Multi-wavelength mock galaxy catalogs of the low-redshift Universe

Aseem Paranjape\(^1\)*, Tirthankar Roy Choudhury\(^2\)† & Ravi K. Sheth\(^3,4\)‡

\(^1\) Inter-University Centre for Astronomy & Astrophysics, Ganeshkhind, Post Bag 4, Pune 411007, India
\(^2\) National Centre for Radio Astrophysics, TIFR, Post Bag 3, Ganeshkhind, Pune 411007, India
\(^3\) Center for Particle Cosmology, University of Pennsylvania, 209 S. 33rd St., Philadelphia, PA 19104, USA
\(^4\) The Abdus Salam International Center for Theoretical Physics, Strada Costiera, 11, Trieste 34151, Italy

**ABSTRACT**

We present a new suite of mock galaxy catalogs mimicking the low-redshift Universe, based on an updated halo occupation distribution (HOD) model and a scaling relation between optical properties and the neutral hydrogen (HI) content of galaxies. Our algorithm is constrained by observations of the luminosity function and luminosity- and colour-dependent clustering of SDSS galaxies, as well as the HI mass function and HI-dependent clustering of massive HI-selected galaxies in the ALFALFA survey. Mock central and satellite galaxies with realistic values of \(r\)-band luminosity, \(g-r\) and \(u-r\) colour, stellar mass and HI mass are populated in an N-body simulation, inheriting a number of properties of the density and tidal environment of their host halos. The host halo of each central galaxy is also ‘baryonified’ with realistic spatial distributions of stars as well as hot and cold gas, along with the corresponding rotation curve. Our default HOD assumes that galaxy properties are a function of group halo mass alone, and can optionally include effects such as galactic conformity and colour-dependent galaxy assembly bias. The mocks predict the relation between the stellar mass and HI mass of massive HI galaxies, as well as the 2-point cross-correlation function of spatially co-located optical and HI-selected samples. These mocks form the basis of a number of applications we will pursue in other work, such as novel null tests for galaxy assembly bias, predictions for the HI velocity width function and testing the universality of the radial acceleration relation.

**Key words:** galaxies: formation - cosmology: theory, dark matter, large-scale structure of Universe - methods: numerical

1 INTRODUCTION

Contemporary studies of galaxy evolution and cosmology must explore a multitude of physical and statistical properties of the observed large-scale structure of the Universe. The computational and technical challenges involved in any such analysis, whether observational or theoretical, mean that mock galaxy catalogs are now a staple tool of cosmological analyses.

Current and upcoming large-volume surveys of the Universe, aiming to extract cosmological information using observables including the redshift space clustering of galaxies at large and small scales, weak lensing, the abundances of clusters and voids, etc., increasingly rely on the use of mock galaxy catalogs for a variety of applications. Apart from calibrating expected measurement covariances and end-to-end pipeline testing (Mao et al. 2018), such catalogs also serve as excellent test-beds for exploring ideas related to the galaxy-dark matter connection, the nature of dark matter or the predictions of alternative gravity theories, and the effects of dark energy. As such, it is critical to develop and calibrate mock-making algorithms, constrained by existing observations, that can accurately account for the multi-scale, multi-probe connection between baryonic matter in and around galaxies and the dark cosmic web in which these galaxies reside. This is the primary motivation behind the present work.

Computationally speaking, the most efficient algorithms are those which model the baryon-dark matter connection using empirical, statistical tools that are motivated by the Halo Model (see Cooray & Sheth 2002, for a review). These include the halo occupation distribution (HOD; Zehavi et al. 2011), conditional luminosity function (CLF; Yang et al. 2018, and references therein) or subhalo abundance matching (SHAM; Behroozi et al. 2019, and references therein) prescriptions via a scaling relation between the luminosity function and the mass function of dark matter halos, or via a prescription that links luminosity and halo mass as a function of group halo mass alone.
that are constrained by observed galaxy abundances and clustering. At the other end of the spectrum, lie full-fledged cosmological hydrodynamical simulations of galaxy formation (e.g., Vogelsberger et al. 2014; Dubois et al. 2014; Schaye et al. 2015; Springel et al. 2018), arguably the most realistic and most expensive tool in computational cosmology. Semi-analytical models (SAMs), which evolve simplified physical descriptions of galaxy formation and evolution within the cosmic web of gravity-only simulations, lie somewhere between full simulations and empirical models, in terms of both computational complexity as well as fidelity to observational constraints (for a recent review, see Somerville & Davé 2015). The present work focuses on HOD models.

Mock-making algorithms based on the HOD or SHAM frameworks have been frequently used in the literature in conjunction with large-volume surveys at low and intermediate redshifts (Manera et al. 2013; de la Torre & Peacock 2013; de la Torre et al. 2013; Kitaura et al. 2016; Mao et al. 2018; Alam et al. 2020; Zhao et al. 2020; Sugiyama et al. 2020). These algorithms typically segregate into those describing stellar populations and overall star formation activity (constrained by a multitude of galaxy surveys at lowand high-redshift spanning wavelengths from the infrared to ultraviolet), and others focused on the distribution of gas, primarily in the form of neutral hydrogen (HI, constrained by radio wavelength observations, typically at low-redshift). We broadly refer to the former category using the label ‘optical’ and the latter as ‘HI’ algorithms. All such algorithms typically rely on dark halos identified in gravity-only cosmological simulations, along with statistical prescriptions to paint galaxies into these halos.

Algorithms for assigning single band optical luminosities (or stellar masses) to mock galaxies using HOD, CLF or SHAM prescriptions have existed for about two decades (Cooray & Sheth 2002; Vale & Ostriker 2004; Reddick et al. 2013), with relatively recent extensions to include colours (or star formation rates) (Skibba & Sheth 2009; Hearin & Watson 2013; Contreras et al. 2020). While the simplest occupation models single out halo mass as the primary driver of observed correlations between galaxy properties and their environments (e.g., Abbas & Sheth 2007; Zu & Mandelbaum 2015; Paranjape et al. 2018b; Alam et al. 2019), recent work has argued for the importance of modelling beyond-mass effects such as assembly bias and galactic conformity (Zentner et al. 2014), leading to tune-able prescriptions for these effects (Masaki et al. 2013; Paranjape et al. 2015; Hearin et al. 2016a; Yuan et al. 2018; Xu et al. 2020; Contreras et al. 2020). State-of-the-art implementations such as the UniverseMachine prescription of Behroozi et al. (2019) employ SHAM on entire merger trees in high-resolution gravity-only simulations, calibrated to reproduce stellar mass functions and star formation rates over a wide range of redshifts.

On the HI side, mock catalogs have been created using a combination of galaxy formation SAMs and a prescription for distributing the HI in disks (see, e.g., Obreschkow et al. 2009), which have been useful for planning upcoming surveys with telescopes such as the Square Kilometre Array (SKA). If one is only interested in the very large-scale correlations of the HI distribution (e.g., for intensity mapping experiments), or in interpreting HI detections based on stacking experiments at high redshift, the algorithms for creating the catalogs are considerably simpler, assigning an HI mass directly to dark halos using a physically motivated prescription (Bagla et al. 2010; Guha Sarkar et al. 2012; Villaescusa-Navarro et al. 2014; Castorina & Villaescusa-Navarro 2017; Padmanabhan et al. 2017).

In this paper, we aim to combine low-redshift ($z \lesssim 0.1$) constraints on galaxy abundances and clustering, from surveys of both optically selected as well as HI-selected galaxies, to generate mock galaxies that are simultaneously assigned multi-band optical information as well as HI masses. To this end, we consolidate recent work in these areas and introduce mock galaxy catalogs constructed using updated halo occupation models and calibration of multi-wavelength low-redshift galaxy optical and HI properties. Each mock galaxy in our catalogs is assigned values of the $r$-band absolute magnitude $M_r$ (detailed definition below), colour indices $g - r$ and $u - r$, HI mass $m_{\text{HI}}$ and stellar mass $m_{*}$, along with a range of environmental properties derived from the dark matter environment of the galaxy’s host halo. Relying on these properties, the mocks reproduce the luminosity function and the luminosity and colour dependence of projected 2-point clustering of SDSS galaxies, the HI mass function and $m_{\text{HI}}$-dependent clustering of massive ALFALFA galaxies, and predict the cross-correlations between these galaxies. The mocks also have tunable implementations of galactic conformity and colour-dependent galaxy assembly bias.

