Draft version June 14, 2023
Typeset using LaTeX **twocolumn** style in AASTeX63

The DESI One-Percent Survey: Evidence for Assembly Bias from Low-Redshift Counts-in-Cylinders Measurements

Alan N. Pearl,[1,2] Andrew R. Zentner,[1,2] Kuan Wang,[3,4] Jeffrey A. Newman,[1] Rachel Bezanson,[1] John Moustakas,[5] and DESI Collaborators

[1] *Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA 15260*
[2] *Pittsburgh Particle Physics, Astrophysics, and Cosmology Center (PITT PACC), University of Pittsburgh, Pittsburgh, PA 15260*
[3] *Department of Physics, University of Michigan, Ann Arbor, MI 48109*
[4] *Leinweber Center for Theoretical Physics, University of Michigan, Ann Arbor, MI 48109*
[5] *Department of Physics and Astronomy, Siena College, 515 Loudon Road, Loudonville, NY 12211*

ABSTRACT

We explore the information on the galaxy-halo connection that is available in low redshift samples from the early data release of the Dark Energy Spectroscopic Instrument (DESI). We model the halo occupation distribution (HOD) from $z = 0.1 - 0.3$ using Survey Validation 3 (SV3; a.k.a., the One-Percent Survey) data of the DESI Bright Galaxy Survey (BGS). In addition to commonly used clustering metrics and number density, we incorporate counts-in-cylinders (CiC) measurements, which tighten HOD constraints drastically. Our analysis is made more efficient through the development of a new open-source Python package, `galtab`, which pretabulates populations of galaxies in halos, enabling the rapid prediction of CiC for any HOD model available in `halotools`. This methodology allows posterior probability distributions from Markov chains to converge much more quickly by reducing the required number of trial points by up to an order of magnitude, in addition to enabling even more drastic speedups due to its GPU portability. Our HOD fits constrain characteristic halo masses tightly and provide statistical evidence for assembly bias, especially at lower luminosity thresholds: the HOD of central galaxies in samples with limiting absolute magnitude $M_r < -20.0$ and $M_r < -20.5$ samples is positively correlated with halo concentration with a significance of 99.9% and 99.5%, respectively. Our models also strongly favor positive central assembly bias for the brighter $M_r < -21.0$ sample at $z \sim 0.25$, but show weaker evidence for assembly bias with the same luminosity threshold at $z \sim 0.15$, with 94.8% and 61.7% significance, respectively. We detect no evidence of assembly bias in the occupation of satellite galaxies. We provide the best fits and confidence intervals for each threshold sample's characteristic halo masses, assembly bias, and other HOD parameters. These constraints are expected to be significantly tightened with future DESI data, which will span an area $100\times$ larger than that of SV3.

1. INTRODUCTION

The large-scale distribution of galaxies in the universe is a powerful probe of cosmological models (e.g., Beutler et al. 2011; Anderson et al. 2012; Abbott et al. 2018). This is because galaxies trace the dark matter distribution, whose distribution is set by cosmological parameters and is well-characterized by modern simulations (e.g., Klypin et al. 2016; Ishiyama et al. 2021). However, for accurate cosmological inference, it is necessary to marginalize over the possible relationships between observational probes and the theoretical matter distribution. Therefore, leveraging large-scale structure to constrain cosmology requires flexible models of the galaxy-halo connection, and necessitates incorporating as much empirical information as possible to tightly constrain such flexible models.

Central and satellite galaxies are thought to form at the dense centers of halo and subhalo potential wells, respectively. Therefore, the spatial clustering of most galaxy samples can be described well by a halo occupation distribution (HOD; e.g., Berlind & Weinberg 2002; Zheng et al. 2007), which probabilistically connects the average number of central and satellite galaxies a dark matter halo hosts to its mass. This formalism can be extended through additional parameters that lead to correlations between galaxy abundance and secondary halo properties (i.e., assembly bias Hearin et al. 2016), which can improve fit quality. As the data improve, further extensions to HOD models may be warranted, e.g., by relaxing the assumption of a log-normal stellar-to-halo-mass relation or of a spatially isotropic Navarro-Frenk-White (NFW) distribution of satellite galaxies.

The most common observables used to constrain the galaxy-halo connection via spectroscopic galaxy samples are the number density and the projected two-point correlation function $w_{\rm p}(r_{\rm p})$ (e.g., Zehavi et al. 2005; Reddick et al. 2013). However, Wang et al. (2019) has shown that the counts-in-cylinders (CiC) distribution $P(N_{\rm CiC})$ offers significant complementary information on the parameters of interest – particularly those that control satellite occupation and assembly bias. As demonstrated by Storey-Fisher et al. (2022), it is also possible to quantify clustering information beyond the two-point function using the underdensity probability function and the density-marked correlation function. These studies highlight that even with existing datasets, incorporating different measurements of the large-scale structure can help optimize model fitting.

In this paper, we extend previous analyses by incorporating a novel spectroscopic dataset; implementing a new, more efficient CiC prediction framework; and demonstrating the gain these provide. We leverage data from the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al. 2022), which will ultimately obtain spectroscopic redshifts of 40 million galaxies in an effort to precisely map the large-scale structure of a large volume of the observable universe. While the full dataset is still being collected, this work utilizes redshift measurements for more than 40,000 galaxies obtained by the Survey Validation 3 (SV3) component of the DESI early data release (DESI Collaboration et al. 2023).

We approximately adopt the best-fit flat-universe cosmology from Planck Collaboration et al. (2020). The relevant cosmological parameters that we use are as follows: $h = 0.6777$, $\Omega_{m,0} = 0.30712$, $\Omega_{b,0} = 0.048252$, and $T_{\rm CMB} = 2.7255$ K. However, we scale all distance and distance-dependent values to units equivalent to setting the Hubble parameter to $h = 1$ (e.g., h^{-1}Mpc).

This paper is organized as follows. We describe the DESI and simulation data in Section 2. We outline the summary statistics used in our analysis in Section 3. We detail our methodology for measuring and predicting CiC, through the `galtab` package, in Section 4. We present our resulting constraints on the HOD in Section 5, and discuss our conclusions in Section 6.

2. DATA

2.1. *DESI BGS*

The DESI Bright Galaxy Survey (BGS) is a highly complete magnitude-limited spectroscopic survey of $z < 0.5$ galaxies, which aims to target galaxies over at least 14,000 square degrees down to a limit roughly two magnitudes fainter than the Sloan Digital Sky Survey (SDSS; Abazajian et al. 2009). Our analyses only use the BGS Bright sample, which is complete down to an apparent r-band magnitude of $m_r < 19.5$. Because the DESI survey is still in progress at the time of this writing, we analyze only data from the Survey Validation 3 (SV3; DESI Collaboration et al. 2023) dataset (also known as the One-Percent Survey as it contains approximately 1% of the anticipated volume of DESI). These data were obtained in over twenty sky regions totaling an area of 173.3 sq deg, as shown in Figure 1. A significantly higher fraction of potential targets was observed in the SV3 fields than will be the case for typical DESI survey data due to the use of a denser tiling strategy, simplifying the corrections needed for our analysis.

We specifically use the SV3 Large Scale Structure (LSS) catalogs, which only include sources with secure spectroscopic redshift measurements, as described in DESI Collaboration et al. (2023). These catalogs are well suited for clustering measurements since they are paired with 18 random realization files, each containing 2500 objects per deg^2 of sky coverage, and weights from 128 fiber assignment realizations. We also utilize r-band absolute magnitude measurements from `fastspecfit` (Moustakas et al. in prep.[1]), which are computed for an SDSS r-band response curve K-corrected to the $z = 0.1$ reference frame. Note that all references to absolute magnitudes in this paper, M_r, are scaled to $h = 1$ units; therefore, they are equivalent to $M_r - 5\log h$ for all other values of the Hubble parameter.

We break this data into three volume-limited samples which each cover the redshift range $0.1 < z < 0.2$, constructed with absolute r-band absolute magnitude limits of $M_r < -20.0$, -20.5, and -21.0. We also define a fourth sample covering a slightly higher redshift range of $0.2 < z < 0.3$ with limit $M_r < -21.0$. We plot each sample cut in Figure 2 and summarize these samples in Table 1. Unless otherwise specified, all observational measurements in this paper are measured from one of these samples.

Table 1. DESI subsamples used for our analyses. The full sample size is given by $N_{\rm tot}$, while $N_{\rm cyl}$ is the number of centers of the cylinders that meet our spatial completeness criteria.

M_r threshold	Redshift range	$N_{\rm tot}$	$N_{\rm cyl}$
-20.0	$0.1 < z < 0.2$	20,241	15,936

Table 1 *continued*

[1] https://fastspecfit.readthedocs.io/

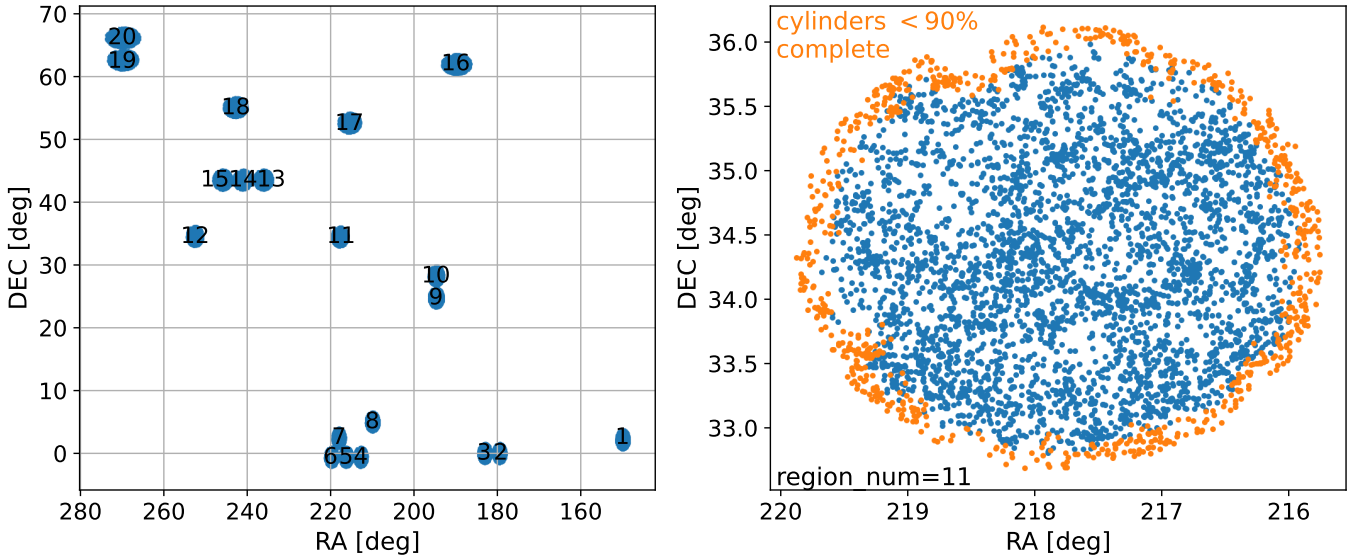

Figure 1. Footprint of the DESI Survey Validation 3 (SV3). The left panel displays the entire survey, broken up into twenty regions that are for the most part spatially isolated from each other. The right panel presents a close-up of the region labeled by the number 11 in the left panel. The points shown in orange, which are located primarily near the edge of the region, indicate objects excluded as cylinder centers in our CiC measurement, as described in Section 4.2

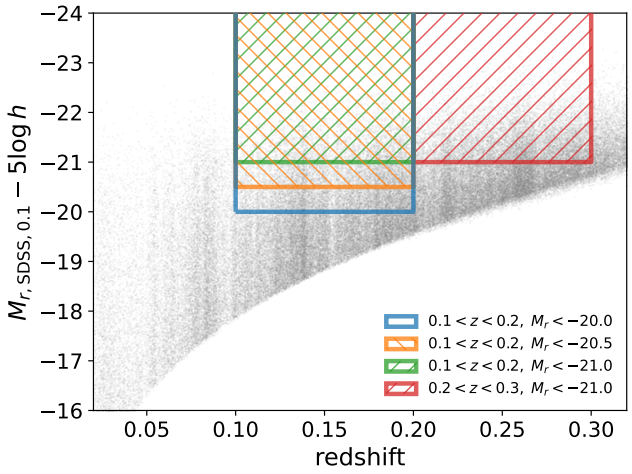

Figure 2. Distribution of r-band absolute magnitude M_r vs. redshift. The full DESI BGS SV3 sample is shown by the grey points. Our four volume-limited, absolute magnitude-thresholded samples are constructed through the cuts represented by the corresponding colored boundaries.

