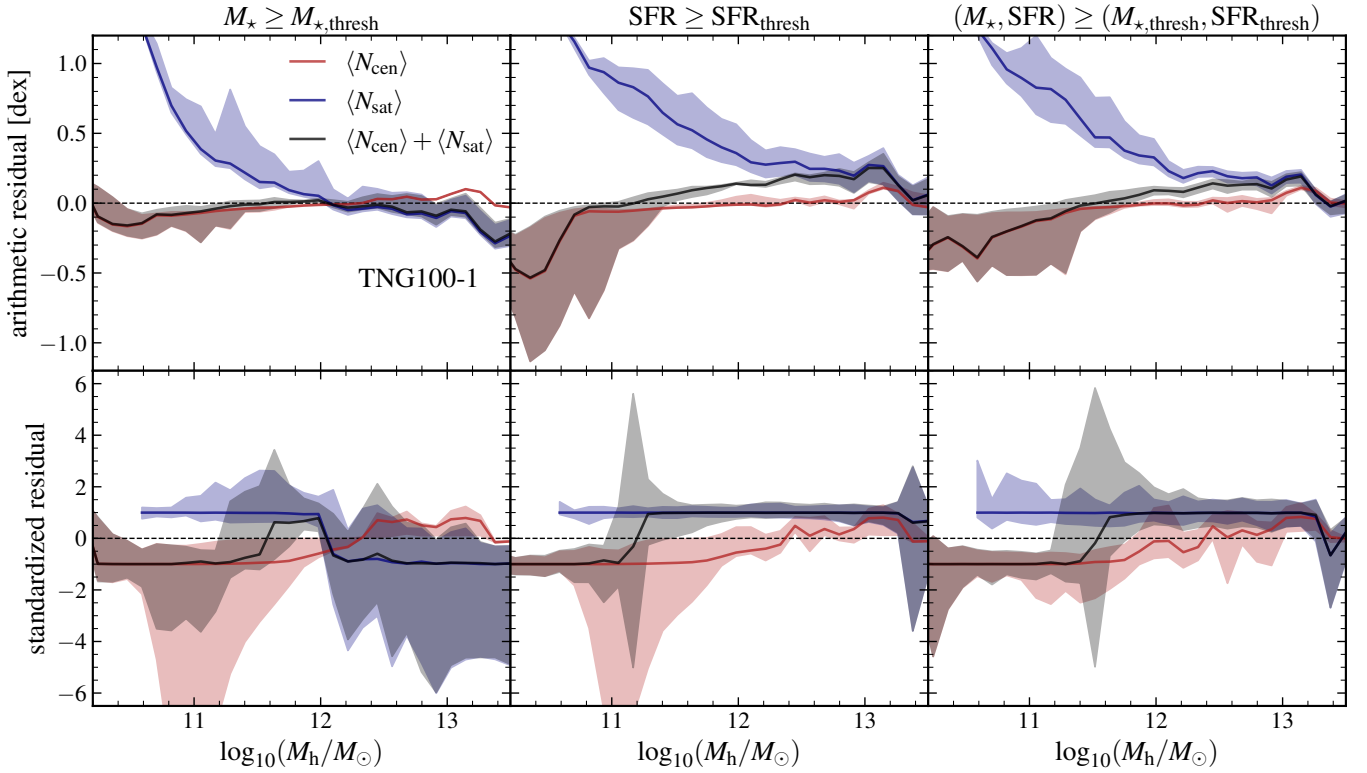


Figure 10. Same as Figure 8, but now for the TNG100-1 mocks. The central HOD is consistently recovered within ≈ 0.2 dex across the three different types of galaxy samples. The satellite HOD recovery performance is significantly worse than for the SC SAM mocks in Figure 8, especially at lower halo masses where the median arithmetic residual frequently exceeds 0.3 dex. To generate the standardized residual curves shown here, we rescale the 68% CIs of the recovered HOD posteriors by the nominal standardized residual in the HOD, producing median standardized residuals that are bounded between $\pm 1\sigma$.

employed in this work. While our approximation is an efficient alternative to numerically estimating the 2PCF, it assumes a scale-independent galaxy bias. As a result, the approximation in Equation (19) tends to underestimate the two-halo contribution in the one- to two-halo transition regime (i.e., $r \approx [0.5, 3]$ cMpc or $\theta \approx [0.008, 0.04]$ deg for $z \sim 0.76$). To mitigate the impact of this on the HOD, we estimate a scale-dependent additive correction to the 2PCF of dark matter. This is calculated as

$$w_{\text{DM}}^{\text{corr}}(\theta) = \left\langle \frac{w^{\text{treecorr}}(\theta) - w^{\text{1h}}(\theta)}{b_g^2} - w_{\text{DM}}^{\text{lin}}(\theta) \right\rangle, \quad (31)$$

where $w^{\text{treecorr}}(\theta)$ is the angular galaxy 2PCF estimated using `treecorr`, $w^{\text{1h}}(\theta)$ is the one-halo term in angular space, b_g is the scale-independent galaxy bias, $w_{\text{DM}}^{\text{lin}}(\theta)$ is the angular 2PCF of dark matter, and $\langle \dots \rangle$ represents the statistical average across the set of 300 SC SAM mock observations analyzed in Section 5. This serves as a semiempirical correction to better match mock galaxy 2PCFs with those of the SC SAM model. The form of the modified angular 2PCF of dark matter used in this analysis is then

$$w_{\text{DM}}^{\text{lin, mod}}(\theta) = w_{\text{DM}}^{\text{lin}}(\theta) + w_{\text{DM}}^{\text{corr}}(\theta). \quad (32)$$

We also note that the linear halo bias $b_h(M_h)$ in Equation (19) assumes a purely halo-mass-dependent relation. While this is a common assumption in the literature (e.g., J. L. Tinker et al. 2010), this formulation implicitly assumes the absence of galaxy assembly bias—the dependence of the HOD on halo properties other than mass on which halo clustering also depends (A. R. Zentner et al. 2014). Many studies have found

that assembly bias is an important ingredient in determining the strength of galaxy clustering (e.g., D. J. Croton et al. 2007). Notably, A. R. Zentner et al. (2014) found that ignoring assembly bias in HOD inference introduces significant systematic errors. Incorporating assembly bias or avoiding the explicit use of the halo bias model altogether is left to future work. Because our one-halo term is empirically derived from simulations that implicitly account for assembly bias, our vulnerability to its exclusion is primarily on large scales.

7.3. Future Extensions: Additional Simulations, Survey Data, and Physical Galaxy Properties

Future work will focus on expanding the applicability of our nonparametric HOD framework to alternative semianalytic and hydrodynamical simulations (e.g., Horizon Run 5: J. Lee et al. 2021; OpenUniverse2024: OpenUniverse Collaboration et al. 2025). Training and testing our method on different simulations will improve its overall robustness and ensure that we are not biased to any single simulation suite. The OpenUniverse2024 simulations are particularly relevant given they are designed to simulate the overlap of Rubin’s Legacy Survey of Space and Time Wide Fast Deep survey (Ž. Ivezić et al. 2019) and the Roman’s High-Latitude Wide-Area survey (Y. Wang et al. 2022). Training our approach on these simulations will prepare our method to analyze the eventual data from these two surveys.

We also intend to validate our nonparametric HOD method against real survey data by applying our methodology to a set of existing galaxy catalogs (e.g., eBOSS: S. Alam et al. 2021; HETDEX: K. Gebhardt et al. 2021; DESI: DESI Collaboration et al. 2025). For spectroscopic galaxy surveys such as HETDEX

and DESI, this will require the incorporation of redshift-space distortions into our clustering analysis, enabling the analysis of the redshift-space 2PCF $\xi(s, \mu)$ (A. J. S. Hamilton 1992, 1998).