Additionally, the host halo of each central galaxy in our mocks is ‘baryonified’, i.e., assigned a realistic spatial distribution of stars and gas and, consequently, a realistic rotation curve for the galactic disk. This is a novel feature of our mocks which, as we discuss later, potentially allows us to explore a number of interesting questions that have not been adequately addressed in the theoretical literature. Among others, these include modelling the observed 21cm line profiles of HI-selected galaxies, along with the associated velocity width distribution, and the nature and universality of the radial acceleration relation in the baryons+cold dark matter (CDM) paradigm.

The rest of the paper is organised as follows. We describe the ingredients of our mock algorithm in section 2, followed by a detailed description of the algorithm itself in section 3. In section 4, we demonstrate the performance of our mocks in reproducing a number of 1-point and 2-point statistical observables. In section 5, we discuss observables whose behaviour is predicted by our mocks, along with possible extensions of our technique that are interesting for future analyses. We conclude in section 6 with a brief discussion of potential applications of our mocks.

Throughout, we consider a flat ΛCDM cosmology with parameters $(\Omega_m, \Omega_{\Lambda}, h, n_s, \sigma_8)$ given by $\{0.276, 0.045, 0.7, 0.961, 0.811\}$ compatible with the 7-year results of the Wilkinson Microwave Anisotropy Probe experiment (WMAP7, Komatsu et al. 2011), with a linear theory transfer function generated by the code CAMB (Lewis et al. 2000). Our convention will be to quote halo masses ($m$) in $h^{-1}M_{\odot}$ and galaxy stellar masses ($m_*$) and HI masses ($m_{\text{HI}}$) in $h^{-2}M_{\odot}$ units. The notation $m$ for halo mass will refer to $m_{200c}$, the mass enclosed in the radius $R_{200c}$, where the enclosed density falls to 200 times the background density. Similarly $m_{\text{vir}}$ will refer to $m_{200c}$, the mass enclosed in the radius $R_{200c}$, where the enclosed density falls to 200 times the critical density.

http://camb.info
2 INGREDIENTS

We start by describing the ingredients used in constructing our mocks. These include the gravity-only simulations that we populate with galaxies, the observed galaxy sample whose optical properties form the basis of the halo occupation distribution (HOD) we use to assign luminosities, colours and stellar masses to the mock galaxies, and the scaling relation between optical properties and neutral hydrogen (H\text{i}) using which we assign H\text{i} masses.

2.1 Simulations

The N-body simulations we rely on are listed in Table 1 of Paranjape & Alam (2020), of which we focus on the WMAP7 configurations. Specifically, we have 2, 10 and 3 realisations each of the L150N1024, L300N1024 and L600N1024 boxes, respectively, corresponding to particle masses $m_p = 2.41 \times 10^8, 1.93 \times 10^8, 1.54 \times 10^{10} \text{M}_\odot$, respectively. The notation L150N1024, for example, indicates a cubic, periodic box of $N_{1024}$ and $L_{600} N_{1024}$ boxes, respectively.

The simulations were performed using the tree-PM code GADGET-2 (Springel 2005)\(^2\) with a PM grid of a factor 2 finer than the initial particle count along each axis, and a comoving force softening length of 1/30 of the mean interparticle spacing. Initial conditions were generated using 2\(^nd\) order Lagrangian perturbation theory (Scoccimarro 1998) with the code MUSIC (Hahn & Abel 2011).\(^3\) Halos were identified using the code ROCKSTAR (Behroozi et al. 2013a)\(^4\) which performs a Friends-of-Friends (FoF) algorithm in 6-dimensional phase space. We discard all sub-halos and further only consider objects whose ‘virial’ energy ratio $\eta = 2T/|U|$ satisfies $0.5 < \eta < 1.5$ (Bett et al. 2007). All the simulations and analysis were performed on the Perseus and Pegasus clusters at IUCAA.\(^5\)

2.2 Galaxy sample

We rely on optical properties of galaxies in the local Universe as provided by Data Release 7 (DR7, Abazajian et al. 2009) of the Sloan Digital Sky Survey (SDSS, York et al. 2000).\(^6\) From the SDSS DR7 Catalog Archive Server (CAS),\(^7\) we obtained galaxy properties including Galactic extinction-corrected apparent magnitudes (luptitudes) in the $u, g$ and $r$ bands for all galaxies with spectroscopic redshifts in the range $0.02 \leq z \leq 0.2$ and satisfying the Petrosian $r$-band apparent magnitude threshold $m_r \leq 17.7$. Both Petrosian and Model magnitudes were obtained from the database. Absolute magnitudes $M_{0,u}, M_{0,g}, M_{0,r}$, were estimated by K-correcting to rest frame bands at $z = 0.1$ using K-CORRECT (Blanton & Roweis 2007) (we used a modified version of the Python wrapper provided by N. Raseliarison\(^8\)) and evolution correcting as described by Blanton et al. (2003). We did not correct for dust extinction in the host, which makes edge-on spirals appear redder. This makes our colour-dependent analysis consistent with that of Zehavi et al. (2011) who reported measurements of colour-dependent clustering using similarly uncorrected colours. Correcting for inclination can, in principle, affect inferences regarding the physics of quenching in satellites, as well as the physics governing the H\text{i} content of optically red galaxies, which we will explore in future work. (Consistency with Zehavi et al. 2011 is also why we do not work with the improved SDSS photometry discussed in Meert et al. 2015.) Flux measurement errors were accounted for when using K-CORRECT, but not explicitly in the Gaussian mixture fitting below.

This analysis yielded values of $M_r \equiv M_{0,r} + 5 \log_{10}(h)$, and similarly $M_g$ and $M_u$, for each galaxy. The latter were converted to the colour indices $g - r = M_g - M_r$ and $u - r = M_u - M_r$, which are therefore rest frame colours, K-corrected and evolution corrected to $z = 0.1$. Below, we use Petrosian absolute magnitudes $M_r$ and Model colours $g - r$ and $u - r$.

2.3 Optical halo occupation distribution

Here we discuss the complete optical HOD we use for assigning luminosities and colours to mock galaxies. This HOD is constrained by measurements of luminosity- and colour-dependent clustering in SDSS DR7 and uses a Gaussian mixture description of bi-variate colour distributions, as described below.

2.3.1 Constraints from luminosity-dependent clustering

We use the standard 5-parameter mass-only HOD calibrated by Paul et al. (2019) for the WMAP7 cosmology, which describes SDSS luminosity-dependent clustering measurements from Zehavi et al. (2011). The HOD was calibrated using simulation-based tables of 2-point halo correlation functions and halo profiles from the $z = 0$ outputs of the simulations described above, following the technique of Zheng & Guo (2016). Satellites were assumed to be distributed according to the spherically averaged dark matter distribution in parent halos, without assuming the (Navarro et al. 1996, NFW) profile (although the latter is an excellent approximation over the spatial and mass dynamic range of interest). For the 2-halo terms, hard-sphere halo exclusion was implemented, but all other correlations were directly measured from the simulations in narrow mass bins, so that non-linear, scale-dependent halo bias was accurately modelled. See Paul et al. (2019) for further discussion.

2.3.2 Constraints from colour-dependent clustering

Paul et al. (2019) also modelled colour-dependent clustering from Zehavi et al. (2011) by treating the ‘red fraction’ $f_{\text{red}}(M_r)$ of satellites of luminosity $M_r$ as a free parameter. Here $f_{\text{red}}(M_r)$ is the fraction of satellites in a luminosity bin (labelled $M_r$) whose $g - r$ colours satisfy

$$g - r > (g - r)_{\text{cut}}(M_r) \equiv 0.21 - 0.03 M_r,$$

(equation 13 of Zehavi et al. 2011). Paul et al. (2019) propagated this cut into the halo model formalism to observationally constrain $f_{\text{red}}(M_r)$ in the four wide luminosity bins for which clustering measurements and jack-knife covariance matrices were available (see their Table 1).

The final product provided by Paul et al. (2019) comprises simple fitting functions for the $M_r$-dependence of the five parameters defining the thresholded HOD and the parameter $f_{\text{red}}$ in the range $-23 < M_r \leq -19$ (see their Figures

\(^{2}\) http://www.mpa-garching.mpg.de/gadget/
\(^{3}\) https://www.n-oca.eu/ohann/MUSIC/
\(^{4}\) https://bitbucket.org/gfcsanford/rockstar
\(^{5}\) http://hpc.iucaa.in
\(^{6}\) www.sdss.org
\(^{7}\) www.skyserver.sdss.org
\(^{8}\) https://github.com/nirinA/kcorrect_python
as the ‘best fit’, with the corresponding standard deviations across the bootstrap samples as errors.