Table 1 *(continued)*

M_r threshold	Redshift range	N_{tot}	N_{cyl}
-20.5	$0.1 < z < 0.2$	11,036	8,686
-21.0	$0.1 < z < 0.2$	5,096	4,031
-21.0	$0.2 < z < 0.3$	14,874	12,543

2.2. *Small MultiDark Planck*

To study the galaxy-halo connection, we must compare DESI galaxy clustering data to an assumed distribution of underlying dark matter halos. For this halo distribution prior, we adopt the Small MultiDark Planck simulation (SMDPL; Klypin et al. 2016), which uses the same Planck cosmology that we assume in this work. This simulation was performed with 3840^3 particles, but our analysis is based only upon the halo catalogs produced by applying the Rockstar halo finder (Behroozi et al. 2013). We adopt the virial mass from Rockstar as our halo mass, M_h.

SMDPL covers a $400h^{-1}$ Mpc periodic cube, which is over ten times the volume of our SV3 samples. This is sufficiently large so that cosmic-variance-like uncertainties from the data dominate over the sample variance of this simulation volume. However, future studies will need to use larger volume simulations to compensate for DESI's volume, which will be 100 times that of SV3.

3. OBSERVABLE SUMMARY STATISTICS

To extract clustering information from the galaxy samples, we use three summary statistics: number density n_{gal}, the projected two-point correlation function $w_{\mathrm{p}}(r_{\mathrm{p}})$, and the CiC distribution $P(N_{\mathrm{CiC}})$. We compare the observations[2] with our best-fit models for these three summary statistics for each sample in Figure 9.

[2] i.e., Bezanson Points

The number density is calculated via the sum of the inverse individual probability (IIP; see Section 4.2) weights of the galaxies in the sample divided by the comoving volume they were sampled from. For the HOD number density predictions, the comoving volume of SMDPL is $400^3 h^{-3}\mathrm{Mpc}^3$, while the volumes of the DESI samples depend on the redshift cuts and the survey area. The DESI SV3 BGS survey area is 173.3 sq deg, which corresponds to comoving volumes of $2.83 \times 10^6 h^{-3}\mathrm{Mpc}^3$ and $6.95 \times 10^6 h^{-3}\mathrm{Mpc}^3$ for samples with redshift ranges of $0.1 < z < 0.2$ and $0.2 < z < 0.3$, respectively.

The projected two-point correlation function is a common way to quantify data clustering at various physical scales. By integrating over the line-of-sight dimension, this statistic decreases the dependence of the inferred clustering on velocity-space distortions. It is defined by

$$w_{\mathrm{p}}(r_{\mathrm{p}}) = 2 \int_0^{\pi_{\mathrm{max}}} \xi(r_{\mathrm{p}}, \pi) d\pi \tag{1}$$

where ξ is the two-point correlation function, π is line-of-sight separation distance, and r_{p} is perpendicular separation distance. For consistency with Wang et al. (2022), we choose $\pi_{\mathrm{max}} = 40 h^{-1}\mathrm{Mpc}$ and use twelve logarithmically spaced bins between r_{p} of $0.158 h^{-1}\mathrm{Mpc}$ and $39.81 h^{-1}\mathrm{Mpc}$. We concatenate all 18 random files from the SV3 LSS catalogs but draw a random 20% subsample to reduce excessive computational time. We utilize the `pycorr`[3] package to apply the Landy & Szalay (1993) estimator, line-of-sight integration, and fiber assignment weights. The performance-critical pair searching is powered by `Corrfunc` (Sinha & Garrison 2020).

Counts-in-cylinders (CiC) is a type of counts-in-cells statistic (i.e., it quantifies the local density of neighbors in a cell around each object; the development of such metrics has a long history; e.g., Hubble 1936; Zwicky 1957; White 1979; Adelberger et al. 1998) that defines neighbors using a cylindrical cell along the line-of-sight direction. As in Wang et al. (2022), we use relatively small-scale cylinders by choosing the radius to be $R_{\mathrm{CiC}} = 2 h^{-1}\mathrm{Mpc}$ and the half-length to be $L_{\mathrm{CiC}} = 10 h^{-1}\mathrm{Mpc}$. Cylinders of this scale primarily probe the number of intra-halo galaxies and are therefore sensitive to satellite occupation. Conveniently, using a small cylindrical volume is also a computationally favorable choice. The CiC distribution $P(N_{\mathrm{CiC}})$ can be evaluated in bins of N_{CiC} – for which we use ten linearly spaced bins between -0.5 and 9.5 plus twenty logarithmically spaced bins between 9.5 and 149.5; alternatively, the majority of available information can be captured by computing the first three to five moments of the N_{CiC} distribution. We describe our methods used to compute counts-in-cylinders in detail in Section 4.

We test the ability of each summary statistic to provide information about the HOD by sampling uniformly from HOD parameters around their 1σ confidence interval from the Wang et al. (2022) $M_r < -20.5$ sample. We predict each of our summary statistics plus noise according to a random draw from the covariance matrix calculated in Section 3.1, including CiC up to the tenth moment. We then train a random forest (Breiman 2001) to predict the HOD parameters from these summary statistics and provide a visualization of the resulting SHapley Additive exPlanations (SHAP; Lundberg & Lee 2017) feature importance in Figure 3. To briefly summarize, number density is highly important for predicting $\log M_{\mathrm{min}}$, the two-point correlation function is broadly informative across all parameters, and the first few CiC moments are particularly important for constraining satellite HOD parameters.

3.1. *Covariance of Summary Statistics*

To constrain our HOD model, we compare the following summary statistics as measured in our data to model predictions: number density; the two-point correlation function (computed in 12 bins in r_{p}); and CiC (for 28 bins in N_{CiC}). We calculate the covariance matrix of these summary statistics by jackknife resampling using the 20 regions displayed in Figure 1.

To do this, we perform a measurement of every summary statistic simultaneously on the subset of data that includes all but one jackknife region. We repeat this process for each combination of 19 jackknife regions to obtain $N_{\mathrm{J}} = 20$ jackknife realizations. The covariance matrix of our summary statistics can then be estimated by

$$\Sigma_{ij} = \frac{N_{\mathrm{J}} - 1}{N_{\mathrm{J}}} \sum_{k=0}^{N_{\mathrm{J}}} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j) \tag{2}$$

where $\bar{x}_i$ is the ith summary statistic measured in the entire dataset, and x_{ik} is the ith summary statistic measured in the kth jackknife realization.

4. COUNTS-IN-CYLINDERS

Counts-in-cylinders (CiC; derived in Peebles 1980 and previously used by Reid & Spergel 2009; Wang et al. 2022) is sensitive to higher-order n-point functions, which makes it complementary to two-point statistics commonly used in the literature. Despite its utility, CiC is not widely adopted in galaxy-halo connection studies, due to difficulties in correcting for systematics, excessive

[3] https://github.com/cosmodesi/pycorr

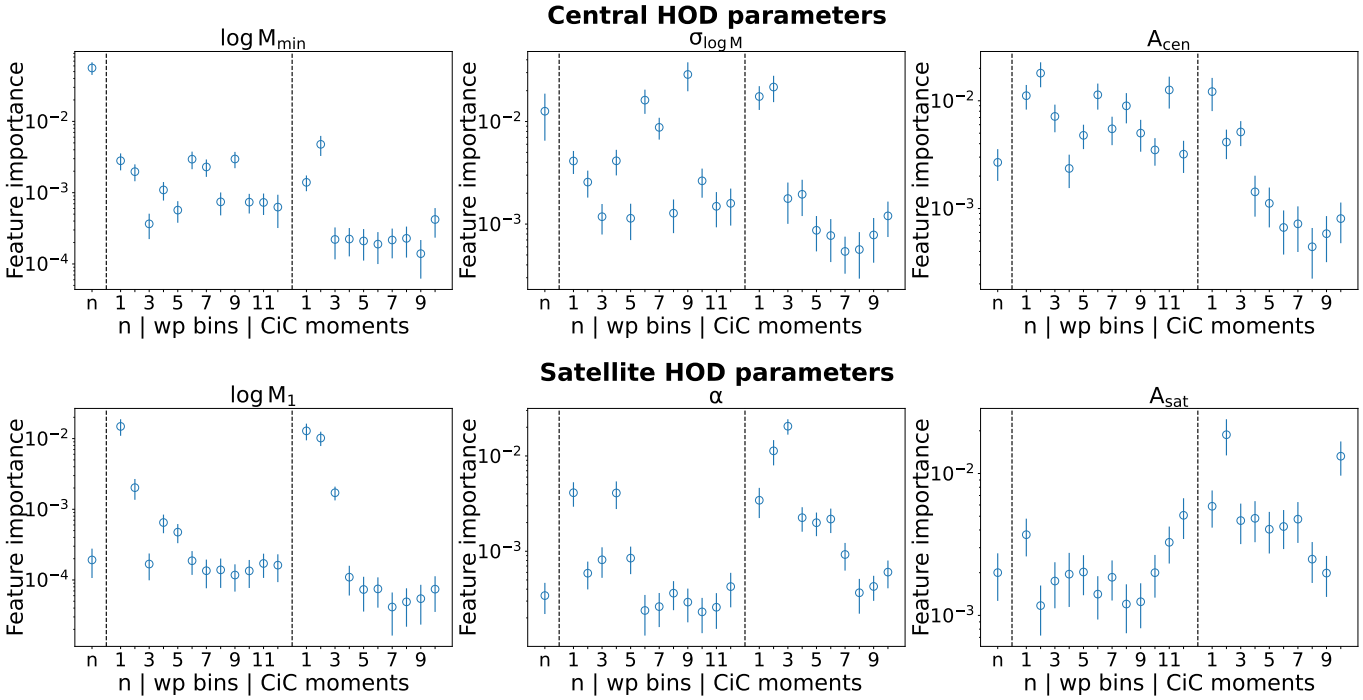

Figure 3. SHAP feature importances for each of our summary statistics for inferring HOD parameters. Each panel plots the importance of each feature (i.e., each quantity that is used to predict the HOD parameters via a machine learning model), calculated by the mean absolute SHAP value for the given HOD parameter. Summary statistics with high feature importance are more useful for predicting the parameter. For the satellite HOD parameters (bottom row), the first few CiC moments provide the majority of the constraining information. See Figure 11 for beeswarm plots of the full distribution of SHAP values of the six most important features for each parameter.

computational time, and significantly increased dimensionality of the full covariance matrix.