In this work, we have used selection thresholds in stellar mass and SFR (as a tuple of M_* and SFR) as nuisance parameters to describe simulated galaxy samples. Observed galaxy samples are most frequently selected to exceed a given value of flux density (i.e., apparent magnitude) in a chosen filter or wavelength. This does not create a direct correspondence with a physical property such as stellar mass, even when the chosen filter falls in the rest-frame near-infrared where dust attenuation is minimal, due to observational incompleteness that is strongest near the selection limit. Inferring the distributions of physical galaxy properties from the observational flux density limit is difficult due to the challenges of forward modeling the galaxies' dust and metallicity and their effects on observed photometry. This suggests a future extension of our method to simultaneously infer both the HOD and the distribution of underlying physical properties that correspond to an observed galaxy sample. The mock galaxy catalog offers a long list of physical properties; if they match the real Universe well, our machinery can be used for this simultaneous inference.

8. Conclusions

The statistical relationship between the galaxies we observe and their host dark matter haloes is a product of both the underlying cosmology and the complex physics of galaxy formation and evolution. HODs present a computationally inexpensive framework to empirically quantify the galaxy-halo connection from measurements of galaxy clustering. However, standard parametric models lack the necessary flexibility to describe the complete space of physically plausible galaxy samples that are anticipated from current and next-generation galaxy surveys.

In this work, we have introduced a nonparametric approach to HOD modeling designed to resolve the issues associated with parametric modeling routines. In this setting, we did not assume a functional form for the HOD. We trained a GP-based algorithm to learn the mappings among the physical galaxy property space, the galaxy 2PCF space, and the HOD space using a single suite of SC SAM light cones. Incorporating this emulator into a Bayesian inference framework, we derived simultaneous constraints on physical galaxy property limits and HODs when provided with a mock measurement of the galaxy 2PCF. In this initial implementation, our nonparametric HOD framework reliably recovers the central, satellite, and total HODs within 0.1 dex for the SC SAM validation set and consistently within 0.2 dex for the TNG100-1 test set, and with accurate uncertainties that have been tuned to estimate those expected in the analysis of observational data.

Using a parametric HOD recovery with the Z. Zheng et al. (2005) and J. E. Geach et al. (2012) parameterizations as benchmark models, we find that our nonparametric HOD model recovers HODs with comparable or better precision and accuracy than parametric inference across most halo mass bins. Our nonparametric HOD approach also benefits from a substantial reduction in computational complexity provided by the GP-based emulator, reducing the average time for a likelihood evaluation.

The development of a nonparametric method for HOD inference is a crucial step toward fully leveraging the statistical power of current and next-generation galaxy surveys. As

observational data become increasingly precise and diverse—probing a wider range of galaxy populations, environments, and redshifts—the limitations of rigid parametric models will become more pronounced. These models, by construction, embed assumptions about the form of the galaxy-halo connection that may not hold across different galaxy selections or cosmological regimes. A flexible approach, such as that introduced in this paper, therefore becomes essential for capturing the diversity and complexity of real galaxy populations.

Robust modeling of the galaxy-halo connection is particularly important when using large-scale structure data for cosmological or astrophysical inference. Inaccurate or oversimplified HOD assumptions can propagate into biased constraints on the quantities of interest. By avoiding strong prior assumptions on the shape of the HOD, our nonparametric HOD framework provides a more reliable and interpretable foundation for such analyses.

Acknowledgments

The authors would like to thank David Weinberg and Joshua Kable for valuable comments on the manuscript. J.K. and E.G. acknowledge support from NSF grant AST-2206222 and from the U.S. Department of Energy, Office of Science, Office of High Energy Physics Cosmic Frontier Research program under Award No. DE-SC0010008. Support for K.I. was provided by NASA through the NASA Hubble Fellowship grant HST-HF2-51508 awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555. A.Y. is supported by a Giacconi Fellowship from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Incorporated, under NASA contract HST NAS5-26555 and JWST NAS5-03127. J.K. and E.G. acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here.

Software: Python, astropy, pyccl, treecorr, GPy, halomod.

Appendix A

Parametric Halo Occupation Distribution Modeling

To demonstrate the improvement in HOD modeling that our nonparametric approach may provide relative to the standard parametric routine, we apply parametric HOD modeling using the Z. Zheng et al. (2005) and J. E. Geach et al. (2012) parameterizations to the same set of 300 TNG100-1 mock observations that we test our nonparametric approach on in Section 6. To perform this investigation, we make use of the Python package `halomod`¹¹ (S. Murray et al. 2013, 2021). `halomod` enables efficient HOD modeling that can be easily interfaced with an MCMC framework to constrain the parameters of a given HOD parameterization from a measurement of the 2PCF. In this work, we consider two HOD parameterizations; (1) the five-parameter model introduced in Z. Zheng et al. (2005) designed to describe stellar-mass-selected galaxy samples, and (2) the eight-parameter model introduced in J. E. Geach et al. (2012) designed to describe

¹¹ <https://github.com/halomod/halomod,v.2.2>.

SFR-selected galaxy samples. The five-parameter Z. Zheng et al. (2005) model has central and satellite HODs of the form

$$\langle N_{\text{cen}}(M_h) \rangle = \frac{1}{2} \left[1 + \text{erf} \left(\frac{\log_{10}(M_h) - \log_{10}(M_{\text{min}})}{\sigma_{\log M}} \right) \right], \quad (\text{A1})$$

$$\langle N_{\text{sat}}(M_h) \rangle = \langle N_{\text{cen}}(M_h) \rangle \Theta(M_h - M_{\text{cut}}) \left(\frac{M_h - M_{\text{cut}}}{M_1} \right)^\alpha, \quad (\text{A2})$$

where erf is the Gauss error function, M_{min} is the minimum halo mass that hosts galaxies, $\sigma_{\log M}$ is the characteristic transition width, Θ is the Heaviside function, M_{cut} is the minimum halo mass that hosts satellites, M_1 is a normalization factor, and α is the power-law index. In Equation (A2), we have chosen the common version of the Z. Zheng et al. (2005) parameterization that conditions the satellite HOD on the central HOD. This ensures that when $\langle N_{\text{cen}}(M_h) \rangle = 0$, $\langle N_{\text{sat}}(M_h) \rangle = 0$ as well. The eight-parameter J. E. Geach et al. (2012) has a central HOD

$$\begin{aligned} \langle N_{\text{cen}}(M_h) \rangle &= F_c^B (1 - F_c^A) \exp \\ &\times \left[-\frac{(\log_{10}(M_h) - \log_{10}(M_c))^2}{2\beta_{\log M}^2} \right] \\ &+ F_c^A \left[1 + \text{erf} \left(\frac{\log_{10}(M_h) - \log(M_c)}{\beta_{\log M}} \right) \right], \end{aligned} \quad (\text{A3})$$

where $F_c^{A,B}$ are normalization factors ranging from zero to one, M_c is the mean of the Gaussian distribution and low-mass cutoff for the step function, and $\beta_{\log M}$ is the width of the Gaussian and transition width of the step function. The J. E. Geach et al. (2012) satellite HOD has the form

$$\begin{aligned} \langle N_{\text{sat}}(M_h) \rangle &= F_s \left(\frac{M_h}{M_{\text{min}}} \right)^\gamma \\ &\times \left[1 + \text{erf} \left(\frac{\log_{10}(M_h) - \log_{10}(M_{\text{min}})}{\delta_{\log M}} \right) \right]. \end{aligned} \quad (\text{A4})$$

where M_{min} is the minimum mass of haloes that host satellites, F_s is the mean number of galaxies at M_{min} , $\delta_{\log M}$ is the satellite transition scale, and γ is the power-law index.