Figure 1 compares the measured bivariate distributions in a few bins of \( M_r \) (coloured histograms) with the best fit Gaussian mixture (contours). Figure 2 summarises the 11 parameters of the mixture for all the luminosity bins (points with errors). The parameters are: (i) the probability \( p(\text{red}|M_r) \) that the galaxy belongs to the ‘red’ mode of the mixture, (ii) the mean vectors \( \langle g - r | M_r \rangle_{\text{red}}, \langle u - r | M_r \rangle_{\text{red}} \) and \( \langle g - r | M_r \rangle_{\text{blue}}, \langle u - r | M_r \rangle_{\text{blue}} \) of the red and blue modes, respectively, (iii) the diagonal elements of the covariance matrices \( \sigma_{\text{sat}}^2(g - r | M_r), \sigma_{\text{sat}}^2(u - r | M_r) \) and \( \sigma_{\text{rew}}^2(g - r | M_r), \sigma_{\text{rew}}^2(u - r | M_r) \) of the red and blue modes, respectively and (iv) the correlation coefficients between \( g - r \) and \( u - r \), for the red and blue modes respectively. The smooth curves in Figure 2 show piece-wise continuous polynomial fits to each parameter. These are used similarly to the HOD fitting functions provided by Paul et al. (2019), to assign colours to mock galaxies (see below).

For assigning colours to centrals and satellites separately, we require the probability \( p(\text{red}|\text{sat},M_r) \) that the \( g - r \) and \( u - r \) colours of a satellite of luminosity \( M_r \) are drawn from the ‘red mode’ of the bivariate double-Gaussian distribution described above. This can be easily obtained by combining the constraints on \( f_{\text{sat}} \) described in section 2.3.2 with (a subset of) the parameters of the Gaussian mixture \( p(g - r,u - r | M_r) \), and is given by

\[
p(\text{red}|\text{sat}, M_r) = \frac{2 f_{\text{sat}} - \mathcal{I}(\text{blue})(M_r)}{\mathcal{I}(\text{red})(M_r) - \mathcal{I}(\text{blue})(M_r)}
\] (2)

where (suppressing the \( M_r \)-dependence of the arguments for brevity)

\[
\mathcal{I}(\text{red}/\text{blue})(M_r) = \text{erfc}\left( \frac{(g - r)_{\text{cut}} - (g - r)_{\text{red/blue}}}{\sqrt{2} \sigma_{\text{red/blue}}(g - r)} \right),
\] (3)

and where \( (g - r)_{\text{cut}}(M_r) \) was given in equation (1).

2.3.4 Stellar masses

Stellar masses are calculated using a mass-to-light ratio calibrated to SDSS DR7 measurements similarly to Paranjape et al. (2015, see their sections 2.3 and 4.2). The measurements are shown as the coloured histogram in Figure 3. We fit a mean relation to these measurements in bins of \( x \equiv (g - r) \), of the form

\[
\langle M/L \rangle(x) = a + b \text{erf}((x - c) d) + e \tanh((x - f)/g),
\] (4)

finding best fitting values

\[
\begin{align*}
& a = 1.3281; \quad b = 0.735; \quad c = 0.5859; \\
& d = 3.38; \quad e = 0.187; \quad f = 0.8976; \quad g = 0.0874.
\end{align*}
\] (5)

We also fit the scatter around the mean in the same bins, of the form

\[
\sigma_{\langle M/L \rangle}(x) = \begin{cases} 
  k_0 + k_1 (x - x_0) & ; x < x_0 \\
  k_0 + k_2 (x - x_0) + k_3 (x - x_0)^2 & ; x \geq x_0
\end{cases}
\] (6)

finding best-fitting values given by

\[
\begin{align*}
& x_0 = 1.019; \quad k_0 = 0.1441; \quad k_1 = -0.182; \\
& k_2 = 1.16; \quad k_3 = -0.2.
\end{align*}
\] (7)
Figure 2. Summary of Gaussian mixture fits for $p(g-r,u-r|M_r)$ (points with errors). Smooth curves show piece-wise continuous polynomial fits to each mixture parameter, which we use in assigning colours to mock galaxies.

Our calibration is shown in Figure 3 and is better behaved than that of Paranjape et al. (2015) for very blue objects; we have checked, however, that both calibrations lead to nearly identical results for the final mocks. For each mock galaxy with colour $g-r$, we calculate a Gaussian random number for the mass-to-light ratio with mean and standard deviation calculated using equations (4) and (6), respectively, setting $x = (g-r)$. This is then combined with the corresponding value of $M_r$ of the galaxy to assign a stellar mass $m_\ast$.

2.4 Neutral hydrogen masses

Paul et al. (2018, hereafter, PCP18) calibrated a lognormal scaling relation (with constant scatter in log-mass) between the neutral hydrogen mass $m_{\text{HI}}$ of a galaxy and its optical properties $M_r$ and $g-r$, using the optical HOD from Guo et al. (2015) and clustering measurements from Guo et al. (2017) of HI-selected galaxies in the ALFALFA survey (Giovanelli et al. 2005). The scaling relation was separately calibrated for central and satellite galaxies using a halo model, along with an overall parameter $f_{\text{HI}}$ which gives the fraction of optically selected galaxies that contain HI. Their default model also assumed that satellites with $m_{\text{HI}} > 10^{10.2} \times (0.678/h) M_{\odot}$ do not exist. The model was constrained using measurements of number counts and projected clustering $w_p(r_p)$ of ALFALFA HI-selected galaxies from Guo et al. (2017) for the thresholds $\log_{10}(m_{\text{HI}}/h^{-2} M_{\odot}) \geq 9.8 + 2 \log_{10}(0.678)$ and $10.2 + 2 \log_{10}(0.678)$.

In the following, we will see that this default implementation of the PCP18 model is in mild disagreement with $w_p(r_p)$ for galaxies with $\log_{10}(m_{\text{HI}}/h^{-2} M_{\odot}) \geq 10.0 + 2 \log_{10}(0.678)$ (which was not used in constraining the scaling relation). This is likely due to our updated optical HOD from Paul et al. (2019) with its improved modelling of scale-dependent halo bias and the self-consistent calibration of the satellite red fraction. (The latter was identified by PCP18 as the parameter most susceptible to systematic effects in their analysis; see their section 4.2.) We therefore explored variations around the PCP18 model, finding that the simple modification of excluding satellites from halos with $m > m_{\text{sat,max}}$, while keeping the scaling relations for centrals and satellites intact otherwise, leads to acceptable descriptions of the measurements for all the thresholds mentioned above, as well as of the HI mass function. A simple $\chi^2$ minimisation exercise, varying $m_{\text{sat,max}}$ and predicting $w_p(r_p)$ for galaxies with $\log_{10}(m_{\text{HI}}/h^{-2} M_{\odot}) \geq 10.0 + 2 \log_{10}(0.678)$ using the
covariance matrix kindly provided by Hong Guo, leads to $n_\text{sat,max} \simeq 10^{14.4} h^{-1} M_\odot$.

Being tied to the optical HOD, this $\text{HI}$ scaling relation suffers from a natural incompleteness in producing $\text{HI}$ masses, determined by the optical completeness limit on $M_r$ for the galaxy population. For a sample limited by $M_r < -19$, for example, the $\text{HI}$ mass function is complete only above $m_{\text{HI}} \gtrsim 10^{6.5} h^{-2} M_\odot$. As such, all our analysis of $\text{HI}$-selected galaxies is restricted to the massive end.

3 ASSIGNING GALAXY PROPERTIES IN MOCKS

We now describe our main algorithm for generating mock galaxies in a gravity-only simulation, along with a ‘baryonisation’ scheme for assigning spatial distributions of stars and gas to each galaxy.

3.1 Algorithm

Our basic algorithm to assign galaxy luminosities and colours to halos in $N$-body simulations is essentially the same as that described by Skibba & Sheth (2009), with a few technical improvements. We also include a number of additional galaxy properties. The algorithm can be summarised as follows:

- **Central occupation and luminosity:** A threshold luminosity $L_{\text{min}}$ (or absolute magnitude $M_{\text{r, max}}$) is dynamically determined by requiring that the HOD of central galaxies $f_{\text{cen}}(> L_{\text{min}}(m))$ be sampled down to a value $1.5 \times 10^{-2}$ for all halo masses $m \geq 40 m_{\text{part}}$, where $m_{\text{part}}$ is the particle mass of the simulation. For example, for the L300_N1024 configuration, this gives us $M_{r, \text{max}} \approx -19$. Halos of mass $m$ are then occupied by a central galaxy with probability $f_{\text{cen}}(> L_{\text{min}}(m))$. Central galaxy luminosities are sampled from the conditional luminosity function $f_{\text{cen}}(> L(m)) / f_{\text{cen}}(> L_{\text{min}}(m))$.