In this section, we present our methodology to mitigate all of these problems. We have implemented each of these methods in an open source Python package `galtab`[4]. After a brief explanation of our observational cylinder geometry in Section 4.1, we present our weighting method in Section 4.2 based on individual inverse probabilities and inverse conditional probabilities (IIP×ICP), which corrects CiC calculations to account for clustering bias in surveys with fiber collisions. This approach is analogous to and inspired by pair inverse probabilities (PIP) weighting (Bianchi & Percival 2017), which we used to correct our $w_{\rm p}(r_{\rm p})$ measurement.

To minimize the increase in dimensionality, we suggest using only the first three to five moments of the CiC distribution, defined in Section 4.3, which retain most of the constraining information. Our analysis uses information from the entire CiC distribution, but our results are not significantly affected by using only the first five CiC moments instead.

[4] https://github.com/AlanPearl/galtab

Additionally, we present a galaxy placeholder pretabulation method in Section 4.4 to speed up our Markov-chain Monte Carlo (MCMC) procedure. This makes our CiC prediction runtime comparable to traditional Monte Carlo $w_{\rm p}(r_{\rm p})$ prediction methods but with the significant advantage of producing precise, deterministic values, which yield much higher MCMC sampling efficiency than stochastic Monte Carlo predictions.

4.1. *Observational Cylinder Geometry*

While a cylinder perfectly aligns with the velocity distortion in an idealized simulation, for observations, we must slightly distort its round face into a truncated cone so that it is always perpendicular to the line-of-sight direction. We also allow a slight curve to this truncated cone's top and bottom faces, to keep them normal to the line of sight. Then there are only two search criteria: angular distance and line-of-sight separation. The line-of-sight separation cut is $L_{\rm CiC}$ and we define the angular radius cut to be

$$\theta_{\rm CiC} = \arccos\left(1 - \frac{3R_{\rm CiC}^2 L_{\rm CiC}}{(d + L_{\rm CiC})^3 - (d - L_{\rm CiC})^3}\right) \quad (3)$$

where d is the comoving distance to the center of our "cylinder". This ensures that its volume is still precisely $2\pi R_{\rm CiC}^2 L_{\rm CiC}$, and $\theta_{\rm CiC} \approx R_{\rm CiC}/d$ as $d \to \infty$.

4.2. *IIP×ICP Weighting*

In order to account for fiber collisions, the DESI Large-Scale Structure catalogs come with "bitweights" columns. These bitweights represent 128 true or false values for each object that correspond to 128 fiber assignment realizations. Therefore, the probability that an object in the catalog would have been assigned a fiber can be obtained by a summation of these 128 bits plus one divided by 129 (since the object was observed, there is an understood true for the 129th realization). We explicitly calculate the probability of assigning a fiber to the ith galaxy using

$$P(i) = \frac{\texttt{sum(bitweights[i])} + 1}{129}, \tag{4}$$

while the probability of simultaneously assigning fibers to both the ith and jth galaxies is

$$P(i\,\&\,j) = \frac{\texttt{sum(bitweights[i] \& bitweights[j])} + 1}{129} \tag{5}$$

where `sum` and `&` are bitwise operations. Thanks to the high fiber completeness of SV3, the average value of $P(i)$ is 0.984.

In order to measure the CiC distribution, we must calculate the expectation value of the number of galaxies we expect to find in the cylinder around every galaxy individually, $N_{{\rm CiC},i}$. For this task, we sum the inverse conditional probabilities (ICPs) of each neighboring galaxy's fiber assignment (conditional on the fiber assignment of the cylinder's central galaxy). Using the definitions from Equations 4 and 5,

$$\mathrm{ICP}_{j|i} = \frac{P(i)}{P(i\,\&\,j)} \tag{6}$$

$$N_{{\rm CiC},i} = \frac{1}{f_{\rm rand}} \sum_{j \in C_i} \mathrm{ICP}_{j|i} \tag{7}$$

where C_i is the set of indices of galaxies contained by the cylinder surrounding the ith galaxy, and $f_{\rm rand}$ is the fraction of randoms enclosed in the cylinder compared to the expected number occupying a circle of angular radius $\theta_{\rm CiC}$ in order to account for incompleteness in spatial coverage. Note that we do not include cylinders with $f_{\rm rand} < 0.9$, and amongst the cylinders we keep, the average value of $f_{\rm rand}$ is 0.959.

We measure $P(N_{\rm CiC})$ from the sample distribution of $N_{{\rm CiC},i}$ values, but we need to overweight the objects in dense regions of the sky that have been undersampled, so therefore, we weight objects by their inverse individual probability (IIP). The IIP of the ith galaxy is simply

$$\mathrm{IIP}_i = \frac{1}{P(i)}. \tag{8}$$

Finally, for our binned histogram measurements of $P(N_{\rm CiC})$, we split each IIP_i into two parts, IIP_{i+} and IIP_{i-}. These weights are applied to the integers above and below $N_{{\rm CiC},i}$, respectively, and are proportional to one minus that integer's distance from $N_{{\rm CiC},i}$ so that

$$\frac{\mathrm{IIP}_{i+}\lceil N_{{\rm CiC},i}\rceil + \mathrm{IIP}_{i-}\lfloor N_{{\rm CiC},i}\rfloor}{\mathrm{IIP}_i} = N_{{\rm CiC},i} \tag{9}$$

4.3. *Calculating the CiC Moments*

In order to decrease the dimensionality of the covariance matrix, one may choose to condense the information contained in the CiC distribution into its first few moments, which we define as

$$\mu_1 = \sum_{i=1}^{N} w_i N_{{\rm CiC},i} \tag{10}$$

$$\mu_2 = \sqrt{\sum_{i=1}^{N} w_i (N_{{\rm CiC},i} - \mu_1)^2} \tag{11}$$

$$\mu_{k>2} = \frac{1}{\mu_2^k} \sum_{i=1}^{N} w_i (N_{{\rm CiC},i} - \mu_1)^k. \tag{12}$$

where $N_{{\rm CiC},i}$ is the number of neighbors inside the cylinder surrounding the $i^{\rm th}$ galaxy in the sample and w_i is the $i^{\rm th}$ IIP weight, but normalized to $\sum w_i = 1$ (see Section 4.2 for details on $N_{{\rm CiC},i}$ and IIP weights). Note that μ_1 is the mean, μ_2 is the standard deviation, and for $k > 2$, μ_k are standardized central moments (skewness, kurtosis, etc.), uncorrected for degree-of-freedom bias, which is a negligible source of systematics for large sample sizes compared to other uncertainties. In figures, we refer to μ_k as CiC_k to be explicit that they are moments of CiC.

4.4. *Pretabulation with Placeholder Galaxies*

Predictions of CiC from Monte Carlo HOD realizations are notoriously slow and noisy. This stochasticity reduces the sampling efficiency of Monte Carlo explorations of model parameter space by decreasing the acceptance rate which, in turn, increases the autocorrelation length of MCMC chains and necessitates longer chains and run times. To remedy this, we have developed a method to calculate precise, deterministic CiC predictions by pretabulating placeholder galaxies inside simulated halos.

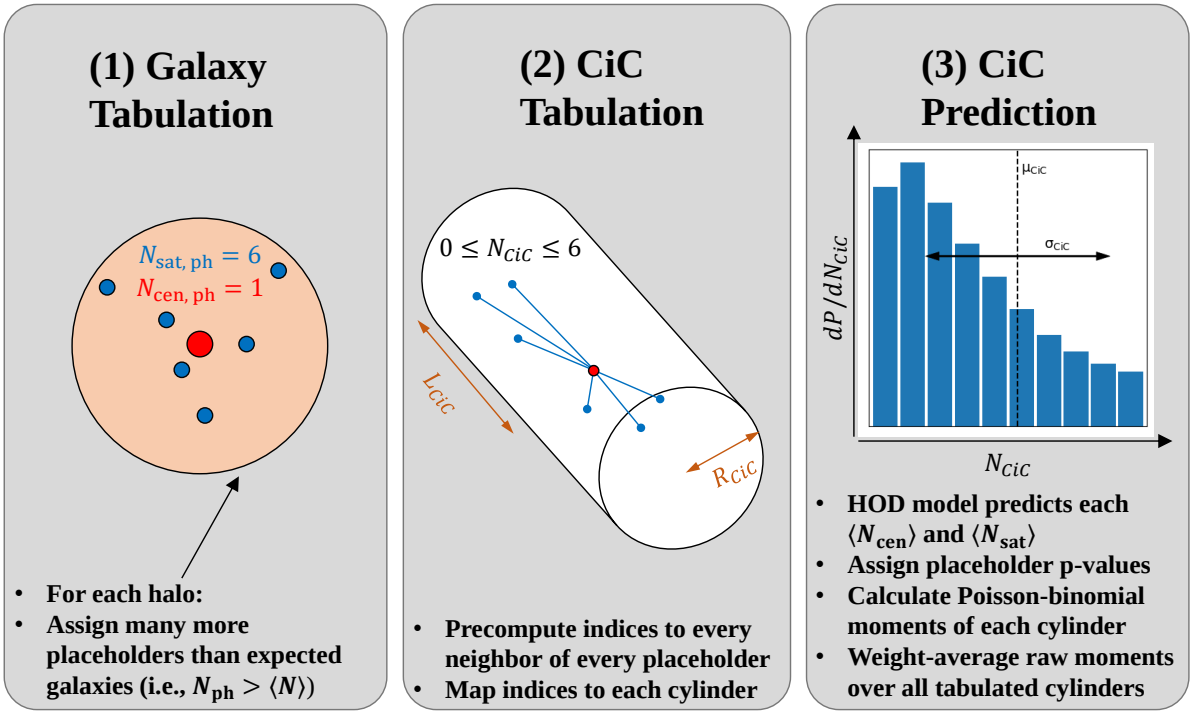

Figure 4. Demonstration of our placeholder algorithm used to pretabulate counts-in-cylinders pair indices. Given a fiducial model, we populate placeholder centrals for most halos with a non-zero probability of hosting a halo. We populate many more placeholder satellites than expected in the fiducial model so that the resulting binomial satellite occupation distribution sufficiently resembles the assumed Poisson distribution. We then tabulate the placeholder indices in each halo for rapid CiC prediction using one of the two modes described in Sections 4.5 and 4.6.

Our procedure is illustrated in Figure 4. Our method requires a fiducial HOD model to compute the expected occupation, $\langle N_{\rm cen} \rangle$ and $\langle N_{\rm sat} \rangle$, for each halo. For our fiducial model, we choose the best fit of Wang et al. (2022) that corresponds to the magnitude threshold of each of our samples. We populate each halo with $N_{\rm cen,ph}$ central placeholders and $N_{\rm sat,ph}$ satellite placeholders. We determine the number of satellite placeholders for each halo with the hyperparameter $W_{\rm max}$ according to the equation

$$N_{\rm sat,ph} = \left\lceil \frac{\langle N_{\rm sat} \rangle}{W_{\rm max}} \right\rceil \tag{13}$$

which ensures that, for fiducial model predictions, there are enough satellite placeholders that their individual weights are less than or equal to $W_{\rm max}$.