To constrain the HOD from a measurement of the 2PCF using either of the above parameterizations, we perform HOD parameter space exploration using the MCMC package `emcee` (D. Foreman-Mackey et al. 2013). This involves constructing a forward model of the 2PCF as a function of the HOD parameters. To do so, we make use of the `halo_model` module in `halomod`. We use the `halo_model` module to compute the one-halo term of the galaxy 2PCF using Equation (14), and the two-halo term using the linear large-scale galaxy bias approximation in Equation (15). This formulation of the one- and two-halo terms is intended to improve the agreement between our approach to estimating $w(\theta)$ detailed in Section 3.1. We specify the normalized radial number density profile of satellites $u_s(r|M_h)$ to be an NFW profile (J. F. Navarro et al. 1996), as is expected for satellites

Table 1

Uniform Top-hat Priors for the Z. Zheng et al. (2005) Parameters

Parameter	Prior
$\log_{10}(M_{\text{min}})$	$\mathcal{U}(9.0, 13.0)$
$\log_{10}(M_1)$	$\mathcal{U}(9.0, 13.5)$
α	$\mathcal{U}(0, 2.0)$
$\sigma_{\log M}$	$\mathcal{U}(0, 0.5)$
$\log_{10}(M_{\text{cut}})$	$\mathcal{U}(9.0, 13.5)$

Table 2

Uniform Top-hat Priors for the J. E. Geach et al. (2012) Parameters

Parameter	Prior
$\log_{10}(M_c)$	$\mathcal{U}(9.0, 13.0)$
$\log_{10}(M_{\text{min}})$	$\mathcal{U}(9.0, 13.5)$
γ	$\mathcal{U}(0, 2.0)$
$\beta_{\log M}$	$\mathcal{U}(0, 0.5)$
F_c^A	$\mathcal{U}(0, 1.0)$
F_c^B	$\mathcal{U}(0, 1.0)$
F_s	$\mathcal{U}(0, 1.0)$
$\delta_{\log M}$	$\mathcal{U}(0, 0.5)$

in TNG100-1 (D. Nelson et al. 2019). We implement the same MCMC configuration as detailed in Section 4.3, albeit with a modified posterior of the form

$$p(\theta_{\text{HOD}} | w^{\text{obs}}(\theta)) \propto p(w^{\text{obs}}(\theta) | \theta_{\text{HOD}}) p(\theta_{\text{HOD}}) \quad (\text{A5})$$

where w^{obs} is the mock observation of the 2PCF being analyzed and θ_{HOD} are the set of HOD parameters corresponding to either of the Z. Zheng et al. (2005) or J. E. Geach et al. (2012) parameterizations. Adopting a Gaussian likelihood for the 2PCF, a Gaussian number density penalty (analogous to that detailed in Section 4.3), and a uniform top-hat prior over the HOD parameters, the logarithm of the posterior the MCMC aims to maximize is

$$\begin{aligned} &\ln p(w^{\text{obs}}(\theta) | \theta_{\text{HOD}}) \propto \\ &-\frac{1}{2} (w^{\text{mod}}(\theta) - w^{\text{obs}}(\theta))^T \mathbf{C}^{\text{obs}-1} (w^{\text{mod}}(\theta) - w^{\text{obs}}(\theta)) \\ &\quad - \frac{(n_{\text{g, mod}} - n_{\text{g, obs}})^2}{2\sigma_{n_{\text{g, obs}}}^2} \\ &\quad - \sum_{i=1}^{N_\theta} \ln \Theta(\theta_i - \theta_{i, \text{min}}) + \ln \Theta(\theta_{i, \text{max}} - \theta_i), \end{aligned} \quad (\text{A6})$$

where $\mathbf{C}^{\text{obs}}$ is the covariance matrix of the mock observation (computed using jackknife resampling), w^{mod} is the 2PCF computed at θ_{HOD} , $n_{\text{g, mod}}$ is the galaxy number density corresponding to w^{mod} , $n_{\text{g, obs}}$ is the galaxy number density corresponding to w^{obs} , and $\sigma_{n_{\text{g, obs}}}^2$ is the variance in $n_{\text{g, obs}}$. The final term enforces the uniform top-hat prior on the HOD parameters, where N_θ is the number of parameters, and $\theta_{i, \text{min}}$ and $\theta_{i, \text{max}}$ denote the minimum and maximum values of the top-hat prior for the i th component of θ_{HOD} , respectively. The values of the parameter bounds used for each of the Z. Zheng et al. (2005) and J. E. Geach et al. (2012) parameterizations are listed in Tables 1 and 2, respectively.