- **Satellite occupation and luminosity:** The number of satellites in an occupied halo is drawn from a Poisson distribution with mean $N_{\text{sat}}(> L_{\text{min}}(m))$, and the luminosities of these satellites are sampled from the conditional luminosity function $N_{\text{sat}}(> L(m)) / N_{\text{sat}}(> L_{\text{min}}(m))$. We do not enforce that satellites be less luminous than their host central. For a luminosity-complete sample of galaxies with $M_r \leq -19$, we find that, in approximately 6.5% of groups containing at least one satellite, the brightest satellite is brighter than the central.

- **Galaxy positions:** The central of a halo is placed at the halo center-of-mass. Satellite positions are distributed as an NFW profile around the central, truncated at $R_{200c}$. Halo concentrations $c_{200c}$ are drawn from a Lognormal distribution with median and scatter at fixed halo mass as calibrated by Diemer & Kravtsov (2015) using very high-resolution simulations, which avoids contamination due to numerical fitting errors in relatively low-resolution boxes. The fitting functions from Diemer & Kravtsov (2015, who used the $m_{200c}$ definition of halo mass) are converted to the $m_{200b}$ definition appropriate for the HOD fits using the prescription of Hu & Kravtsov (2003).

To preserve correlations between halo concentration and large-scale environment, the Lognormal concentrations of halos in narrow mass bins are rank ordered by the actual estimated halo concentrations in each bin (the assumption being that this ranking would be approximately preserved even in relatively coarsely sampled halos, although see Ramakrishnan et al. 2020). The scale radius $r_s = R_{200b}/c_{200b}$, inferred for each halo from this exercise is stored for later use (see section 3.2).

- **Galaxy velocities:** The central of a halo is assigned the bulk velocity of the halo. The satellites are assigned random velocities drawn from a 3-dimensional isotropic Gaussian distribution with mean equal to the central velocity and 1-d velocity dispersion appropriate for the NFW profile at the location of the satellite.

Additionally, we implement the following modifications:

- **Galaxy colours:** To assign $g - r$ and $u - r$ colours to galaxies, we extend the algorithm proposed by Skibba & Sheth (2009). As in their case, the first step is to determine whether a galaxy is ‘red’ or ‘blue’, which is done separately for satellites and centrals, using the probability $p(\text{red}, sat, M_r)$ for satellites (see sections 2.3.3) and the corresponding probability for centrals which can be derived using $p(\text{red}, cen, M_r)$, $p(\text{red}, cen, M_r)$ and the HOD (this also requires an integral over the halo mass function, for which we use the fitting function from Tinker et al. 2008). In the next step, we assign $g - r$ and $u - r$ colours by sampling the appropriate (bivariate) mode of the double Gaussian $p(g - r, u - r | M_r)$ calibrated in section 2.3.3. Operationally, we first sample the univariate mode $p_{\text{red/blue}}(g - r | M_r)$ (obtained by marginalising the respective bivariate distribution over $u - r$) and then sample the corresponding conditional distribution $p_{\text{red/blue}}(u - r | g - r, M_r)$, for each red/blue galaxy. Optionally, we also include galactic conformity by correlating $g - r$ with halo concentration (or an unspecified Gaussian-distributed halo property) at fixed halo mass, using the tunable prescription of Paranjape et al. (2015).

- **Galaxy stellar masses:** As discussed by Paranjape et al. (2015) and described in detail in section 2.3.4, we calculate stellar masses $m_\star$ using a $(g - r)$-dependent mass-to-light ratio calibrated for SDSS DR7 galaxies.

- **Galaxy neutral hydrogen masses:** We assign $\text{HI}$ masses to a uniformly sampled fraction $f_{\text{HI}}$ of galaxies, with a lognormal distribution at fixed $M_r$ and $g - r$, using the model from PCP18. This model, which we refer to as the ‘minimal PCP18’ model below, additionally discards $\text{HI}$-satellites which have $\log_{10}(m_{\text{HI}}/h^{-2} M_\odot) > 10.2 + 2 \log_{10}(0.678)$. As described in section 2.4, we modify this model by instead discarding $\text{HI}$-satellites in parent halos having $m \geq m_{\text{sat,max}}$. This is done after the uniform downsampling for HI assignment described above. For our default model, which we call ‘PCP18 mod-sat’, we set $m_{\text{sat,max}} = 10^{14.4} h^{-1} M_\odot$ (see below for a comparison between the models). As discussed by PCP18, for optically selected galaxies with $M_r \leq -18$, this optical-$\text{HI}$ scaling relation places the majority of HI mass in faint blue galaxies (see their Figure 9).

- **Neutral hydrogen disks:** For each galaxy containing $\text{HI}$, we assign a comoving disk scale length $h_0$ for an assumed thin disk with HI surface density $\Sigma_{\text{HI}}(r_z) \propto e^{-r_z/r_{\text{HI}}}$ (here $r_{\text{HI}}$ is the radial distance in the disk plane), using the scaling relation

$$h_{\text{HI}} = 7.49 h^{-1}\text{kpc} \left(m_{\text{HI}}/10^{10} h^{-2} M_\odot\right)^{0.5},$$

10 All such ‘discarded’ satellites continue to have their assigned optical properties, only their HI mass is set to zero.
with a scatter of 0.06 dex, consistent with the measurements reported by Wang et al. (2016). These are useful in modelling ‘baryonified’ rotation curves, as we describe later.

- **Galaxy environment:** We assign to each galaxy the value of its host-centric dark matter overdensity \( \delta \) and tidal anisotropy \( \alpha \), each Gaussian smoothed at an adaptive scale \( 4R_{200b,\text{host}}/\sqrt{5} \) and at the fixed scale 2h\(^{-1}\)Mpc, as well as the bias \( b_1 \) of the host (see Paranapse et al. 2018a and Paranapse & Alam 2020 for how these variables are calculated in the N-body simulation).

### 3.2 ‘Baryonified’ density profiles and rotation curves

The galaxy properties described above, namely, luminosity, colour, stellar mass and Hi mass, are all assigned by our algorithm by treating each galaxy as a point object. Combined with the information on the host halo mass and concentration, however, these properties can also be used to model the spatial distribution of stars and of hot and cold gas in the galaxy and its halo. This in turn can be used to construct a rotation curve for the galaxy, which has several interesting applications as we discuss later.

Since the circular velocity \( v_{\text{circ}}(r) \) at a halo-centric distance \( r \) depends on the total mass \( m_{\text{total}}(r) \) enclosed in this radius, we must model the spatial distribution of all matter components inside the host halo. This is particularly relevant for the inner parts of the halo which are typically baryon-dominated. We restrict this analysis to central galaxies and will return in future work to satellite galaxies, which require additional modelling of processes such as tidal and ram pressure stripping, strangulation, etc. (see, e.g., van den Bosch et al. 2008; Behroozi et al. 2019) that are beyond the scope of the present work.

We follow the prescription of Schneider & Teyssier (2015, henceforth, ST15) to ‘baryonify’ each host halo. This method, and extensions thereof, have been shown to successfully account for baryonic effects in the matter power spectrum (Chisari et al. 2018; Schneider et al. 2019; Aricò et al. 2020b) and bispectrum (Aricò et al. 2020a) at relatively small scales over a range of redshifts. We have modified the ST15 prescription to include the Hi disk and have simplified it by truncating all profiles at the halo radius (see, e.g., Aricò et al. 2020b). Following the general practice for this method, we use the mass \( m_{\text{vir}} \equiv m_{200c} \) and radius \( r_{\text{vir}} \equiv R_{200c} \) for all baryonification scaling relations below. We are primarily interested here in low redshifts and length scales \( \lesssim R_{\text{vir}} \). We briefly summarise the method next.

- Before baryonification, each halo starts with its total matter as a single component distributed according to an NFW profile consistent with the gravity-only simulation in which the mock catalog is being generated, using the mass \( m_{\text{vir}} \) and concentration \( c_{\text{vir}} \equiv r_{\text{vir}}/r_s \) (see section 3.1 for details of determining the scale radius \( r_s \) for each halo).

- We divide the total mass of each baryonified halo into 5 components: bound gas (‘bgas’), stars in the central galaxy (‘cgal’), neutral hydrogen in its disk (‘Hi’), gas expelled due to feedback (‘egas’) and the dark matter which quasi-adiabatically relaxes in the presence of the baryons (‘rdm’).