For centrals, we define a hyperparameter $Q_{\rm min}$ that sets the minimum quantile of central galaxies for which to populate a central placeholder. In practice, we set $N_{\rm cen,ph} = 1$ for all halos with $\langle N_{\rm cen} \rangle \geq \langle N_{\rm cen} \rangle_{\rm min}$, and $N_{\rm cen,ph} = 0$ otherwise. To solve for $\langle N_{\rm cen} \rangle_{\rm min}$, we numerically integrate and invert

$$Q_{\rm min} = \frac{\int_{\langle N_{\rm cen} \rangle_{\rm min}}^{1} \Phi(\langle N_{\rm cen} \rangle) \langle N_{\rm cen} \rangle d\langle N_{\rm cen} \rangle}{\int_{0}^{1} \Phi(\langle N_{\rm cen} \rangle) \langle N_{\rm cen} \rangle d\langle N_{\rm cen} \rangle} \tag{14}$$

where $\Phi(\langle N_{\rm cen} \rangle) d\langle N_{\rm cen} \rangle$ is the number density of halos with expected central occupation between $\langle N_{\rm cen} \rangle$ and $\langle N_{\rm cen} \rangle + d\langle N_{\rm cen} \rangle$.

To balance accuracy and runtime (see Figure 5), we set $W_{\rm max} = 0.05$ and $Q_{\rm min} = 10^{-4}$. In `galtab`, these hyperparameters can be tuned via the `max_weight` and `min_quant` keyword arguments, respectively.

We may choose any parameters for our HOD model and obtain a new prediction of $\langle N_{\rm X} \rangle$ for each halo and for each galaxy type denoted by X: central or satellite. Each placeholder galaxy is then assigned a weight, or probability, value $P_i = \langle N_{\rm X} \rangle / N_{\rm X,ph}$.

As is usually done in Monte Carlo HOD realizations, these galaxy probability values are assumed to be independent. Therefore, the halo occupation of centrals follows a Bernoulli distribution, the same as typical Monte Carlo frameworks. However, the halo occupation of satellites follows a binomial distribution in our framework, which only converges to the desired Poisson dis-

tribution in the low $P_i \lesssim 0.05$ limit, hence our choice of $W_{\rm max} = 0.05$.

Finally, a single counts-in-cylinder search is required (we use the `halotools` implementation for this) to obtain a list of the indices of possible neighbors for each placeholder. This allows us to rapidly calculate our CiC metric, as described in the following sections.

4.5. *Pretabulated CiC Prediction: Monte Carlo Mode*

In order to calculate the CiC distribution $P(N_{\rm CiC})$ from the probability values of our pretabulated galaxies, we must consider the probability of each possible value of $N_{{\rm CiC},i}$ for each cylinder i, from which the full CiC distribution is the weighted superposition of each $N_{{\rm CiC},i}$ distribution. We write this as

$$P(N_{\rm CiC}) = \frac{\sum_{i=1}^{N} P_i P(N_{{\rm CiC},i})}{\sum_{i=1}^{N} P_i}. \tag{15}$$

In general, each $P(N_{{\rm CiC},i})$ is a Poisson binomial distribution, whose calculation scales exponentially with the number of neighbors in the ith cylinder, which is infeasible. Therefore, the full distribution can only be calculated using our Monte Carlo mode prediction. In this mode, we also pretabulate $n_{\rm MC}$ random seeds over $[0, 1)$ for each galaxy, which we use as Bernoulli quantiles after assigning the P_i of each placeholder. This allows us to effectively create $n_{\rm MC}$ independent realizations that can still produce quasi-deterministic and almost continuous (but non-differentiable) predictions. We find that using $n_{\rm MC} = 10$ random seeds produces reasonably stable results without excessive runtime. We will show in Section 4.6 that predictions of the CiC moments can be made without invoking random seeds, allowing them to be perfectly continuous and differentiable.

4.6. *Pretabulated CiC Prediction: Analytic Mode*

Although the full $P(N_{\rm CiC})$ distribution cannot be calculated analytically from our galaxy placeholders, the moments of this distribution can. As a simple example, the mean of this distribution is simply the weighted average of the individual means

$$\langle N_{\rm CiC} \rangle = \frac{\sum_{i=1}^{N} P_i \langle N_{{\rm CiC},i} \rangle}{\sum_{i=1}^{N} P_i} \tag{16}$$

where

$$\langle N_{{\rm CiC},i} \rangle = \sum_{j \in C_i} P_j \tag{17}$$

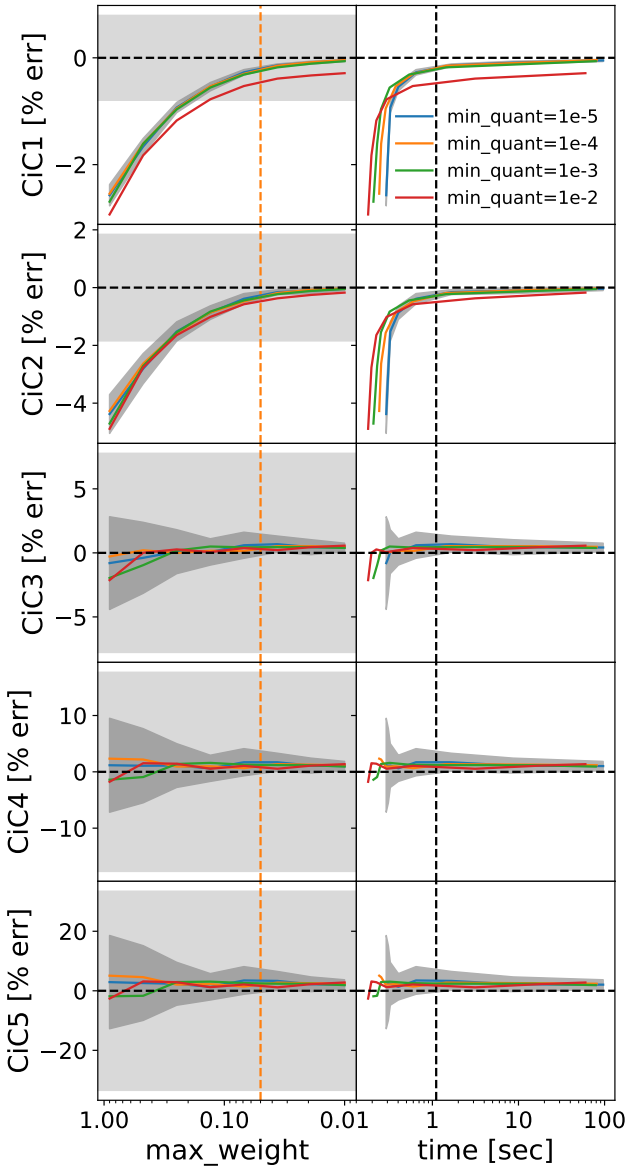

Figure 5. Hyperparameter tuning of `galtab` to achieve sufficient accuracy of CiC moments. The left panels show the tuning of the $W_{\rm max}$ parameter, which is translated to a CPU runtime in the panels on the right side of the figure, with lower values of $W_{\rm max}$ requiring longer times, but achieving higher accuracy. Line colors correspond to the denoted value of $Q_{\rm min}$, the dark grey bands correspond to a standard deviation due to tabulation stochasticity, horizontal dashed lines correspond to truth values from `halotools`, and the light grey band corresponds to a `halotools` standard deviation. The vertical dashed line in the left panels corresponds to our chosen value of $W_{\rm max} = 0.05$, which intentionally yields a similar runtime as `halotools`: approximately one CPU-second, as specified by the vertical dashed line in the right panels.

and C_i is the set of indices of galaxies contained by the cylinder surrounding the ith galaxy.

It is possible to calculate a similar relation for the standard deviation and the higher standardized moments we have defined in Equations 11 and 12. However, these relations are much more complicated. Note that the mean is a special case because it is the first raw moment (which allows Equation 16) as well as the first cumulant (which allows Equation 17).

Cumulants are a type of moment that have a special property that they are additive for random variables which are the sum of other random variables. For example, the number of neighbors in the ith cylinder is a random variable, which is the sum of the occupation of each of its pretabulated placeholder companions, which themselves are random variables:

$$N_{\mathrm{CiC},i} = \sum_{j \in C_i} X_j \tag{18}$$

where X_j is the occupation of the jth placeholder, which follows a Bernoulli distribution (0 or 1) with mean P_j. Therefore, the first cumulant of this Bernoulli distribution is $\kappa_1(X_j) = P_j$, and the subsequent Bernoulli cumulants can be derived from the recursion relation

$$\kappa_{k+1}(X_j) = P_j(1 - P_j)\frac{d\kappa_k(X_j)}{dP_j}. \tag{19}$$

Given the first $k_{\max}$ Bernoulli cumulants of each placeholder, we can calculate the first $k_{\max}$ Poisson binomial cumulants of the ith cylinder. We can take the kth cumulant of each random variable on both sides of Equation 18:

$$\kappa_k(N_{\mathrm{CiC},i}) = \sum_{j \in C_i} \kappa_k(X_j). \tag{20}$$

From the moments of each $N_{\mathrm{CiC},i}$, we would like the moments of the combined CiC distribution, which is a weighted superposition of each individual cylinder's distribution, as expressed in Equation 15. For this step, the most convenient set of moments to use are raw moments. The kth raw moment of $N_{\mathrm{CiC},i}$ can be obtained from its first k cumulants according to

$$\langle N^k_{\mathrm{CiC},i} \rangle = \kappa_k(N_{\mathrm{CiC},i}) + \sum_{j=1}^{k-1} \kappa_j(N_{\mathrm{CiC},i}) \langle N^{k-j}_{\mathrm{CiC},i} \rangle. \tag{21}$$

From these individual kth raw moments, we can calculate the kth raw moment of their superposition using a simple weighted average:

$$\langle N^k_{\mathrm{CiC}} \rangle = \frac{\sum_{i=1}^{N} P_i \langle N^k_{\mathrm{CiC},i} \rangle}{\sum_{i=1}^{N} P_i}. \tag{22}$$

The first raw moment is μ_1, but the remaining μ_k for $2 \leq k \leq k_{\max}$ depend on central moments. Therefore, the final nontrivial step of our analytic prediction framework is to calculate the central moments using the following binomial expansion:

$$\langle (N_{\mathrm{CiC}} - \langle N_{\mathrm{CiC}} \rangle)^k \rangle = \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j} \langle N^j_{\mathrm{CiC}} \rangle \langle N_{\mathrm{CiC}} \rangle^{k-j} \tag{23}$$

from which we can calculate the standard moments given in Equations 10 through 12 using

$$\mu_1 = \langle N_{\mathrm{CiC}} \rangle, \tag{24}$$

$$\mu_2 = \sqrt{\langle (N_{\mathrm{CiC}} - \langle N_{\mathrm{CiC}} \rangle)^2 \rangle}, \tag{25}$$

and

$$\mu_{k>2} = \frac{1}{\mu_2^k} \langle (N_{\mathrm{CiC}} - \langle N_{\mathrm{CiC}} \rangle)^k \rangle. \tag{26}$$

4.7. *Computational Performance*

In Section 4.4 and Figure 5, we have described our hyperparameter tuning of $W_{\max}$ and $Q_{\min}$ to balance runtime and accuracy. These parameters control the number of placeholders, N, as well as the average number of placeholders per cylinder, C. To store all pretabulated indices, the memory usage of `galtab` scales with $\mathcal{O}(NC)$.