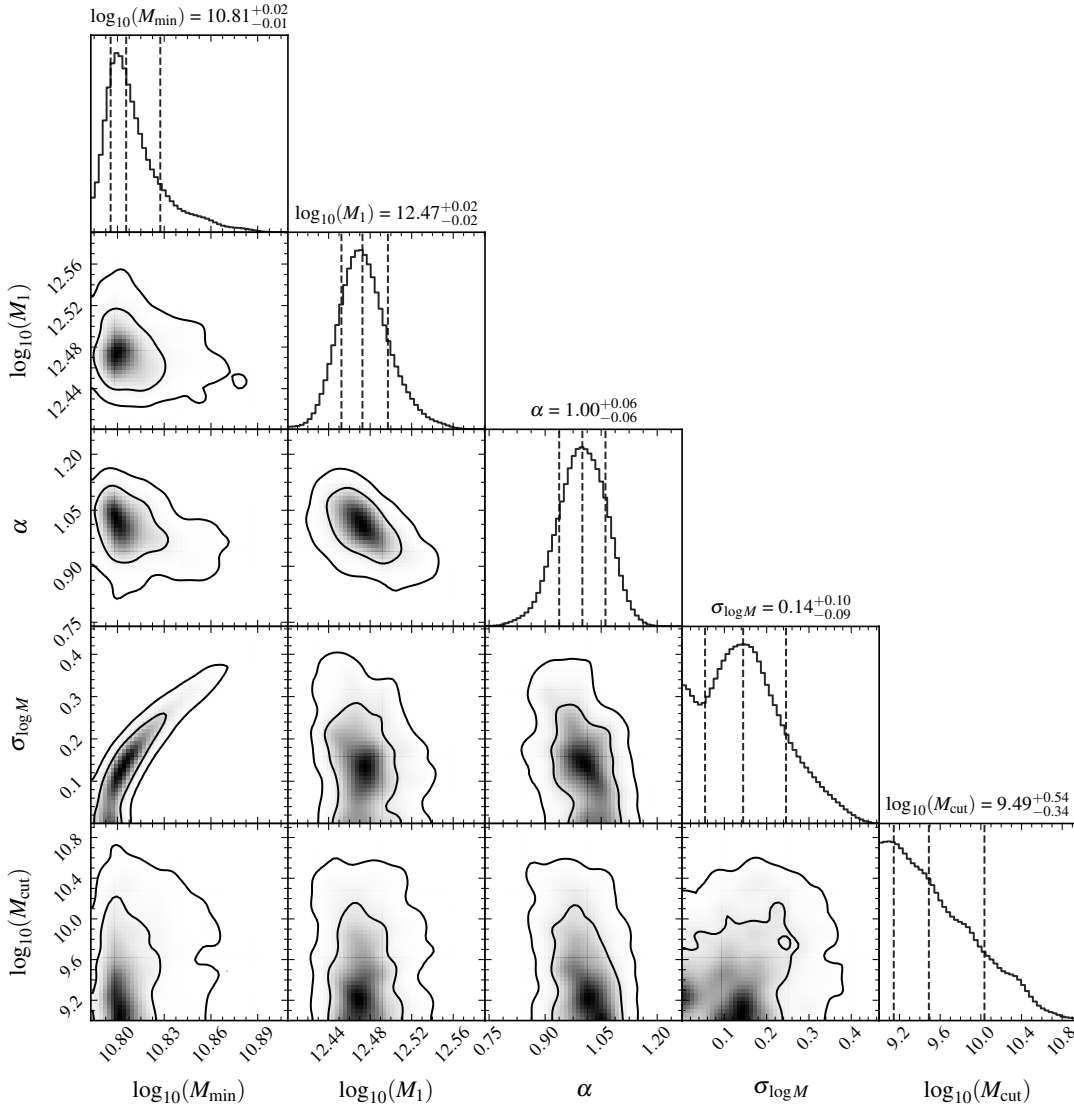


Figure 11. An MCMC corner plot for the Z. Zheng et al. (2005) parameterization using a TNG100-1 mock 2PCF observation with $(\log_{10}(M_*/M_\odot), \log_{10}(\text{SFR}/M_\odot \text{ yr}^{-1})) \geq (8.45, -2.45)$ (i.e., the same input 2PCF as used in Figure 5). The black lines in the one-dimensional posteriors indicate the 16th (left dashed), 50th (middle dashed), and 84th (right dashed) percentiles. The inner and outer contours in the two-dimensional posterior contain 68% and 95% of the samples, respectively. Because the galaxy 2PCFs we are fitting were generated from a nonparametric HOD, no “true” set of HOD parameters is available to plot.

As a baseline comparison for our nonparametric HOD model, we run MCMC recoveries using both the Z. Zheng et al. (2005) and J. E. Geach et al. (2012) parameterizations for the 300 TNG100-1 mock observations discussed in Section 6. Equipped with the parameter chains corresponding to each mock observation, we randomly draw 1000 samples and compute the median central, satellite, and total HODs and their respective 68% CIs. We then characterize the parametric HOD recovery performance by computing the median and 68% CIs for both the arithmetic and standardized residuals as done for the nonparametric HOD recoveries conducted in Section 6. In Figures 11, 12, 13, and 14 we show the MCMC corner plots and corresponding HOD posteriors for both the Z. Zheng et al. (2005) and J. E. Geach et al. (2012) parameterizations when provided with the same TNG100-1 mock observation used in Figure 9. In Figures 13 and 14, the recovered “best-fit” parametric HODs generally fail to capture the unique shape of the central and satellite HODs. Figures 15 and 16 demonstrate

the statistical behavior of the HOD recovery performance for the Z. Zheng et al. (2005) and J. E. Geach et al. (2012) parameterizations, respectively. In Figure 15, the Z. Zheng et al. (2005) parametric recoveries of the M_* -selected samples exhibit comparable arithmetic residuals to our nonparametric recoveries in the central HOD, and larger arithmetic residuals in the satellite HOD (excluding $\log_{10}(M_h/M_\odot) \lesssim 11$). For the SFR- and joint (M_* , SFR)-selected samples in Figure 15, the Z. Zheng et al. (2005) parametric recoveries have comparable or larger arithmetic residuals to our nonparametric recoveries in the central HOD and in the satellite HOD for $\log_{10}(M_h/M_\odot) \gtrsim 11.5$. In Figure 15, the standardized residuals in the central HOD for the Z. Zheng et al. (2005) parametric recoveries are reasonably consistent with our nonparametric recoveries, however, the standardized residuals for the satellite HOD are heavily biased for $\log_{10}(M_h/M_\odot) \gtrsim 11$ at $\gtrsim 3\sigma$, indicating that the predicted uncertainties for the Z. Zheng et al. (2005) satellite HODs are underestimated. In Figure 16,

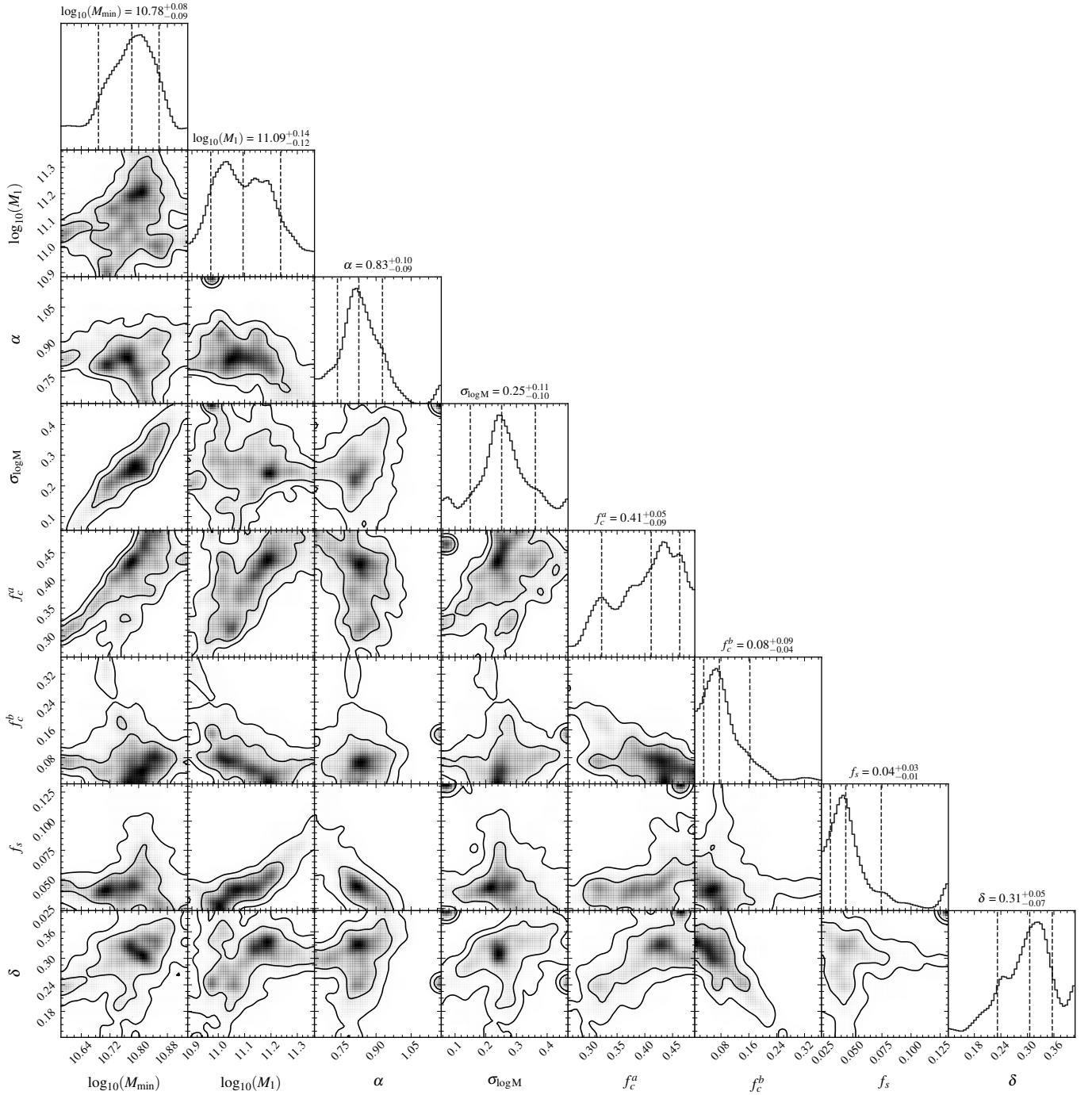


Figure 12. The same as Figure 11, but now for the J. E. Geach et al. (2012) parameterization, with the same input TNG100-1 mock 2PCF observations.