Each baryonic component, denoted by index \( \alpha \in \{\text{bgas, cgal, Hi, egas}\} \), is assigned a mass fraction \( f_{\alpha} \) subject to the constraint \( \sum_{\alpha} f_{\alpha} = 1 \) due to conservation of baryonic mass. In practice, we set \( f_{\text{bgas}} = m_{\text{bgas}}/m_{\text{vir}}, \\ f_{\text{HI}} = m_{\text{HI}}/m_{\text{vir}} \) and \( f_{\text{gas}} \) using

\[
f_{\text{bgas}} = (\Omega_{\text{b}}/\Omega_{\text{m}}) \times \left[ 1 + (M_c/m_{\text{vir}})^{\beta} \right]^{-1},
\]

with \( M_c = 1.2 \times 10^{14} h^{-1} M_\odot \) and \( \beta = 0.6 \) as described by ST15 (see their equation 2.19 and figure 2). We comment on possible variations in this relation later. The ejected gas fraction \( f_{\text{gas}} \) is then set by the baryonic mass conservation constraint.

Each baryonic component is given its own mass profile \( \rho_{\alpha}(r) \). The choices below for \( \rho_{\text{bgas}}, \rho_{\text{cgal}} \) and \( \rho_{\text{gas}} \) are identical to those in ST15. We briefly describe these below and refer the reader to ST15 for more details and original references.

- The bound gas profile \( \rho_{\text{bgas}} \) has the form \( \rho_{\text{bgas}} \propto \left[ \ln(1 + r/r_s)/r_s \right]^{\alpha}, \) set assuming hydrostatic equilibrium and a polytropic equation of state in the inner halo and matched to the original NFW profile in the outer halo. Here \( r_s, \) and \( \Gamma \) are the scale radius of the NFW profile and the polytropic index of the gas, respectively. The matching is performed at a radius \( r_{\text{match}} = \sqrt{5} r_s \) and fixes the value of \( \Gamma. \) For hosts of centrals with \( M_c \lesssim 19 \), we find typical values of \( \Gamma \approx 1.19 \) with a dispersion of \( \approx 0.015. \)

- The stellar profile is assumed to follow \( \rho_{\text{cgal}} \propto r^{-2} e^{-r^2/4r_{\text{hl}}^2} \) with half-light radius \( R_{\text{hl}} = 0.015 R_{\text{vir}} \) (Kravtsov 2013). In principle, this can be extended to include a scatter and/or accommodate a dependence on halo angular momentum as predicted by disk formation models (Mo et al. 1998, see the discussion in Kravtsov 2013); we ignore this here for simplicity. Strictly speaking, we should treat the stellar profile as a combination of a central bulge and a 2-dimensional disk (with a relative contribution that correlates with galaxy colour), rather than the purely spherically symmetric form assumed here. In this work, we will follow the previous literature on the subject and assume the form given above, leaving a more self-consistent description of the stellar disk to future work.

In this sense, the stellar profile we model is better thought of as a pure bulge.

- The ejected gas profile is taken to be \( \rho_{\text{gas}}(r) \propto e^{-r^2/2r_e^2} \) with \( r_e = 0.5 \sqrt{200} h_0 R_{\text{vir}} \), setting \( h_0 = 0.5 \), so that \( r_e \approx 3.5 R_{\text{vir}}. \) As discussed by ST15, the modelling of \( \rho_{\text{gas}}(r) \) in the halo outskirts is rather uncertain and observationally ill-constrained. However, at the scales of our interest (\( r \lesssim R_{\text{vir}} \)), \( \rho_{\text{gas}}(r) \approx \) constant and its contribution to the total mass is therefore completely determined by baryonic mass conservation inside the halo. Our results are therefore expected to be very robust to any minor variations in the shape of \( \rho_{\text{gas}}(r) \) at scales \( \gtrsim R_{\text{vir}}. \)

- The sphericalised profile of Hi is obtained by integrating a thin exponential disk of surface density \( 2n_{\text{H}_1}(r) \propto e^{-r^2/b_{\text{H}_1}^2} \) (with \( r_{\text{H}_1} \) being the radial distance in the disk plane) to get \( \rho_{\text{Hi}}(r) \propto r^{-1} e^{-r/b_{\text{H}_1}}. \) The disk scale length

\(^{11}\) Wang et al. (2016) provide a scaling relation for the quantity \( D_{\text{HI}} \) defined as the diameter of the contour corresponding to a surface density of 1M\(_{\odot}\)pc\(^{-2}\), which we relate to \( h_0 \) using the provided scaling relation itself (which implies a constant average surface density, independent of \( m_{\text{HI}} \)) along with an integral over the exponential disk profile \( Z_{\text{HI}}(r). \)

\(^{12}\) We assume that the stellar and Hi disks are decoupled and do not model time-dependent warps, etc. in the Hi disk.
h_{\text{Hi}} is assigned as described in section 3.1. Note: This sphericalisation contribution only affects the calculation of the relaxed dark matter component. The contribution of the Hi disk to the rotation curve itself is treated separately as described below.

Each profile function $\rho_{b}(r)$ is normalised so as to enclose the total mass $m_{\text{vir}}$ inside $R_{\text{vir}}$. The total baryonic profile is then $f_{\text{bary}} \rho_{\text{bary}}(r) = \sum_{i} f_{i} \rho_{i}(r)$, with an enclosed baryonic mass $m_{\text{bary}}(<r) = 4\pi \int_{0}^{r} dr' r'^{2} f_{\text{bary}} \rho_{\text{bary}}(r')$.

- The dark matter component is assumed to quasi-adiabatically respond to the presence of baryonic mass and relax to a new shape while approximately conserving angular momentum. The details of the iterative procedure used to calculate the resulting relaxed dark matter profile $\rho_{\text{radm}}(r)$ (normalised similarly to the baryonic profile functions) are in Appendix A, which is based on section 2.3 of ST15. The mass fraction $f_{\text{relm}}$ is set simply by mass conservation to be $f_{\text{relm}} = 1 - \Omega_{b}/\Omega_{m}$. The dark matter mass enclosed in radius $r$ is then $m_{\text{radm}}(<r) = 4\pi \int_{0}^{r} dr' r'^{2} f_{\text{relm}} \rho_{\text{radm}}(r')$.

- The thin exponential Hi disk leads to a mid-plane circular velocity contribution $v_{\text{Hi}}(r)$ satisfying (see section 2.6 of Binney & Tremaine 1987)

\[ v_{\text{Hi}}(r) = \frac{2 \Omega_{\text{Hi}} V_{\text{vir}}}{(2\Omega_{\text{Hi}}/V_{\text{vir}})} y^{2} \left[ I_{0}(y) K_{0}(y) - I_{1}(y) K_{1}(y) \right] , \]

where $y \equiv r/(2\Omega_{\text{Hi}})$, $V_{\text{vir}} = \sqrt{Gm_{\text{vir}}/R_{\text{vir}}}$ is the virial velocity and $I_{\alpha}(y)$ and $K_{\alpha}(y)$ are modified Bessel functions of the first and second kind, respectively.

- The total mass of dark matter and all baryonic components except the Hi disk enclosed in radius $r$ is $m_{\text{tot-in}}(<r) = m_{\text{relm}}(<r) + m_{\text{bary}} - m_{\text{Hi}}(<r)$, where the notation ‘bary-Hi’ refers to summing over all baryonic components except Hi. Since we truncate all profiles at the radius $R_{\text{vir}}$ of the host halo, the total halo mass satisfies $m_{\text{vir}} = m_{\text{tot-in}}(<R_{\text{vir}}) + m_{\text{Hi}}$. Finally, the rotational velocity $v_{\text{rot}}(r)$ of a test particle at halo-centric distance $r$ in the galaxy mid-plane is

\[ v_{\text{rot}}(r) = \sqrt{Gm_{\text{tot-in}}(<r)/r + v_{\text{Hi}}^{2}(r)} . \]  

Figure 4 shows baryonified density profiles and rotation curves for hypothetical dwarf-like and Milky Way-like halos hosting an NGC 99-like galaxy, comparing the baryonified result with the original NFW result in each case. Only for these examples, we have set the galaxy stellar mass using the abundance matching prescription of Behroozi et al. (2013b) with updated parameters taken from Kravtsov et al. (2018), and have fixed the Hi mass to $m_{\text{Hi}} = 10^{9.83}\ h^{-2}M_{\odot}$.