There are also additional runtime considerations specific to each prediction mode. For the Monte Carlo mode, the runtime scales with the number of effective Monte Carlo realizations, n_{MC}, so the time complexity is $\mathcal{O}(n_{\mathrm{MC}}NC)$. For the analytic mode, the runtime scales with the highest calculated moment, $k_{\max}$, so the time complexity is $\mathcal{O}(k_{\max}NC)$.

By far, the most computationally expensive step of our procedure is the summation of occupations (or cumulants, for the analytic mode; see Equation 20) of placeholders per cylinder. To fully optimize this calculation, we employ just-in-time (JIT) compilation via the JAX library (Bradbury et al. 2018). This also automatically ports the computation to the GPU, if available, which can speed up the predictions by at least an order of magnitude faster than the times reported in Figure 5.

Table 2. Maximum-likelihood HOD parameters for each sample. For each set of best-fit parameters, the goodness of fit is given by the Akaike Information Criterion (AIC), the chi-squared (χ^2), the degrees of freedom (DoF), the p value corresponding to the probability of measuring $\geq \chi^2$ by chance, and the corresponding z score measure of tension. The fits without CiC, and without assembly bias are included for comparison.

Threshold	$\log M_{\rm min}$	$\sigma_{\log M}$	α	$\log M_1$	$\log M_0$	$A_{\rm cen}$	$A_{\rm sat}$	AIC	χ^2	DoF	p value	Tension
-20.0	12.227	0.990	0.681	12.739	12.339	0.966	-0.156	-292.68	12.15	19	0.879	0.15σ
(no CiC)	12.114	0.884	0.858	12.946	12.430	0.540	-0.795	49.66	10.20	13	0.678	0.42σ
(no $A_{\rm bias}$)	11.968	0.481	0.778	12.763	12.459			-284.95	23.88	19	0.201	1.28σ
-20.5	12.285	0.527	0.765	13.140	12.657	0.911	-0.223	-214.70	20.51	19	0.364	0.91σ
(no CiC)	12.923	1.387	0.566	12.935	12.930	0.164	-0.317	52.74	7.70	13	0.863	0.17σ
(no $A_{\rm bias}$)	12.244	0.381	0.661	13.020	12.912			-208.36	30.85	19	0.042	2.03σ
-21.0	12.467	0.211	0.475	13.323	13.068	0.853	0.050	-233.42	54.76	42	0.090	1.70σ
(no CiC)	12.411	0.063	0.819	13.618	12.643	0.885	-0.249	58.89	3.88	13	0.992	0.01σ
(no $A_{\rm bias}$)	12.453	0.045	0.409	13.226	13.116			-234.72	57.46	42	0.056	1.91σ
-21.0 (high z)	12.388	0.271	1.005	13.565	12.813	0.817	-0.072	-141.88	25.89	19	0.133	1.50σ
(no CiC)	12.415	0.398	0.758	13.475	12.836	0.890	-0.549	57.89	17.13	13	0.193	1.30σ
(no $A_{\rm bias}$)	12.360	0.059	0.852	13.431	13.099			-136.33	35.43	19	0.012	2.50σ

Table 3. Confidence intervals of the HOD parameters from the 16th, 50th, and 84th percentiles of the marginalized posteriors. The confidence intervals without CiC constraints, and without assembly bias, are included for comparison.

Threshold	$\log M_{\rm min}$	$\sigma_{\log M}$	α	$\log M_1$	$\log M_0$	$A_{\rm cen}$	$A_{\rm sat}$
-20.0	$12.026^{+0.087}_{-0.069}$	$0.587^{+0.159}_{-0.136}$	$0.748^{+0.059}_{-0.065}$	$12.833^{+0.073}_{-0.094}$	$12.315^{+0.163}_{-0.145}$	$0.848^{+0.115}_{-0.210}$	$-0.028^{+0.211}_{-0.226}$
(no CiC)	$12.151^{+1.047}_{-0.274}$	$0.845^{+1.701}_{-0.635}$	$0.784^{+0.125}_{-0.149}$	$12.833^{+0.177}_{-0.287}$	$12.566^{+0.156}_{-0.329}$	$0.613^{+0.288}_{-0.556}$	$-0.260^{+0.502}_{-0.423}$
(no $A_{\rm bias}$)	$11.951^{+0.080}_{-0.063}$	$0.454^{+0.155}_{-0.164}$	$0.744^{+0.063}_{-0.057}$	$12.759^{+0.088}_{-0.084}$	$12.427^{+0.127}_{-0.185}$		
-20.5	$12.252^{+0.074}_{-0.056}$	$0.471^{+0.126}_{-0.122}$	$0.707^{+0.065}_{-0.065}$	$13.102^{+0.088}_{-0.104}$	$12.728^{+0.121}_{-0.142}$	$0.862^{+0.102}_{-0.205}$	$-0.113^{+0.217}_{-0.222}$
(no CiC)	$12.518^{+1.300}_{-0.367}$	$0.916^{+1.572}_{-0.715}$	$0.681^{+0.182}_{-0.257}$	$13.094^{+0.224}_{-0.500}$	$12.886^{+0.152}_{-0.275}$	$0.462^{+0.412}_{-0.771}$	$-0.072^{+0.607}_{-0.576}$
(no $A_{\rm bias}$)	$12.213^{+0.074}_{-0.052}$	$0.389^{+0.150}_{-0.172}$	$0.691^{+0.055}_{-0.043}$	$13.017^{+0.080}_{-0.065}$	$12.837^{+0.067}_{-0.113}$		
-21.0	$12.450^{+0.015}_{-0.012}$	$0.083^{+0.108}_{-0.057}$	$0.423^{+0.108}_{-0.071}$	$13.292^{+0.140}_{-0.100}$	$13.091^{+0.046}_{-0.098}$	$0.229^{+0.533}_{-0.758}$	$0.047^{+0.154}_{-0.228}$
(no CiC)	$12.464^{+0.125}_{-0.038}$	$0.272^{+0.293}_{-0.191}$	$0.719^{+0.165}_{-0.232}$	$13.569^{+0.112}_{-0.206}$	$12.871^{+0.236}_{-0.253}$	$0.333^{+0.501}_{-0.779}$	$-0.012^{+0.562}_{-0.522}$
(no $A_{\rm bias}$)	$12.455^{+0.022}_{-0.011}$	$0.098^{+0.160}_{-0.077}$	$0.414^{+0.091}_{-0.097}$	$13.291^{+0.119}_{-0.090}$	$13.080^{+0.048}_{-0.088}$		
-21.0 (high z)	$12.365^{+0.036}_{-0.027}$	$0.222^{+0.126}_{-0.144}$	$0.895^{+0.089}_{-0.090}$	$13.494^{+0.095}_{-0.099}$	$12.944^{+0.133}_{-0.173}$	$0.759^{+0.185}_{-0.380}$	$-0.200^{+0.200}_{-0.214}$
(no CiC)	$12.356^{+0.048}_{-0.024}$	$0.178^{+0.175}_{-0.120}$	$0.959^{+0.078}_{-0.118}$	$13.563^{+0.052}_{-0.097}$	$12.597^{+0.217}_{-0.154}$	$0.640^{+0.276}_{-0.683}$	$-0.218^{+0.252}_{-0.270}$
(no $A_{\rm bias}$)	$12.366^{+0.035}_{-0.025}$	$0.244^{+0.118}_{-0.149}$	$0.929^{+0.067}_{-0.064}$	$13.479^{+0.078}_{-0.073}$	$12.964^{+0.113}_{-0.143}$		

5. CONSTRAINING THE HOD

5.1. *HOD Model*

We employ a decorated HOD model based on the formulation of Zheng et al. (2007). In this framework, the halo occupations of central and satellite galaxies over a given magnitude threshold are described by Bernoulli and Poisson distributions, respectively. Their means are functions of halo mass M_h, described by

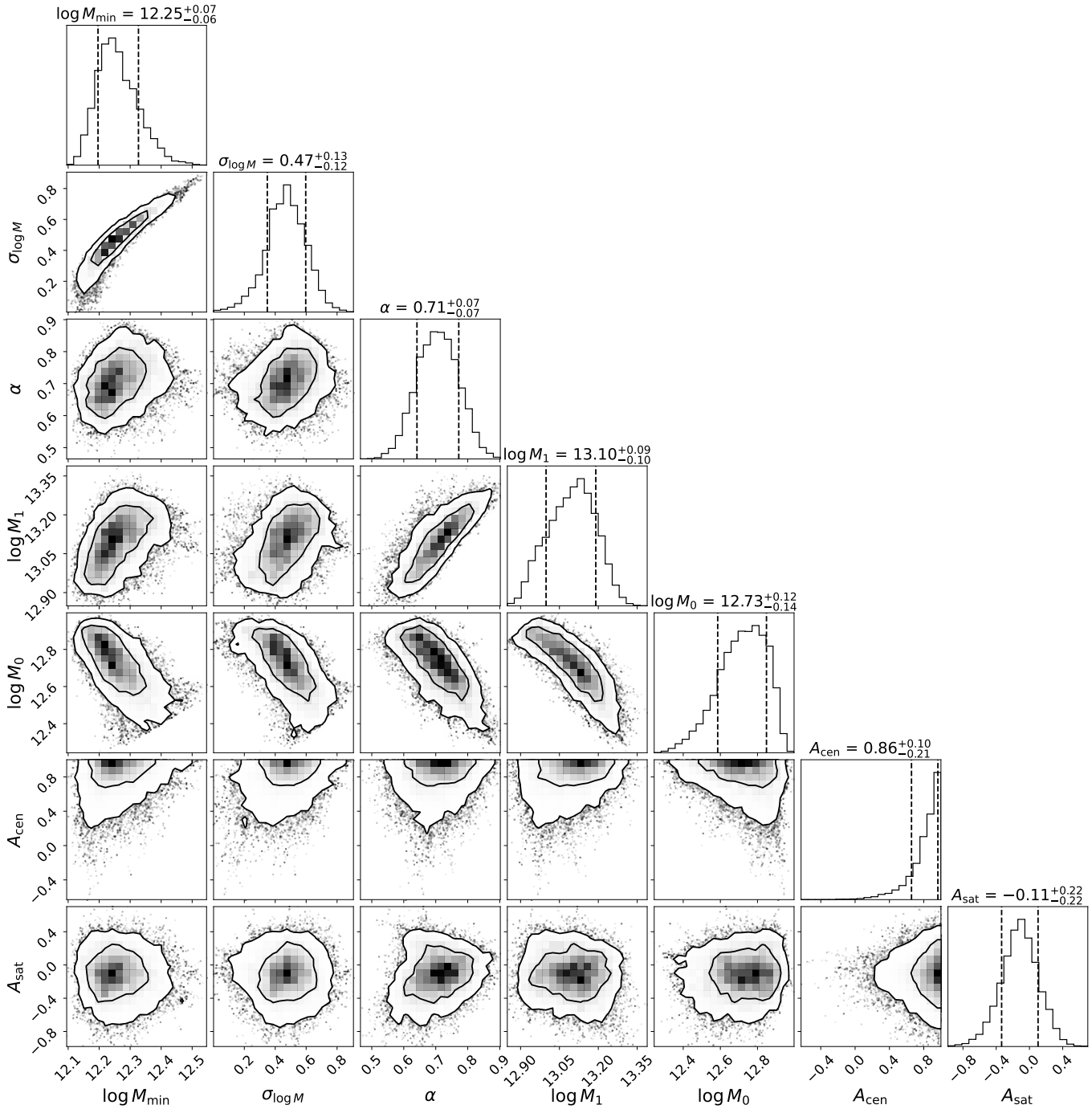

Figure 6. Posterior distribution of the HOD parameters of the -20.5 threshold sample from MCMC sampling. The 68% and 95% confidence regions are displayed by contour lines for each two-dimensional projection, and the 68% confidence intervals are marked with dashed vertical lines for each one-dimensional projection.