the J. E. Geach et al. (2012) parametric recoveries of the M_* -selected samples exhibit comparable or larger arithmetic residuals to our nonparametric recoveries in the central HOD and the satellite HOD for $\log_{10}(M_h/M_\odot) \gtrsim 11$. For the SFR- and joint (M_* , SFR)-selected samples in Figure 16, the J. E. Geach et al. (2012) parametric recoveries have

comparable or larger arithmetic residuals to our nonparametric recoveries in the central HOD and in the satellite HOD for $\log_{10}(M_h/M_\odot) \gtrsim 11.5$. In Figure 16, the standardized residuals in the central and satellite HODs for the J. E. Geach et al. (2012) parametric recoveries are biased toward overestimating the HOD ($\approx 2\sigma$ for $\langle N_{\text{cen}} \rangle$ and $\gtrsim 1\sigma$ for $\langle N_{\text{sat}} \rangle$).

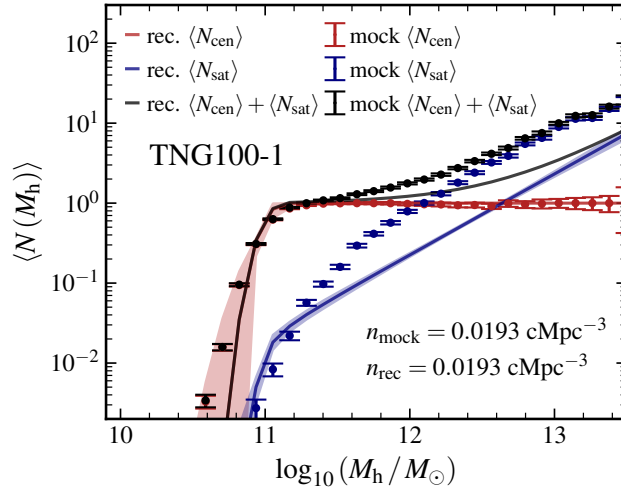


Figure 13. The median recovered central (solid red), satellite (solid blue), and total (solid black) HODs corresponding to the corner plot in Figure 11 for the Z. Zheng et al. (2005) parameterization. The 68% CIs for each curve are indicated as shaded regions. The “true” HODs of the mock observation—the same as in Figure 9—are shown as data points. The Z. Zheng et al. (2005) parameterization closely reproduces the central HOD, yet significantly underestimates the satellite HOD of the TNG100-1 mock.

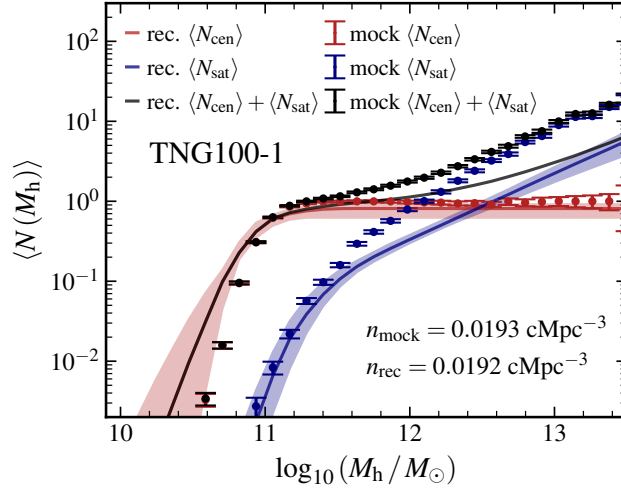


Figure 14. The same as Figure 13, but now for the J. E. Geach et al. (2012) parameterization. The J. E. Geach et al. (2012) parameterization consistently underestimates both the central HOD and the satellite HOD for $\log_{10}(M_h/M_\odot) \gtrsim 11$.

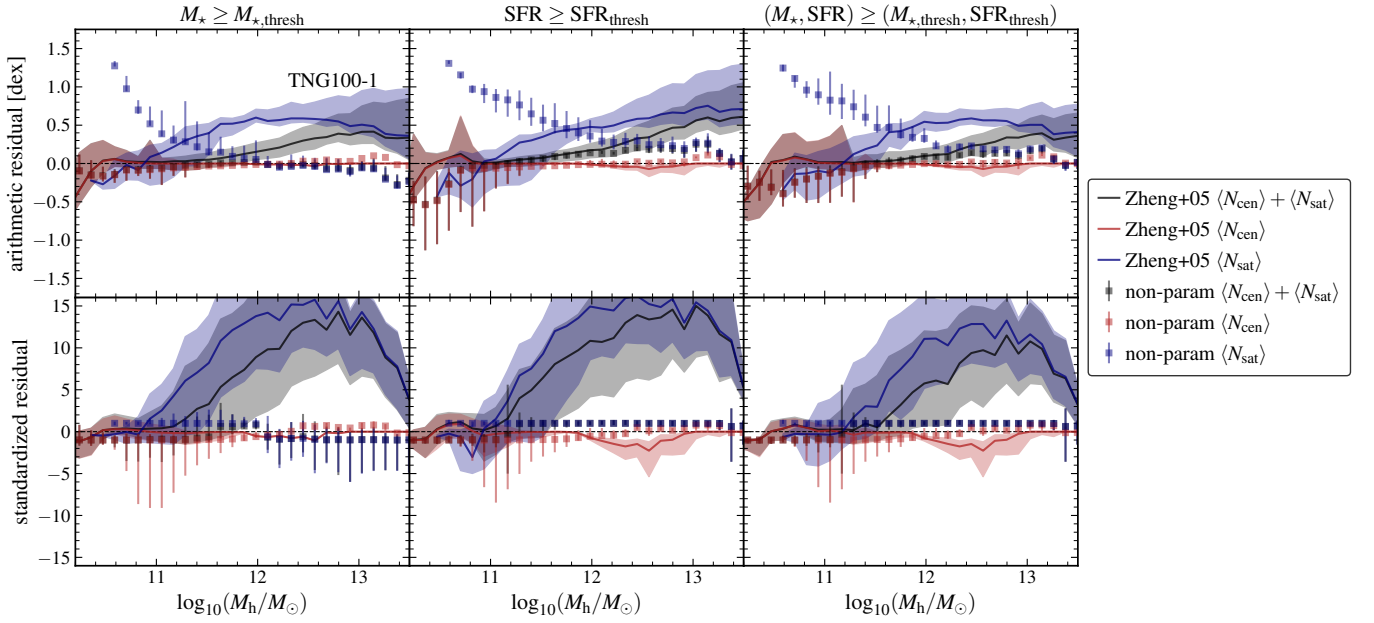


Figure 15. Same as Figure 10, although now using the parametric HOD modeling routine with the Z. Zheng et al. (2005) parameterization. The arithmetic and standardized residuals corresponding to the nonparametric HOD recoveries over the TNG100-1 mocks are plotted as data points for direct comparison with the parametric recovery curves. The error bars on the nonparametric data points represent 68% CIs (previously represented as shaded regions in Figure 10).