We clearly see that the ejected gas profile is essentially constant in each case. More importantly, we see that the baryonified rotation curves are substantially flatter and also more diverse in shape than their purely NFW counterparts (see also section 5.1).
Figure 5. Sanity check on HOD. (Left panel:) Mocks versus input functions. For each threshold on $M_r$, we separately show the contribution of central and satellite galaxies in the mock (histograms) and in the analytical HOD (smooth curves). (Right panel:) Thresholded luminosity function in the mock (separately showing the contribution of central, satellite and all galaxies) versus SDSS data from Zehavi et al. (2011).

Figure 6. Sanity check on clustering. Projected 2-point correlation function (2pcf) $w_p(r_p)$ for red/blue/all galaxies in luminosity bins in the mock (solid lines with error bars) compared with data from Zehavi et al. (2011, points with errors). Mock measurements used 6 realisations of the L300_N1024 box for the three fainter bins and 3 realisations of the L600_N1024 box for the brightest bin. Lines show the mean and error bars reflect the respective standard deviations over all available realisations. Similarly to the data, mock galaxies were classified as red and blue based on their $g-r$ values in comparison to equation (1). The 2pcf for mock galaxies was calculated using equation (12) with $\pi_{max} = 60 \, h^{-1}\text{Mpc}$ to match the Zehavi et al. (2011) measurements.

4 RESULTS

We now report the results of generating mock catalogs using the algorithm of section 3 on the simulations described in section 2.1.

4.1 Optical properties

As a sanity check, the left panel of Figure 5 compares the input fitting functions for the HOD from Paul et al. (2019) with the output of the mock algorithm applied to a single L300_N1024 box. The right panel of the Figure compares the thresholded luminosity function averaged over 6 realisations of the L300_N1024 box with the measurements from Zehavi et al. (2011) that were used as constraints by Paul et al. (2019).

Figure 6 similarly compares the projected clustering of red/blue/all galaxies in mock catalogs with the measurements from Zehavi et al. (2011). Mock galaxies were classified as red and blue based on their $g-r$ values in comparison to equation (1) to ensure a fair comparison with the data. The projected 2-point correlation function (2pcf) $w_p(r_p)$ for mock galaxies was calculated by integrating the real space 2pcf $\xi(r)$ using

$$w_p(r_p) = 2 \int_{r_p}^{\infty} \frac{\pi_{max}}{r_p} dr \xi(r), \quad (12)$$

where we set $\pi_{max} = 60 \, h^{-1}\text{Mpc}$ to match the Zehavi et al. (2011) measurements. We see generally good agreement in all cases, although the clustering of red galaxies tends to be lower in the mock than in the data. This is very likely due to the limited volume of our 300$h^{-1}\text{Mpc}$ boxes, which do not include the effects of faint (predominantly red) satellites in very massive halos.

Figure 7 shows the joint distributions of $M_r$, $g-r$ and $u-r$ in one mock using the L300_N1024 box. We clearly see the well-known colour-magnitude and colour-colour bimodality, another sanity check on the mock algorithm.
Turning to somewhat more detailed tests, Figure 8 compares the differential luminosity and stellar mass functions averaged over 6 realisations of the L300_N1024 box with fitting functions to SDSS measurements from the literature. In each case, for the mock measurements we show results separately for red/blue/all central/satellite/all galaxies. The total luminosity function is in reasonable agreement with the Schechter function fit from Blanton et al. (2003), which is not very surprising since the thresholded luminosity function was used as a constraint in the HOD calibration. The total stellar mass function of the mocks also agrees reasonably well with the corresponding fit from Peng et al. (2012). In this case, we also have individual fits for centrals and satellites, which were produced by Peng et al. (2012) using the SDSS group catalog of Yang et al. (2007), which similarly agree well with the stellar mass functions of mock centrals and satellites, respectively. Considering the substantial amount of systematic uncertainty involved in extracting stellar mass functions from data, as well as inherent systematics in the galaxy classification algorithm used to produce the SDSS group catalog, we conclude that the mocks are in good agreement with the data here as well.

4.2 Neutral hydrogen properties

Figure 9 compares the H\textsubscript{i} mass function and projected 2pcf in the ‘minimal PCP18’ model (see section 3.1) with measurements in the ALFALFA survey. The left panel shows the differential H\textsubscript{i} mass function of red/blue/all central/satellite/all galaxies averaged over 6 realisations of the L300_N1024 box, compared with the fit to Hi-selected galaxies in the ALFALFA survey by Martin et al. (2010). We see reasonable agreement above the completeness limit of the catalog (set by the luminosity completeness threshold; see
Figure 9. (Left panel:) Differential H\textsubscript{i} mass function of red/blue/all central/satellite/all galaxies in averaged over 6 ‘minimal PCP18’ model mocks using the L300\_N1024 configuration. Colour segregation was based on equation (1) applied to $g - r$ and $M_r$ for each galaxy. For comparison, the solid green curves show the corresponding fit from Martin et al. (2010) to H\textsubscript{i}-selected galaxies in the ALFALFA survey. (Right panel:) Projected 2pcf in the ‘minimal PCP18’ model (averaged over the same mocks as in the left panel) for three H\textsubscript{i} thresholds, compared with corresponding ALFALFA measurements from Guo et al. (2017). The labels indicate the values of $\chi^2$ when comparing each mock result with the corresponding 12 data points from the ALFALFA measurements using the covariance matrices kindly provided by Hong Guo.

Figure 10. Same as figure 9, showing results for the ‘PCP18 mod-sat’ model described in section 3.1, which discards H\textsubscript{i}-selected satellites in parent halos with $m \geq m_{\text{sat,max}}$. The value of $m_{\text{sat,max}}$ was set to $10^{14.4}h^{-1}M_\odot$ by minimising the $\chi^2$ between the mock 2pcf results and corresponding ALFALFA measurements and covariance for the lowest mass threshold shown. We adopt this as our default model for assigning H\textsubscript{i} mass to mock galaxies.

As mentioned earlier, this slight disagreement is likely due to our use of an updated and improved optical HOD. We therefore explore the modification of the PCP18 model described in section 3.1 and discard H\textsubscript{i}-selected satellites in parent halos with $m \geq m_{\text{sat,max}}$. To set the value of the threshold, we attempted to minimise the $\chi^2$ between mocks and data for the threshold $\log_{10}(m_{\text{HI}}/h^{-2}M_\odot) \geq 10.0 + 2\log_{10}(0.678)$. We found that the $\chi^2$ has a very broad minimum in the vicinity of $m_{\text{sat,max}} = 10^{14.4}h^{-1}M_\odot$, which we use as our default value. The resulting H\textsubscript{i} mass function section 2.4 and the discussion in Paul et al. 2018).
and projected 2pcf are shown in Figure 10; we see a mild improvement in the 2pcf of the lowest threshold, and also some improvement in the higher thresholds. Since this halo thresholded model, which we refer to as ‘PCP18 mod-sat’, allows for the existence of massive satellites while still agreeing with observations, we choose to adopt it as our default model. For comparison, the number densities in units of \((h^{-1}\text{Mpc})^{-3}\) in this model for each thresholded sample (in order of increasing threshold \(\text{HI} mass\)) in the mocks are, respectively, 
\[2.27 \pm 0.003, 0.810 \pm 0.002, 0.203 \pm 0.0007 \times 10^{-3}\] 
(with errors estimated using the scatter across 6 realisations), while the corresponding values from Table 1 of Guo et al. (2017) are 
\[2.68, 0.92, 0.22 \times 10^{-3}\] 
(no errors are provided on these values).

We have also explored several ‘beyond halo mass’ modifications of the PCP18 model by changing the criterion used for discarding HI-selected satellites. In particular, we considered thresholds on (i) halo mass and concentration jointly (i.e., only allowing satellites in low mass and high concentration halos), (ii) halo concentration alone (only high concentration halos allowed), (iii) large-scale linear halo bias (low bias halos allowed) and (iv) discarding all ‘red mode’ satellites. These are generally inspired by the results of Guo et al. (2017), who found that abundance matching preferentially younger (sub)halos with HI-selected galaxies led to good descriptions of ALFALFA clustering. Of these, a joint threshold on halo mass and concentration performs the best, but leads to minimum \(\chi^2\) values nearly identical to those for the ‘PCP18 mod-sat’ model above, at the cost of one additional parameter. We therefore conclude that the ALFALFA data for massive HI galaxies do not require ‘beyond halo mass’ effects in modelling the mass function and projected 2pcf, provided that HI mass is assigned through an optical scaling relation.

5 PREDICTIONS AND EXTENSIONS

Having demonstrated that our mock catalogs reproduce the basic 1-point and 2-point observables associated with galaxy samples selected by optical or HI properties, in this section we discuss certain predictions of our mocks, along with a few possible extensions.