$$\langle N_{\rm cen}\rangle(M_h) = \frac{1}{2}\left(1 + {\rm erf}\left(\frac{\log(M_h/M_{\rm min})}{\sigma_{\log M}}\right)\right) \quad (27)$$

and

$$\langle N_{\rm sat}\rangle(M_h) = \left(\frac{M_h - M_0}{M_1}\right)^{\alpha} \quad (28)$$

where $\log M_{\rm min}$, $\sigma_{\log M}$, α, $\log M_1$, and $\log M_0$ are free parameters controlling the shape of the mean occupation functions. These parameters must be tuned separately for each magnitude threshold and redshift sample. We further parameterize $\log M_0$ into Q_0 using

$$\log M_0 = \log M_{\rm min} + Q_0(\log M_1 - \log M_{\rm min}) \quad (29)$$

which helps us ensure that $\log M_0$ always stays between $\log M_{\rm min}$ and $\log M_1$ to preserve its sensitivity to, and the stability of, our summary statistics.

Adding further flexibility into our model, we include assembly bias parameters $A_{\rm cen}$ and $A_{\rm sat}$ to introduce a halo occupation dependence on the NFW concentration. These parameters both range from [-1, 1], and allow for redistribution of the central and satellite occupation, respectively, from low to high concentration halos for positive A, or vice versa. See Hearin et al. (2016) for further details on the decorated HOD parameterization.

5.2. *MCMC Fits*

We use Markov-chain Monte Carlo (MCMC) to constrain the HOD model using each galaxy sample. We make use of the `emcee` (Foreman-Mackey et al. 2013) implementation, in which several walkers simultaneously sample a likelihood function throughout parameter space, and occasionally trade locations to construct MCMC chains. Ignoring the normalization constant, the log-likelihood is given by

$$\ln\mathcal{L} = -\frac{1}{2}(\vec{x}_{\rm model} - \vec{x}_{\rm data})^{\intercal}\Sigma^{\dagger}(\vec{x}_{\rm model} - \vec{x}_{\rm data}) \quad (30)$$

where Σ is the covariance matrix from Equation 2 and $\Sigma^{\dagger}$ is its Moore-Penrose pseudo-inverse (Penrose 1955), which prevents the reduced dimensionality of our likelihood from affecting the likelihood numerically. Loss in dimensionality occurs when we use the full CiC distribution (but not when we reduce this information into CiC moments) due to some eigenvalues in the covariance matrix equaling zero when there are at least 20 summary statistics, which is our number of jackknife realizations. We use the implementation available in the `logpdf` method of the `multivariate_normal` class from SciPy (Virtanen et al. 2020).

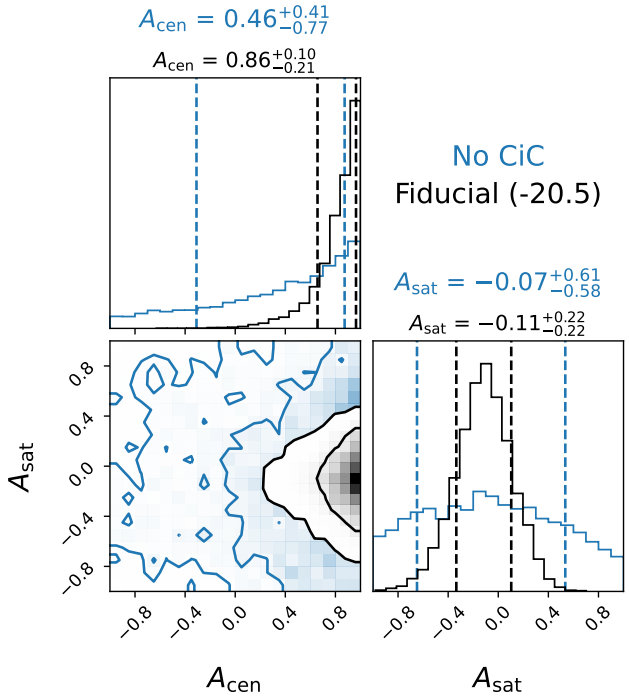

Figure 7. Posterior distribution of the assembly bias parameters of the -20.5 threshold sample from MCMC sampling. Overplotted in blue is the result we obtain without including any constraints from CiC, yielding very little information about assembly bias.

In addition, we rescale the summary statistics such that their covariance matrix has a diagonal of ones. Mathematically, this has no effect and is equivalent to an arbitrary change of units. However, this circumvents machine precision errors where the pseudo-inverse will delete the constraints of summary statistics with low orders of magnitude, like number density.

We initialize our MCMC chains around the best-fit parameters of the corresponding magnitude threshold sample from Wang et al. (2022), with very slight variation between the MCMC walkers. We let these chains run for 60,000 trial points (3,000 iterations × 20 walkers), and conservatively remove a burn-in of 2,000 trial points to calculate our posteriors displayed in Figures 6, 7, and 8, as well as the maximum-likelihood points and confidence regions reported in Tables 2 and 3, respectively. Our relatively small number of trial points is acceptable thanks to our deterministic likelihood evaluations and our prior on $\log M_0$ that confines the MCMC to a stable region of parameter space. The autocorrelation lengths of our chains ended up ranging from 100-300. This is about a factor of two shorter than the autocorrelation lengths we obtain using Monte Carlo CiC evaluations, and pos-

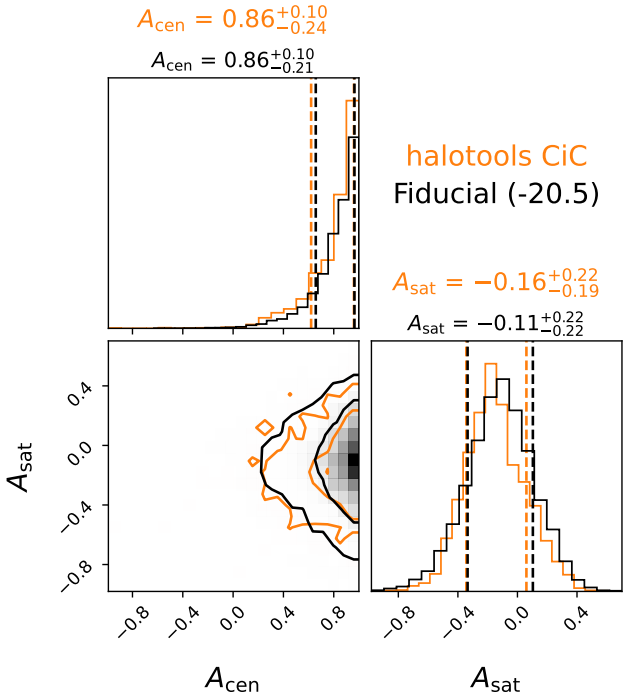

Figure 8. Posterior distribution of the assembly bias parameters of the -20.5 threshold sample from MCMC sampling. Overplotted in orange is the result we obtain from calculating CiC from `halotools` instead of `galtab` for the same number of MCMC iterations. Due to the stochasticity of the `halotools` predictions, its acceptance rate was three times lower in this case, causing much slower posterior convergence.

sibly orders of magnitudes shorter than the result from Monte Carlo $w_p(r_p)$ evaluations.

To quantify how well our maximum-likelihood models agree with the data, we calculate χ^2 along with the probability of measuring data with at least this value of χ^2 by chance using the chi-squared cumulative distribution function[5]. In Table 2, we report this probability and translate it into the z-score of a Gaussian to quantify the "number of sigmas" of tension that exists between our model and data.

6. RESULTS AND DISCUSSION

The measurements from the DESI One-Percent Survey already produce reasonably tight constraints on the HOD. For each of the four threshold samples defined in Table 1, the corresponding best-fit HOD parameters are given in Table 2, and 1σ confidence intervals are given in Table 3. We have also summarized these constraints as a function of M_r threshold and redshift into easier-to-digest plots in Figure 10. In this figure, we show that as luminosity increases from M_r of -20.0 to -21.0, the characteristic halo mass for central galaxies gradually increases from roughly $10^{12.0}$ to $10^{12.4}$ $M_\odot$. We find a similar increasing trend for the characteristic halo masses containing one (and two) satellite galaxies for each sample; the inferred slope α of the $\langle N_{sat}\rangle(M_{halo})$ relation does not evolve significantly compared to the shown error bars. Finally, we show the parameters which trace assembly bias; these are very significantly greater than zero for centrals in the lower two magnitude threshold samples, while satellite assembly bias is consistent with zero throughout. With only one $z = 0.25$ sample, we find no significant signals of redshift evolution.

Given the current relatively small sample sizes, the tightness of our constraints can be attributed to the power of combining information from w_p and CiC. We find a 3σ detection of assembly bias for central galaxies in the two lower luminosity bins. More precisely, the strength of the evidence for central assembly bias in each sample is as follows:

- For our -20.0 and -20.5 samples, the posterior probability that $A_{cen} > 0$ is 0.9987 and 0.995, respectively. Without CiC constraints, these probabilities are only 0.860 and 0.737.
- Positive assembly bias at $M_r < -21.0$ is favored significantly only in the $z \sim 0.25$ sample. For it, we find a posterior probability for $A_{cen} > 0$ of 0.948 (or 0.828 without CiC constraints).
- There are very poor constraints on assembly bias at $M_r < -21.0$ in our $z \sim 0.15$ sample whether or not we include CiC in the sample.

The constraints we find on assembly bias are consistent with the findings from studies based on SDSS data. Despite the smaller sample size currently available from DESI, our $w_p(r_p)$ + CiC analysis produces much stronger constraints than characterizing SDSS clustering with $w_p(r_p)$ alone (e.g., Zentner et al. 2019; Vakili & Hahn 2019). In fact, we achieve very similar constraining power to Wang et al. (2022), even though we use the same set of summary statistics. This may imply that the assembly bias signal is stronger at the higher redshifts probed by BGS. Additionally, the purity of the DESI samples may be higher due to the high targeting completeness in the 1% survey, which allows us to avoid having to assign redshifts to untargeted galaxies based upon the nearest neighbors in the sky.