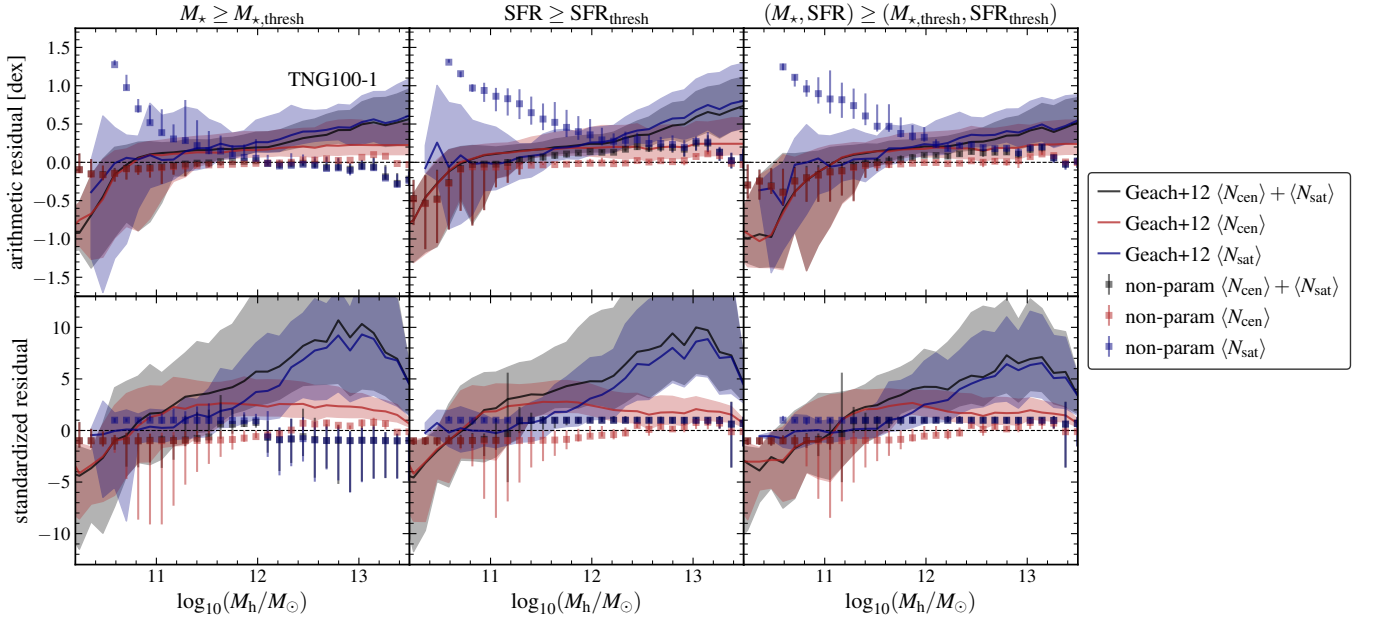


Figure 16. The same as Figure 15, although now using the parametric HOD modeling routine with the J. E. Geach et al. (2012) parameterization.

Appendix B Radial Distribution of Satellites

The distributions of central–satellite separations in the SC SAM and TNG100-1 are shown in the left panel of Figure 17. Notably, the SC SAM has a maximum that is $\approx 4\times$ smaller than TNG100-1. Consequently, the distribution of host halo virial radius normalized central–satellite separations shown in the right panel of Figure 17 indicates that the radial profile of satellites falls off much more steeply in the SC SAM than in

TNG100-1, which exhibits a steady decline out to $r/R_{\text{vir}} \approx 3.0$. We believe that the discrepancies between the radial distributions of satellites in haloes in the SC SAM and TNG100-1, coupled with the enhanced quenching of satellites in the SC SAM (as described in Section 7), are responsible for the marked difference in nonparametric HOD recovery performance between the SC SAM and TNG100-1 mocks, given that they modify the nominal $w(\theta)$ –HOD correspondence our method has identified.

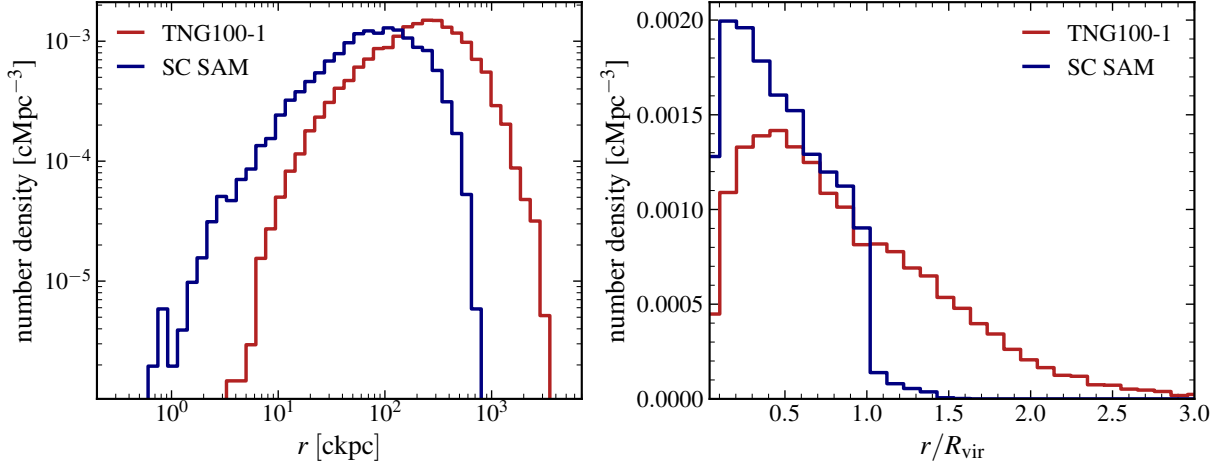


Figure 17. Left panel: distribution of central–satellite separations for galaxies with $(\log_{10}(M_*/M_\odot), \log_{10}(\text{SFR}/M_\odot \text{ yr}^{-1})) \geq (7, -3)$ in TNG100-1 (red) and the SC SAM (blue). The TNG100-1 distribution is shifted toward larger separations, reaching a maximum at $r \approx 4 \times 10^2$ ckpc, while the SC SAM distribution reaches a maximum at $r \approx 10^2$ ckpc. Right panel: the same as the left panel, although now normalizing the central–satellite separations by the host halo virial radius. The SC SAM exhibits a distinctive drop-off around $r/R_{\text{vir}} \sim 1.0$, while TNG declines steadily over the range $r/R_{\text{vir}} \sim [0.5, 3.0]$.

Appendix C

SC Semianalytic Model and TNG100-1 Halo Occupation Distribution Comparison

Figure 18 displays the ratio between the TNG100-1 and SC SAM $\langle N_{\text{cen}} \rangle$, $\langle N_{\text{sat}} \rangle$, and $\langle N_{\text{sat}}^2 \rangle$ for the ranges of M_* and SFR lower bound thresholds considered in this work. Notably, there are significant differences in the HODs of the two simulations for the same lower bound thresholds. The discrepancy is especially significant at low to intermediate halo masses (e.g., $\log_{10}(M_*/M_\odot) \lesssim 12$), and generally decreases as halo mass increases. In particular, the differences in magnitude of $\langle N_{\text{cen}} \rangle$,

$\langle N_{\text{sat}} \rangle$, and $\langle N_{\text{sat}}^2 \rangle$ across the two simulations increase as the lower bound threshold on the SFR increases, indicating that the corresponding galaxy samples are indeed very different. As a result, an emulator (such as ours) trained on any single simulation (e.g., the SC SAM) would predictably perform poorer on an alternative simulation (e.g., TNG100-1) if the space of HODs covered by the test set lies outside of the space of HODs covered by the training set—such is the case that we have identified here. Improving the robustness of the emulator by training on a range of simulations, spanning a larger space of HODs, would likely improve our framework’s generalizability.