5.1 Rotation curves and baryon mass-halo mass relations

Figure 11 shows the rotation curves of 150 randomly chosen mock central galaxies containing HI disks, with the curves in each panel being coloured by one of \(m_\ast \), \(M_r \), \(m_{\text{HI}}\) or \(g-r\). As noted earlier, these are generally flat but show considerable diversity, in qualitative agreement with observed rotation curves (Persic et al. 1996; McGaugh et al. 2001). We defer a more quantitative comparison with observations to future work.

Our mocks also predict the relations between group halo mass and stellar mass \((m_\ast-m)\) as well as HI mass \((m_{\text{HI}}-m)\). Figure 12 shows the \(m_\ast-m\) relation (top panel) and the \(m_{\text{HI}}-m\) relation (bottom panel) for all central galaxies with \(M_r \leq -19\) in one mock using the L300_N1024 configuration. The top panel is essentially the same as Figure A4 of Paranjape et al. (2015), except that we have used an updated HOD. Similarly, the bottom panel can be compared with Figure B1 of PCP18, who showed the median \(m_{\text{HI}}-m\) relation using their analytical halo model. Both sets of results are consistent with these earlier works, showing a steep \(m_\ast-m\) relation but a much shallower \(m_{\text{HI}}-m\) relation. The latter feature also emphasizes the need for caution when painting HI directly into halos: the weak correlation between HI mass and halo mass can amplify systematic errors in any calibration.

For comparison, the purple curves in the top panel show the median \(m_\ast-m\) relations calibrated using SHAM by Behroozi et al. (2013b, solid) and Kravtsov et al. (2018, dash-dotted), converted in each case to the \(m_{\text{HI}}\) mass definition appropriate for this work. We see that, for stellar masses above the completeness limit of our mocks (horizontal dotted line), the mock result is closer to the Behroozi et al. (2013b) relation for \(m \lesssim 10^{13.5} h^{-1} M_\odot\) and lies between the two SHAM calibrations at larger halo masses. A similar comparison with the literature for the \(m_{\text{HI}}-m\) relation is complicated by the fact that different authors have used different conventions for defining this relation (e.g., Padmanabhan et al. 2017; Guo et al. 2017). We have checked that our results are qualitatively similar to these calibrations, leaving a more detailed analysis to future work.

5.2 HI-optical cross-correlations

A primary strength of our mock algorithm is its ability to paint realistic optical and HI properties in the same galaxies. This means that we can go beyond previous studies and predict or forecast expectations for the joint distribution of, say, stellar and HI mass in low-redshift galaxies, along with the corresponding spatial correlations.

Figure 13 shows the \(m_{\text{HI}}-m_\ast\) relation (left panel) and the spatial cross-correlation function between galaxy samples selected by luminosity and HI mass (right panel) in our mocks. These are genuine predictions of our algorithm; comparing these with corresponding measurements forms a test of the various underlying assumptions. Indeed, as already noted by PCP18, we see in the left panel that the predicted \(m_{\text{HI}}-m_\ast\) relation is in reasonable agreement with the results of Maddox et al. (2015, purple points with errors), although the median trend in the mocks is slightly lower than in the data.

To date, the only measurements of cross-correlations similar to those in the right panel are by Papastergis et al. (2013), who studied the projected cross-2pcf between HI-selected galaxies in ALFALFA and colour-selected galaxies in SDSS (see their Figures 17 and 18). As pointed out by Guo et al. (2017), however, the weights used by Papastergis et al. (2013) in their 2pcf measurements did not accurately account for sample variance effects, which are substantial in the small volume (\(z \lesssim 0.05\)) probed by the ALFALFA survey. E.g., Guo et al. (2017) reported a significant \(m_{\text{HI}}\)-dependence of clustering, which was not detected by Papastergis et al. (2013).

It will therefore be very interesting to confront our mock catalogs with more robust cross-2pcf measurements. For example, the choices controlling HI-satellites in our model (namely, the value of the threshold halo mass \(m_{\text{sat,max}}\)) affect the shape of the cross-correlation at small separations between bright optical galaxies and HI-selected galaxies and can therefore be tested by such observations.

MNRAS 000, 1–17 (0000)
Figure 11. Rotation curves of 150 central galaxies containing HI disks, chosen at random from one mock using the L300_N1024 configuration. Each curve is coloured by the galaxy’s stellar mass $m_*$ (top left), luminosity $M_r$ (bottom left), HI mass $m_{\text{HI}}$ (top right) and colour $g-r$ (bottom right). Since each rotation curve is only generated for $r \leq R_{\text{vir}}$, the curves truncate at the dotted line which shows $v(r) = (V_{\text{vir}}/R_{\text{vir}}) r$ in each panel.

Figure 12. Correlation with halo mass. Histograms show the $m_* - m$ (top panel) and $m_{\text{HI}} - m$ (bottom panel) relation for all central galaxies with $M_r \leq -19$ in one mock using the L300_N1024 configuration. Solid yellow lines in each panel show the median relation in bins of halo mass, while dashed yellow lines show the corresponding 16th and 84th percentiles. For calculating these curves, in the top panel, we ignore halos which do not contain a central galaxy and in the bottom panel, we further ignore halos whose central does not contain any HI mass. The solid purple curve in the top panel shows the SHAM calibration from Behroozi et al. (2013b), while the dash-dotted purple curve shows the same relation with parameters taken from Kravtsov et al. (2018), converted to the $m_{200b}$ mass definition in each case. The horizontal dotted lines in each panel indicate the approximate completeness thresholds for $m_*$ and $m_{\text{HI}}$ in the mock.

5.3 Possible extensions

Although the mocks we have presented here provide fairly realistic descriptions of the distribution of optical and HI properties of local Universe galaxies, they contain several ingredients which can be potentially improved upon or extended. We list some of these here.

5.3.1 Assembly bias

We noted in section 3.1 that our assignment of halo concentrations $c_{200b}$ preserves spatial correlations of $c_{200b}$ at fixed halo mass (also called assembly bias) under the assumption that these correlations are accurately tracked even by poorly resolved halos. Recently, Ramakrishnan et al. (2020) have shown that this assumption fails at worse than $\sim 20\%$ for halos resolved with $\lesssim 150$ particles. Instead, they showed that the spatial correlations of the local tidal anisotropy $\alpha$ are accurately preserved even for halos with as few as 30 particles. Using their technique of sampling a conditional distribution $p(c_{200b}|m,\alpha)$, therefore, will be a promising extension of our algorithm that would endow mocks built on low-resolution N-body simulations with accurate representations of halo assembly bias.

Using the fitting functions provided by Ramakrishnan et al. (2020), the same technique can be trivially extended, over a similar dynamic range in halo mass, to halo properties such as spin and asphericity, with the secondary bias of these variables also accurately accounted for. Halo spin, for example, can then be used to inform the assignment of the stellar spatial distribution (section 3.2), while halo asphericity could conceivably be used to model aspherical spatial distributions of satellites.

Another interesting extension, also easy to include in our mocks, would be an environment-dependent modulation of the HOD itself, as discussed by Xu et al. (2020). Combined with the conditional sampling of halo properties, this would lead to full flexibility in modelling galaxy assembly bias.
5.3.2 Stellar disk-bulge decomposition

We also noted above that our treatment of the stellar spatial profile is, strictly speaking, inconsistent because it assumes a spherically symmetric distribution of stars rather than, say, an axially symmetric disk. This has not been a concern in the literature on baryonic effects in the matter power spectrum (see section 3.2 for references), presumably because of the angular averaging inherent in calculating the latter. Since our primary intention is to produce a rotation curve for each galaxy, however, the difference between a spherical bulge and an axial disk can potentially be a large effect (see, e.g., Figure 2-17 of Binney & Tremaine 1987). We intend to explore this further in a forthcoming work, by simultaneously modelling a stellar disk and bulge using realistic bulge-to-disk mass ratios (e.g., Bernardi et al. 2014), along with their correlations with galaxy colours and the presence of an HI disk.

5.3.3 Gas fractions

In this work, we used the expression (9) for the bound gas fraction \( f_{\text{bgas}} \), with parameters adopted from ST15, for modelling all central galaxies with \( M_r \leq -19 \). Strictly speaking, this relation holds for the central galaxies of halos with \( m_{\text{vir}} \geq 10^{13} h^{-1} M_\odot \), since it is calibrated using X-ray observations of galaxy clusters. Our choice therefore corresponds to an extrapolation of this relation into an unobserved regime of halo mass.