While the HOD model can consistently produce good fits to w_p and n simultaneously (possibly to the point of overfitting), incorporating CiC measurements results

[5] i.e., the Newman Score

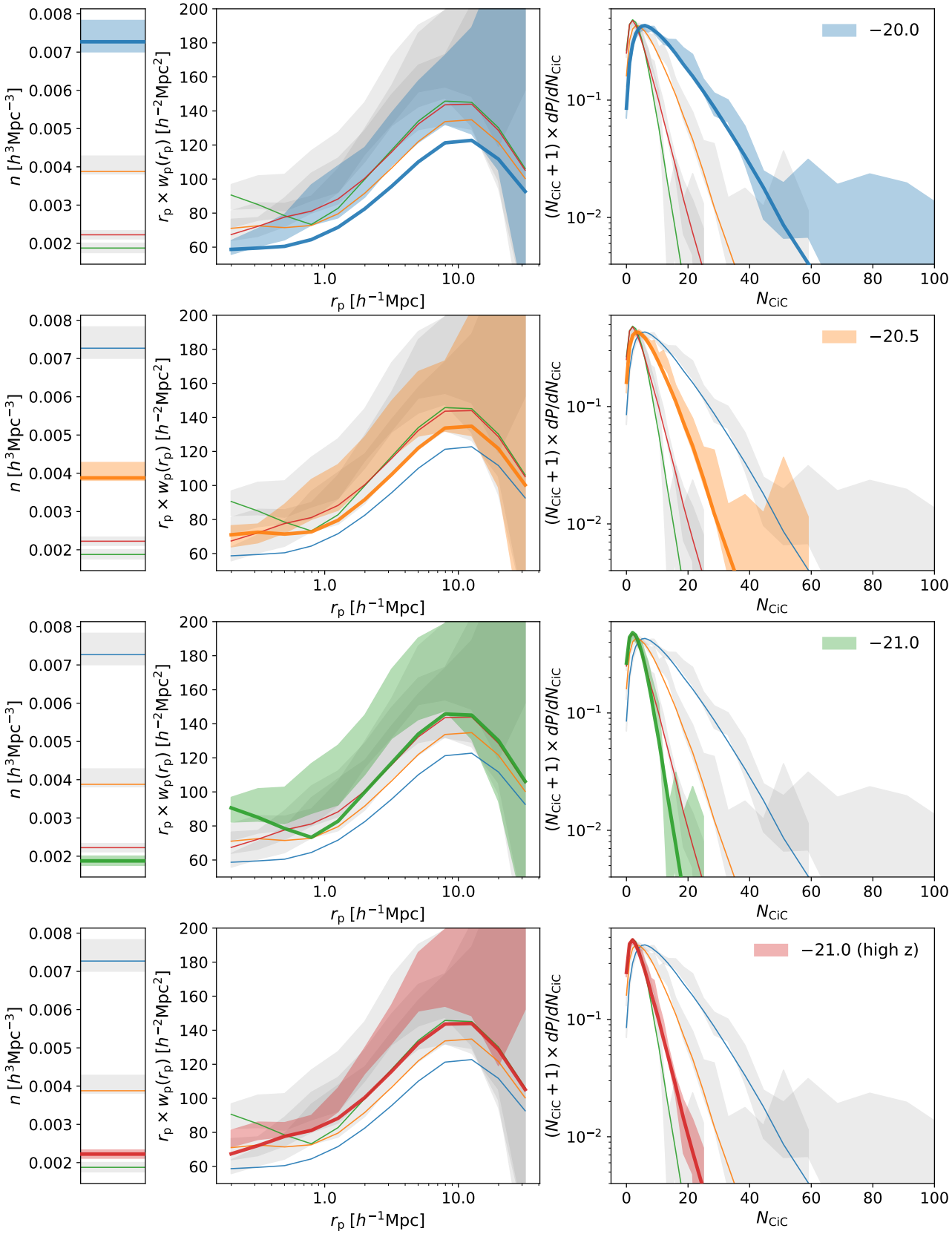

Figure 9. Measurements and our maximum-likelihood predictions of number density (left panels), the projected correlation function (center panels), and the CiC distribution (right panels). The 1σ confidence intervals from the measurements of a given quantity are represented by shaded regions of the color corresponding to the sample, while the maximum-likelihood predictions are represented by solid lines following the same color scheme. The parameters of the best-fit models and their tensions versus the data are reported in Table 2.

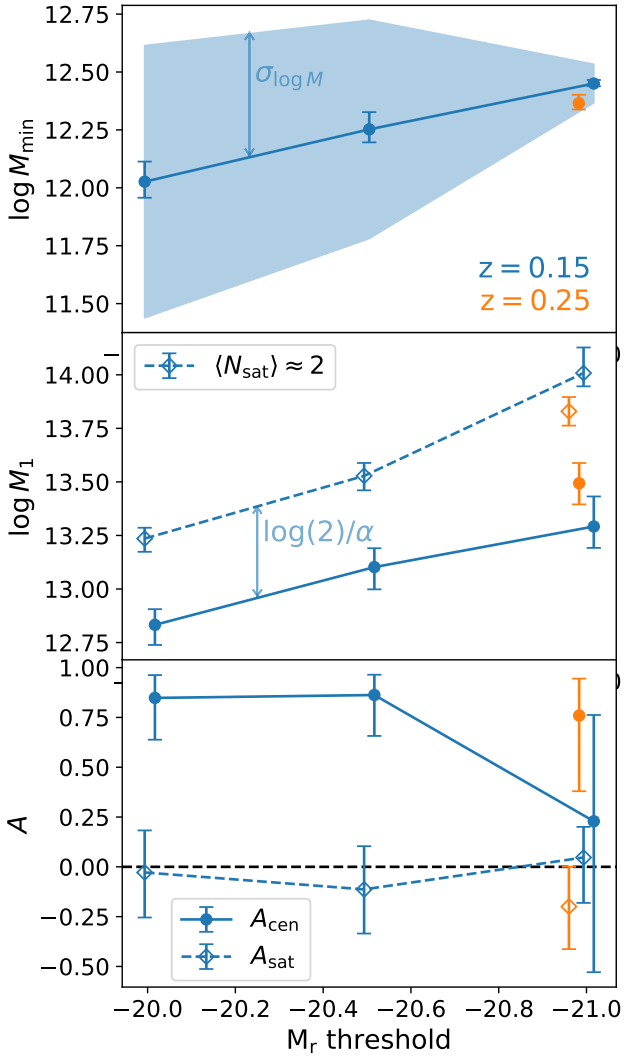

Figure 10. Variation of HOD parameters with luminosity and redshift. Median values of the one-dimensional marginalized posteriors for the characteristic masses, $\log M_{\rm min}$ (top panel) and $\log M_1$ (middle panel) are plotted, as well as the assembly bias parameters $A_{\rm cen}$ and $A_{\rm sat}$ (bottom panel). The capped error bars on these points span the 16th to the 84th percentile of the posterior for a given parameter. Median values derived from our posteriors of other HOD parameters $\sigma_{\log M}$ (top panel) and α (middle panel) are labeled; $\sigma_{\log M}$ characterizes the spread in the M_r-$M_{\rm halo}$ relation, and $\log(2)/\alpha$ characterizes the log-difference between the halo masses corresponding to $\langle N_{\rm sat} \rangle \approx 1$ and $\langle N_{\rm sat} \rangle \approx 2$. We apply small x-offsets to easily distinguish the points, but all M_r thresholds are exactly -20.0, -20.5, or -21.0.

in mismatches between the model and data in some cases. Although introducing assembly bias parameters has slightly reduced this tension, the $M_r < -21.0$ sample at $z \sim 0.15$ still exhibits a tension of nearly 2σ between our models and the data. This tension is reported in Table 2 and is readily apparent in Figure 9 (though one must use caution when assessing the mismatch by eye since the summary statistics can be strongly covariant).

Significant tension in only one of our four samples by no means rules out the HOD model used, but it should incentivize us to consider what else the model might be missing. In the coming years, the size of the DESI sample will grow by a factor of 100 compared to what was used here, so we can expect that the constraints will tighten significantly and tensions may grow. Our model is not sufficiently flexible to fit early data samples well; therefore, it is plausible that these models could be ruled out convincingly with the full dataset. Future studies should explore additional ways to make the HOD more flexible such that they can produce better fits to the DESI data; we describe a few plausible extensions here, but by no means exhaust the possibilities.

As one example, the HOD we have used in this work assumes that the stellar-to-halo-mass relation has a log-normal scatter, but the UniverseMachine simulations (Behroozi et al. 2019) exhibit a slight skew to this scatter in several tested samples. In principle, it is simple to test the addition of one more parameter to allow a skew-log-normal scatter.

Another modification that may be justified is to relax the assumed isotropic NFW distribution of satellite galaxies. This is a common assumption, yet it has long been known that the distribution of subhalos is anisotropic, due to the preferential accretion of mergers along filaments (Zentner et al. 2005). Additionally, recent studies have found a significant difference in the radial profile of the halo mass associated with subhalos from NFW (Fielder et al. 2020; Mezini et al. 2023). Such modifications would be more complex but will be particularly important as small-scale clustering measurements improve since they are sensitive to the spatial distribution of satellites.

Additionally, we have only tested for assembly bias tied to halo concentration, and have ignored other occupation correlations that may be based upon halo spin or age (Contreras et al. 2021; Sato-Polito et al. 2019). Another possibility is that the occupation of satellites is correlated with the occupation of the central in the same halo due to galactic conformity (Berti et al. 2017; Kauffmann et al. 2013). Both of these possibilities would likely produce similar statistical imprints. However, a primary question to investigate is whether these alternate assumptions lead to a biased inference of HOD parameters such as characteristic halo masses and as-

sembly bias. If so, all of our results could be overly confident[6].

While CiC plays a crucial role in the HOD constraints obtained via our analysis, it is also our computational bottleneck. However, we have significantly sped up this process with `galtab`, particularly by removing the stochasticity of likelihood evaluations, which greatly improves the MCMC convergence rate. Using a stochastic estimator, convergence is especially problematic for the lowest-number-density, brightest-threshold samples, which exhibit order-of-magnitude increases in the acceptance rates of their MCMC chains.

Depending on the computing resources available and the dimensionality of the analysis, `galtab` may provide even more drastic speedups. Due to the implementation in JAX, the expensive steps are automatically executed on a GPU when available. Additionally, our framework allows the predictions to be differentiable with respect to HOD parameters (assuming the occupation model is compatible with JAX arrays, for which those available in `halotools` require slight modifications). In principle, this allows for the use of alternative MCMC methods with improved scalability to high-dimensional or strongly covariant posterior estimation, such as Hamiltonian Monte Carlo (Neal 2011).

Our development of the `galtab` package provides a useful tool for further analyses of the galaxy-halo connection that may require differentiable predictions. By combining these new tools with upcoming enlarged samples from DESI, we anticipate that coming studies will soon shift focus from mere detections of assembly bias to studying its implications for galaxy formation in much finer detail.

APPENDIX

A. SHAP FEATURE IMPORTANCE CALCULATIONS

As briefly discussed in Section 3 and plotted in Figure 3, we have roughly quantified the importance of each summary statistic in inferring the HOD model parameters by testing how influential each quantity for machine learning-based predictions. We performed this test using an artificial dataset based upon uniformly sampling 1000 sets of HOD parameters via Latin Hypercube Sampling over the projected one-dimensional 1σ confidence interval of the fiducial fits for the $M_r < -20.5$ threshold sample of Wang et al. (2022).