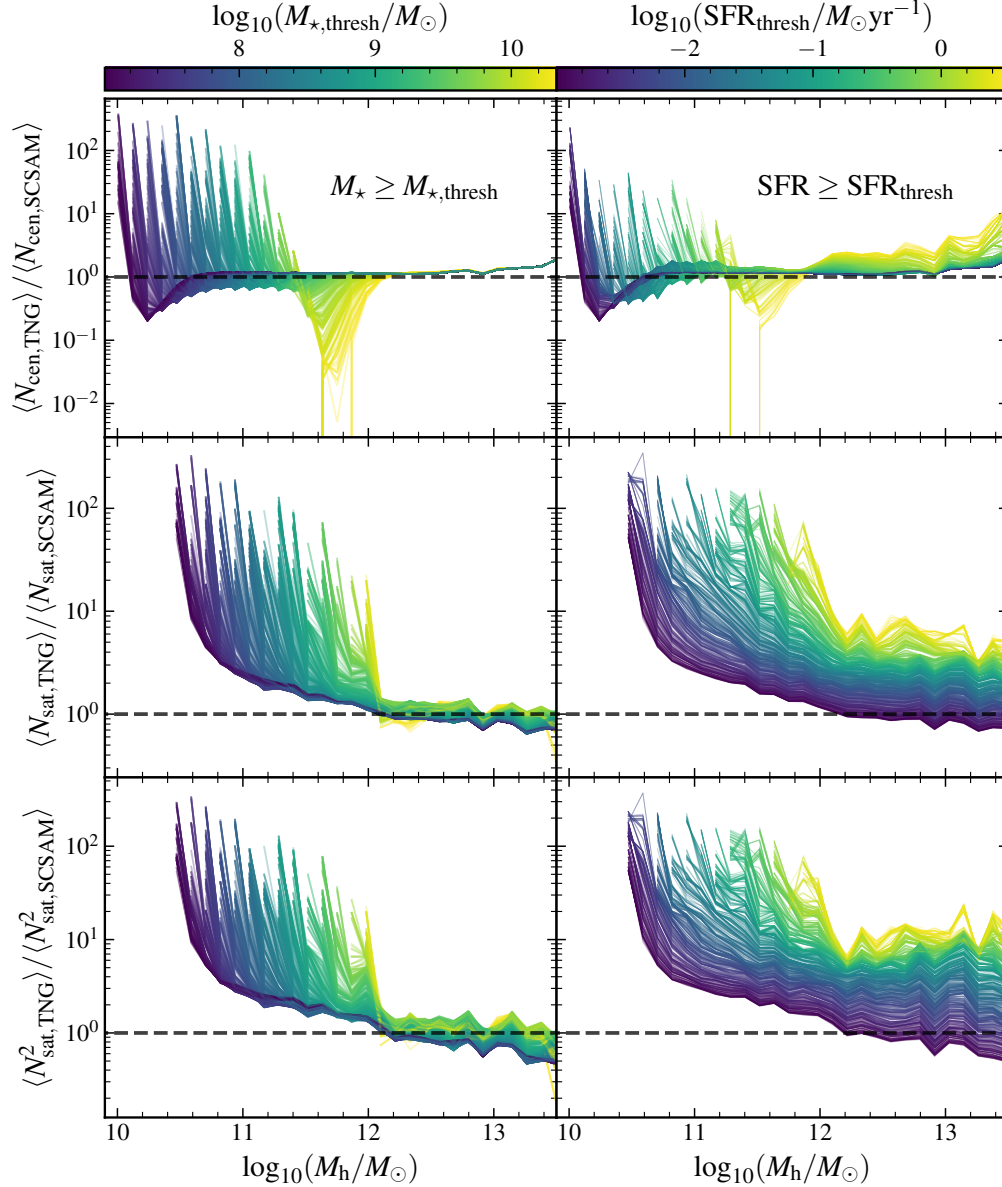


Figure 18. Left column: the ratios between the TNG100-1 and SC SAM central HODs (upper row), satellite HODs (middle row), and $\langle N_{\text{sat}}^2 \rangle$ (bottom row) for variable lower bound thresholds of M_* . Right column: same as the left column, but now for variable lower bound thresholds of the SFR. The range of amplitudes covered by the curves is indicative of the substantial differences in HODs between the two simulations.

ORCID iDs

Jacob Kennedy  <https://orcid.org/0009-0001-4745-3555>
 Eric Gawiser  <https://orcid.org/0000-0003-1530-8713>
 Kartheik G. Iyer  <https://orcid.org/0000-0001-9298-3523>
 L. Y. Aaron Yung  <https://orcid.org/0000-0003-3466-035X>