For each of the 1000 sets of HOD parameters, we predicted the values of all of the summary statistics via the methods described in Section 4.6. We then trained a scikit-learn (Pedregosa et al. 2011) random forest regression model to perform the inverse mapping (i.e., predicting HOD parameters from the values of the summary statistics).

We then calculate SHAP feature importance values for each feature (i.e., each quantity used as an input to the random forest). SHAP values are explained in detail in Lundberg & Lee (2017). In brief, they attempt to quantify the amount of "impact" each feature has on model predictions. To be explicit, a large positive SHAP value corresponds to a feature for which increases in the feature value cause large increases in model predictions, and vice versa. This allows us to analyze and distinguish the effects of positive or negative changes in each feature on the model predictions that result.

We show the full beeswarm distribution of SHAP values for each HOD parameter in Figure 11. We assign importance values shown in Figure 3 by taking the mean absolute values of these distributions. Features that have large SHAP importances will correspond to those quantities which are most useful for predicting a given HOD parameter.

ACKNOWLEDGMENTS

This research has made extensive use of the arXiv and NASA's Astrophysics Data System. This research has made use of adstex (https://github.com/yymao/adstex).

Software: Halotools (Hearin et al. 2016), JAX (Bradbury et al. 2018), Corrfunc (Sinha & Garrison 2020), emcee (Foreman-Mackey et al. 2013), corner.py (Foreman-Mackey 2016), scikit-learn (Pedregosa et al. 2011), SciPy (Virtanen et al. 2020), matplotlib (Hunter 2007), Astropy (Astropy Collaboration et al. 2018), NumPy (van der Walt et al. 2011),

REFERENCES

[6] i.e., Zentner Points™

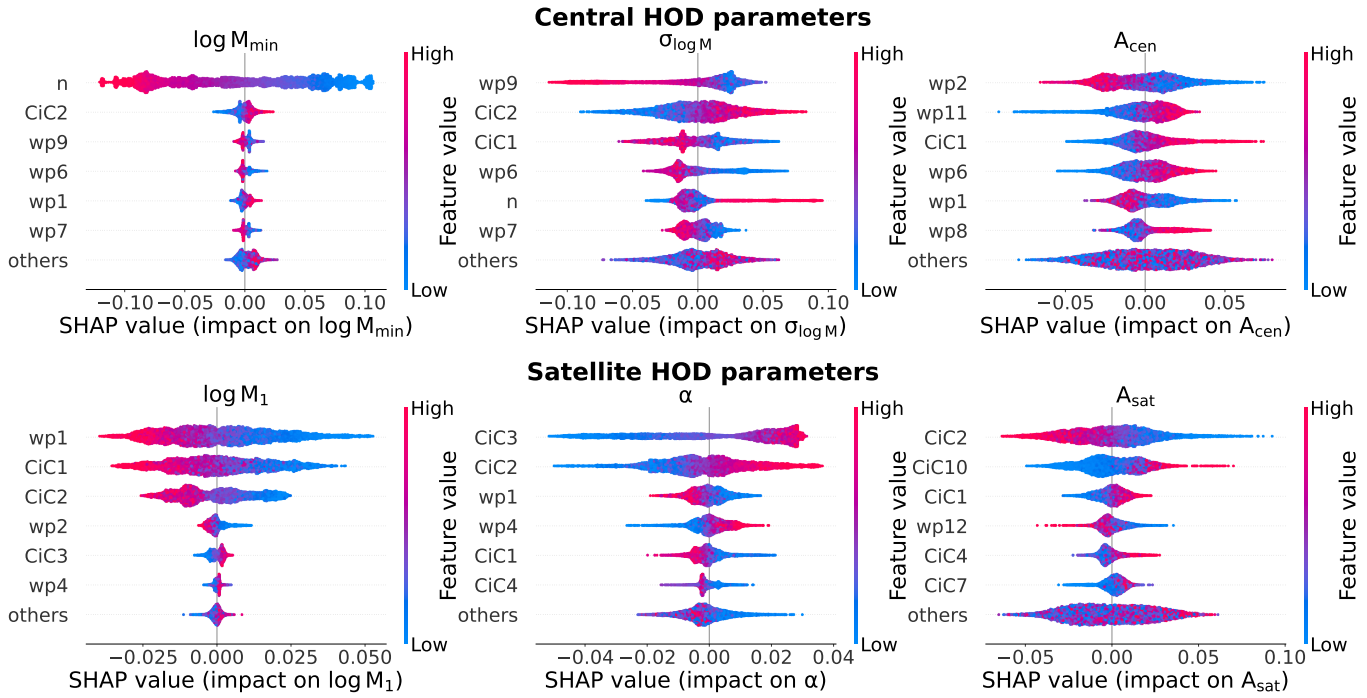

Figure 11. The impact of each of our summary statistics on HOD inference, based upon SHAP feature importances. The upper panel of each sub-figure shows beeswarms of the SHAP values for each feature's impact on predicting the given HOD parameter. Each panel shows the six most important quantities in order of importance, and the panels are organized in the same way those in Figure 3. See Figure 3 for a more condensed version of this information which focuses on the mean absolute SHAP value as an importance metric.

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543, doi: 10.1088/0067-0049/182/2/543

Abbott, T. M. C., Abdalla, F. B., Alarcon, A., et al. 2018, PhRvD, 98, 043526, doi: 10.1103/PhysRevD.98.043526

Adelberger, K. L., Steidel, C. C., Giavalisco, M., et al. 1998, ApJ, 505, 18, doi: 10.1086/306162

Anderson, L., Aubourg, E., Bailey, S., et al. 2012, MNRAS, 427, 3435, doi: 10.1111/j.1365-2966.2012.22066.x

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123, doi: 10.3847/1538-3881/aabc4f

Behroozi, P., Wechsler, R. H., Hearin, A. P., & Conroy, C. 2019, MNRAS, 488, 3143, doi: 10.1093/mnras/stz1182

Behroozi, P. S., Wechsler, R. H., & Wu, H.-Y. 2013, ApJ, 762, 109, doi: 10.1088/0004-637X/762/2/109

Berlind, A. A., & Weinberg, D. H. 2002, ApJ, 575, 587, doi: 10.1086/341469

Berti, A. M., Coil, A. L., Behroozi, P. S., et al. 2017, ApJ, 834, 87, doi: 10.3847/1538-4357/834/1/87

Beutler, F., Blake, C., Colless, M., et al. 2011, MNRAS, 416, 3017, doi: 10.1111/j.1365-2966.2011.19250.x

Bianchi, D., & Percival, W. J. 2017, MNRAS, 472, 1106, doi: 10.1093/mnras/stx2053

Bradbury, J., Frostig, R., Hawkins, P., et al. 2018, JAX: composable transformations of Python+NumPy programs, 0.3.13. http://github.com/google/jax

Breiman, L. 2001, Machine Learning, 45, 5, doi: 10.1023/A:1010933404324

Contreras, S., Chaves-Montero, J., Zennaro, M., & Angulo, R. E. 2021, MNRAS, 507, 3412, doi: 10.1093/mnras/stab2367

DESI Collaboration, Abareshi, B., Aguilar, J., et al. 2022, AJ, 164, 207, doi: 10.3847/1538-3881/ac882b

DESI Collaboration, Adame, A. G., Aguilar, J., et al. 2023, arXiv e-prints, arXiv:2306.06308. https://arxiv.org/abs/2306.06308

Fielder, C. E., Mao, Y.-Y., Zentner, A. R., et al. 2020, MNRAS, 499, 2426, doi: 10.1093/mnras/staa2851

Foreman-Mackey, D. 2016, The Journal of Open Source Software, 1, 24, doi: 10.21105/joss.00024

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306, doi: 10.1086/670067

Hearin, A. P., Zentner, A. R., van den Bosch, F. C., Campbell, D., & Tollerud, E. 2016, MNRAS, 460, 2552, doi: 10.1093/mnras/stw840

Hubble, E. P. 1936, Realm of the Nebulae

Hunter, J. D. 2007, Computing in Science and Engineering, 9, 90, doi: 10.1109/MCSE.2007.55

Ishiyama, T., Prada, F., Klypin, A. A., et al. 2021, MNRAS, 506, 4210, doi: 10.1093/mnras/stab1755

Kauffmann, G., Li, C., Zhang, W., & Weinmann, S. 2013, MNRAS, 430, 1447, doi: 10.1093/mnras/stt007

Klypin, A., Yepes, G., Gottlöber, S., Prada, F., & Heß, S. 2016, MNRAS, 457, 4340, doi: 10.1093/mnras/stw248

Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64, doi: 10.1086/172900

Lundberg, S., & Lee, S.-I. 2017, arXiv e-prints, arXiv:1705.07874, doi: 10.48550/arXiv.1705.07874

Mezini, L., Fielder, C. E., Zentner, A. R., et al. 2023, arXiv e-prints, arXiv:2304.13809, doi: 10.48550/arXiv.2304.13809

Neal, R. 2011, in Handbook of Markov Chain Monte Carlo, 113–162

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825, doi: 10.48550/arXiv.1201.0490

Peebles, P. J. E. 1980, The large-scale structure of the universe

Penrose, R. 1955, Proceedings of the Cambridge Philosophical Society, 51, 406, doi: 10.1017/S0305004100030401

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, A&A, 641, A6, doi: 10.1051/0004-6361/201833910

Reddick, R. M., Wechsler, R. H., Tinker, J. L., & Behroozi, P. S. 2013, ApJ, 771, 30, doi: 10.1088/0004-637X/771/1/30

Reid, B. A., & Spergel, D. N. 2009, ApJ, 698, 143, doi: 10.1088/0004-637X/698/1/143

Sato-Polito, G., Montero-Dorta, A. D., Abramo, L. R., Prada, F., & Klypin, A. 2019, MNRAS, 487, 1570, doi: 10.1093/mnras/stz1338

Sinha, M., & Garrison, L. H. 2020, MNRAS, 491, 3022, doi: 10.1093/mnras/stz3157

Storey-Fisher, K., Tinker, J., Zhai, Z., et al. 2022, arXiv e-prints, arXiv:2210.03203, doi: 10.48550/arXiv.2210.03203

Vakili, M., & Hahn, C. 2019, ApJ, 872, 115, doi: 10.3847/1538-4357/aaf1a1

van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science and Engineering, 13, 22, doi: 10.1109/MCSE.2011.37

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature Methods, 17, 261, doi: 10.1038/s41592-019-0686-2

Wang, K., Mao, Y.-Y., Zentner, A. R., et al. 2022, MNRAS, 516, 4003, doi: 10.1093/mnras/stac2465

—. 2019, MNRAS, 488, 3541, doi: 10.1093/mnras/stz1733

White, S. D. M. 1979, MNRAS, 186, 145, doi: 10.1093/mnras/186.2.145

Zehavi, I., Zheng, Z., Weinberg, D. H., et al. 2005, ApJ, 630, 1, doi: 10.1086/431891

Zentner, A. R., Hearin, A., van den Bosch, F. C., Lange, J. U., & Villarreal, A. 2019, MNRAS, 485, 1196, doi: 10.1093/mnras/stz470

Zentner, A. R., Kravtsov, A. V., Gnedin, O. Y., & Klypin, A. A. 2005, ApJ, 629, 219, doi: 10.1086/431355

Zheng, Z., Coil, A. L., & Zehavi, I. 2007, ApJ, 667, 760, doi: 10.1086/521074

Zwicky, F. 1957, Morphological astronomy