References

- Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, *ApJS*, **182**, 543
- Ade, P. A. R., Aghanim, N., Arnaud, M., et al. 2016, *A&A*, **594**, A13
- Akeson, R., Armus, L., Bachelet, E., et al. 2019, arXiv, arXiv:1902.05569
- Alam, S., Aubert, M., Avila, S., et al. 2021, *PhRvD*, **103**, 083533
- Asgari, M., Mead, A. J., & Heymans, C. 2023, *OJAp*, **6**, 39
- Berlind, A. A., & Weinberg, D. H. 2002, *ApJ*, **575**, 587
- Berlind, A. A., Weinberg, D. H., Benson, A. J., et al. 2003, *ApJ*, **593**, 1
- Bernardeau, F., Colombi, S., Gaztañaga, E., & Scoccimarro, R. 2002, *PhR*, **367**, 1
- Chaves-Montero, J., Angulo, R. E., Schaye, J., et al. 2016, *MNRAS*, **460**, 3100
- Chisari, N. E., Alonso, D., Krause, E., et al. 2019, *ApJS*, **242**, 2
- Cochrane, R. K., Best, P. N., Sobral, D., et al. 2017, *MNRAS*, **469**, 2913
- Cole, S., & Lacey, C. 1996, *MNRAS*, **281**, 716
- DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv:1611.00036
- DESI Collaboration, Abdul-Karim, M., Aguilar, J., et al. 2025, *PhRvD*, **112**, 083514
- Colless, M., Dalton, G., Maddox, S., et al. 2001, *MNRAS*, **328**, 1039
- Conroy, C., Wechsler, R. H., & Kravtsov, A. V. 2006, *ApJ*, **647**, 201
- Cooray, A., & Sheth, R. 2002, *PhR*, **372**, 1
- Coupon, J., Kilbinger, M., McCracken, H. J., et al. 2012, *A&A*, **542**, A5
- Croton, D. J., Gao, L., & White, S. D. M. 2007, *MNRAS*, **374**, 1303
- Dark Energy Survey Collaboration, Abbott, T., Abdalla, F. B., et al. 2016, *MNRAS*, **460**, 1270
- Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, *ApJ*, **292**, 371
- Davis, M., & Peebles, P. J. E. 1983, *ApJ*, **267**, 465
- de Jong, J. T. A., Verdoes Kleijn, G. A., Kuijken, K. H., & Valentijn, E. A. 2012, *ExA*, **35**, 25
- Delgado, A. M., Wadkar, D., Hadzhiyska, B., et al. 2022, *MNRAS*, **515**, 2733
- Dumerchat, T., & Bautista, J. 2024, *A&A*, **684**, 57
- Durkalec, A., Le Fèvre, O., Pollo, A., et al. 2015, *A&A*, **583**, A128
- Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, *AJ*, **142**, 72
- Eke, V. R., Cole, S., Frenk, C. S., & Navarro, J. F. 1996, *MNRAS*, **281**, 703
- Euclid Collaboration, Mellier, Y., Abdurro'uf, et al. 2025, *A&A*, **697**, A1
- Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, *PASP*, **125**, 306
- Gabrielpillai, A., Somerville, R. S., Genel, S., et al. 2022, *MNRAS*, **517**, 6091
- Gardner, J. P., Mather, J. C., Abbott, R., et al. 2023, *PASP*, **135**, 068001
- Geach, J. E., Sobral, D., Hickox, R. C., et al. 2012, *MNRAS*, **426**, 679
- Gebhardt, K., Mentuch Cooper, E., Ciardullo, R., et al. 2021, *ApJ*, **923**, 217
- Goodman, J., & Weare, J. 2010, *CAMCS*, **5**, 65
- GPy 2012, GPy: A Gaussian process framework in python, <http://github.com/SheffieldML/GPy>
- Guo, H., Zheng, Z., Zehavi, I., et al. 2015, *MNRAS*, **453**, 4368
- Hadzhiyska, B., Liu, S., Somerville, R. S., et al. 2021, *MNRAS*, **508**, 698
- Hamilton, A. J. S. 1992, *ApJL*, **385**, L5
- Hamilton, A. J. S. 1993, *ApJ*, **417**, 19
- Hamilton, A. J. S. 1998, *Linear Redshift Distortions: A Review* (Springer), **185**
- Hauser, M. G., & Peebles, P. J. E. 1973, *ApJ*, **185**, 757
- Hearin, A. P., & Watson, D. F. 2013, *MNRAS*, **435**, 1313
- Hogg, D. W. 2000, arXiv, arXiv:astro-ph/9905116
- Hogg, D. W., Eisenstein, D. J., Blanton, M. R., et al. 2005, *ApJ*, **624**, 54
- Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, *ApJ*, **873**, 111
- Jana, A., & Samushia, L. 2024, *A&C*, **49**, 100883
- Jarvis, M., Bernstein, G., & Jain, B. 2004, *MNRAS*, **352**, 338
- Kaiser, N. 1984, *ApJL*, **284**, L9
- Klypin, A., Yepes, G., Gottlöber, S., Prada, F., & Heß, S. 2016, *MNRAS*, **457**, 4340
- Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., et al. 2004, *ApJ*, **609**, 35
- Landy, S. D., & Szalay, A. S. 1993, *ApJ*, **412**, 64
- Lee, J., Shin, J., Snaith, O. N., et al. 2021, *ApJ*, **908**, 11
- Ma, C.-P., & Fry, J. N. 2000, *ApJ*, **543**, 503
- McKay, M. D., Beckman, R. J., & Conover, W. J. 1979, *Techn*, **21**, 239
- Mo, H. J., Jing, Y. P., & White, S. D. M. 1996, *MNRAS*, **282**, 1096
- Murray, S., Diemer, B., Chen, Z., et al. 2021, *A&C*, **36**, 100487
- Murray, S., Power, C., & Robotham, A. 2013, *A&C*, **3**, 23
- Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, *ApJ*, **462**, 563
- Nelson, D., Springel, V., Pillepich, A., et al. 2019, *ComAC*, **6**, 2
- Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, *ApJS*, **208**, 5
- Norberg, P., Baugh, C. M., Gaztañaga, E., & Croton, D. J. 2009, *MNRAS*, **396**, 19
- OpenUniverse Collaboration T. L. D. E. S. Team T. R. H. P. I., et al. 2025, *MNRAS*, **544**, 3799
- Ouchi, M., Ono, Y., & Shibuya, T. 2020, *ARA&A*, **58**, 617
- Peacock, J. A., & Smith, R. E. 2000, *MNRAS*, **318**, 1144
- Peebles, P. J. E. 1980, *The Large-scale Structure of the Universe* (Princeton Univ. Press)
- Rasmussen, C. E., & Williams, C. K. I. 2006, *Gaussian Processes for Machine Learning* (MIT Press)
- Scoccimarro, R., Sheth, R. K., Hui, L., & Jain, B. 2001, *ApJ*, **546**, 20
- Seljak, U. 2000, *MNRAS*, **318**, 203
- Sheth, R. K., & Tormen, G. 1999, *MNRAS*, **308**, 119
- Sobral, D., Best, P. N., Geach, J. E., et al. 2010, *MNRAS*, **404**, 1551
- Somerville, R. S., Gilmore, R. C., Primack, J. R., & Domínguez, A. 2012, *MNRAS*, **423**, 1992
- Somerville, R. S., Hopkins, P. F., Cox, T. J., Robertson, B. E., & Hernquist, L. 2008, *MNRAS*, **391**, 481
- Somerville, R. S., & Kolatt, T. S. 1999, *MNRAS*, **305**, 1
- Somerville, R. S., Popping, G., & Trager, S. C. 2015, *MNRAS*, **453**, 4337
- Somerville, R. S., & Primack, J. R. 1999, *MNRAS*, **310**, 1087
- Somerville, R. S., Olsen, C., Yung, L. Y. A., et al. 2021, *MNRAS*, **502**, 4858
- Springel, V., White, S. D. M., Tormen, G., & Kauffmann, G. 2001, *MNRAS*, **328**, 726
- Storey-Fisher, K., Tinker, J. L., Zhai, Z., et al. 2024, *ApJ*, **961**, 208
- Tegmark, M., Blanton, M. R., Strauss, M. A., et al. 2004, *ApJ*, **606**, 702
- Tinker, J. L., Robertson, B. E., Kravtsov, A. V., et al. 2010, *ApJ*, **724**, 878
- Vale, A., & Ostriker, J. P. 2004, *MNRAS*, **353**, 189
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *NatMe*, **17**, 261
- Wadkar, D., Villaescusa-Navarro, F., Ho, S., & Perreault-Levasseur, L. 2020, arXiv, arXiv:2012.00111
- Wang, Y., Zhai, Z., Alavi, A., et al. 2022, *ApJ*, **928**, 1
- White, M., Blanton, M., Bolton, A., et al. 2011, *ApJ*, **728**, 126
- White, M., Zheng, Z., Brown, M. J. I., Dey, A., & Jannuzi, B. T. 2007, *ApJ*, **655**, L69
- Yadav, J., Bharadwaj, S., Pandey, B., & Seshadri, T. R. 2005, *MNRAS*, **364**, 601
- York, D. G., Adelman, J., Anderson, J. E. J., et al. 2000, *AJ*, **120**, 1579
- Yung, L. Y. A., Somerville, R. S., Finkelstein, S. L., et al. 2021, *MNRAS*, **508**, 2706
- Yung, L. Y. A., Somerville, R. S., Finkelstein, S. L., Popping, G., & Davé, R. 2019, *MNRAS*, **483**, 2983
- Yung, L. Y. A., Somerville, R. S., Ferguson, H. C., et al. 2022, *MNRAS*, **515**, 5416
- Yung, L. Y. A., Somerville, R. S., Ferguson, H. C., et al. 2022, *MNRAS*, **515**, 5416
- Yung, L. Y. A., Somerville, R. S., Finkelstein, S. L., et al. 2023, *MNRAS*, **519**, 1578
- Zentner, A. R., Hearin, A. P., & van den Bosch, F. C. 2014, *MNRAS*, **443**, 3044
- Zhai, Z., Tinker, J. L., Becker, M. R., et al. 2019, *ApJ*, **874**, 95
- Zheng, Z., Berlind, A. A., Weinberg, D. H., et al. 2005, *ApJ*, **633**, 791
- Zheng, Z., Coil, A. L., & Zehavi, I. 2007, *ApJ*, **667**, 760
- Zheng, Z., & Weinberg, D. H. 2007, *ApJ*, **659**, 1