Since the resulting value of \( f_{\text{bgas}} \) for each central is typically substantially smaller than its \( f_{\text{gas}} \) (which is set by baryonic mass conservation in this work), we do not expect this extrapolation to lead to any significant systematic error for any of the statistics explored in this paper. One possibility for improving this calibration would be to focus on modelling \( f_{\text{gas}} \) first, using observations of the circum-galactic medium, and then fix \( f_{\text{bgas}} \) using baryonic mass conservation.

5.3.4 Predictions at higher redshift

All of our results have been restricted to the local Universe, a consequence of using clustering constraints from the low-redshift \((z \lesssim 0.1)\) surveys SDSS and ALFALFA. There are a number of reasons to motivate extending our results to redshifts \(0.5 \lesssim z \lesssim 1\).

This redshift range is interesting from the point of view of both astrophysics and cosmology; it is not only close to the epoch of peak star formation activity in the Universe (see Madau & Dickinson 2014, for a review), but also encompasses the transition from matter domination to dark energy domination and the onset of cosmic acceleration. As a result, this redshift range is the target of several completed, ongoing and upcoming galaxy surveys such as GAMA (Driver et al. 2009), VIPERS (Guzzo et al. 2014), DES (Flaugher 2005), HSC (Aihara et al. 2018), DESI (DESI Collaboration et al. 2016), J-PAS (Benitez et al. 2015), Euclid (Laureijs et al. 2011), Rubin LSST (LSST Science Collaboration et al. 2009) and SKA (Blyth et al. 2015), that seek to both study the galaxy-dark matter connection and exploit it in extracting cosmological information.

From the practical viewpoint, redshifts \(z \geq 0.5\) should be sufficiently close to \(z = 0\) that a simple extension of our algorithm, ignoring the effects of halo and galaxy mergers, might be sufficient to robustly predict (central) galaxy properties. This would involve a ‘backwards’ evolution of multi-band galaxy luminosities and colours using tools such as K-CORRECT, along with some assumptions regarding the evolution of the optical HOD and the optical-HI scaling relation we use, possibly informed by simplified galaxy evolution models (e.g., Lilly et al. 2013). This approach would be...
complementary to SHAM prescriptions such as UniverseMachine (Behroozi et al. 2019) which rely on high-resolution merger trees that our approach, in principle, can do without.

If successful, such an exercise would allow us to address a range of issues. For optically selected samples, this includes the expected distribution of galaxies observed by surveys such as DES and Euclid, the interplay between scale-dependent galaxy bias and redshift space distortions for these galaxies, etc. For neutral hydrogen studies, such mocks could inform expectations from stacking analyses, predictions for intensity mapping experiments, etc. We leave a more detailed assessment of the feasibility of this technique to future work.

6 CONCLUSION

The ability to realistically reproduce, in a simulated universe, the properties and spatial distribution of galaxies observed in the actual Universe, opens the door to addressing a number of interesting astrophysical and cosmological questions. We have presented an updated algorithm that produces catalogs of mock galaxies in simulated halos at $z \approx 0$, realistically endowed with a variety of properties including $r$-band luminosities, $g - r$ and $u - r$ colours, stellar masses $m_*$, neutral hydrogen (HI) masses $m_{\text{HI}}$, as well as (for central galaxies) the spatial distribution of gas and stars, leading to realistic rotation curves. Our mock galaxies additionally inherit a number of environmental properties from their host dark matter halos, including the halo-centric overdensity, tidal anisotropy and large-scale halo bias.

Our algorithm, which relies on an HOD which assigns galaxy properties based on halo mass $m$ alone, can optionally include effects such as galactic conformity and colour-dependent galaxy assembly bias, and is easily extendable to include effects such as environment-dependent modulations of the HOD. By construction, the basic mocks we presented here reproduce the luminosity function, colour-luminosity relation and the luminosity- and colour-dependent 2-point clustering of optically selected SDSS galaxies with $M_r \leq -19$, as well as the HI mass function and HI-dependent 2-point clustering of HI-selected ALFALFA galaxies with $m_{\text{HI}} > 10^{9.7} h^{-2} M_\odot$. The mocks then reproduce the SDSS stellar mass function and the SDSS-ALFALFA $m_r - m_{\text{HI}}$ relation reasonably well (these were not used when constraining the parameters of the algorithm), while predicting the spatial 2-point cross-correlation function of low-redshift optical and HI galaxies (which has not yet been robustly measured; Figure 13), and the $m_* - m$ and $m_{\text{HI}} - m$ relations (Figure 12). The calibrations we used lead to volume-completeness thresholds of $M_r \leq -19$, $m_r > 10^{8.5} h^{-2} M_\odot$ and $m_{\text{HI}} > 10^{9.7} h^{-2} M_\odot$, thus representing the population of massive galaxies in the low-redshift Universe. Our algorithm represents a consolidation of the results of Paranjape et al. (2015), Paul et al. (2018) and Paul et al. (2019).

Our mocks are potentially useful for a number of applications, some of which we list here.

- In their study of environment-dependent clustering in SDSS, Paranjape et al. (2018b) noted some small (∼20%) but significant differences between their ‘mass-only HOD’ mocks and SDSS galaxies in the most anisotropic tidal environments. These differences were ultimately inconclusive due to the comparable level of systematic uncertainties associated with the HOD calibration used by Paranjape et al. (2018b).

The mocks we have presented, which are based on updated calibrations, largely mitigate many of these uncertainties. It will therefore be interesting to revisit this analysis to assess the level of beyond-mass effects induced by the tidal environment in the SDSS field (see also Alam et al. 2019).

- As we noted earlier, our algorithm successfully describes Hi-dependent clustering at the massive end in ALFALFA without the need of assembly bias, unlike earlier studies (e.g. Guo et al. 2017). Our mocks can therefore serve as useful null tests for galaxy assembly bias using interesting new combinations of observables, such as the large-scale bias of galaxies split by HI mass in bins of stellar mass.

- Galactic conformity, a putative non-local connection between the satellites and central galaxy of the same halo, continues to pose a puzzle for galaxy formation models (Weinmann et al. 2006; Hearin et al. 2016b). Our mocks can be used to explore new tests of this phenomenon, e.g., the potential dependence of galactic conformity on tidal environment (which is otherwise an excellent indicator of halo assembly bias, see Ramakrishnan et al. 2019).

- Our mocks can predict the HI mass function of (massive) galaxies selected by optical luminosity or colour. The corresponding measurements have only recently become available (Dutta et al. 2020; Dutta & Khandai 2021), and will be very useful for testing our basic assumptions regarding the connection between optical properties and HI.

- The rotation curves of our mock central galaxies, along with their HI disks when present, can be used to model the observed 21cm velocity profiles of HI-selected galaxies. The distribution of the widths of these profiles has been measured in the ALFALFA survey (Papastergis et al. 2011; Moorman et al. 2014) and constitutes an exciting and hitherto unexplored new probe of the small-scale distribution of baryonic matter.

- The ‘radial acceleration relation’ (RAR) between the acceleration profiles due to dark matter and baryons in disk galaxies (McGaugh et al. 2016; Lelli et al. 2017) has emerged as an intriguing new probe of gravitational theories at galactic length scales. The rotation curves and mass profiles of baryonic and dark matter in our mock central galaxies will allow us to explore the nature of the RAR for large samples of galaxies in the CDM+baryons framework.

We intend to pursue these applications in the near future and will report on the results in forthcoming publications.

ACKNOWLEDGMENTS

We thank R. Srianand for his collaboration and for many valuable discussions and suggestions which improved the presentation of this paper. AP also thanks Nishikanta Khandai and Kandaswamy Subramanian for insightful discussions. The research of AP is supported by the Associateship Scheme of ICTP, Trieste and the Ramanujan Fellowship awarded by the Department of Science and Technology, Government of India. TRC acknowledges support of the Department of Atomic Energy, Government of India, under project no. 12-R&D-TFR-5.02-0700 and the Associateship Scheme of ICTP. This work made extensive use of the open source computing packages NumPy (Van Der Walt et al. 2011),13 SciPy (Vir-
DATA AVAILABILITY

The mock catalogs generated by our algorithm will be shared upon reasonable request to the authors.

REFERENCES

Abadi M. G., Navarro J. F., Fardal M., Babul A., Steinmetz M., 2010, MNRAS, 407, 435
Abazajian K. N., et al., 2009, ApJS, 182, 543
Abbas S., Sheth R. K., 2007, MNRAS, 378, 641
Aihara H., et al., 2018, PASJ, 70, 54
Alici G., Angulo R. E., Hernández-Monteagudo C., Contreras S., Zennaro M., 2020a, arXiv e-prints, p. arXiv:2007.09004
Alici G., Angulo R. E., Hernández-Monteagudo C., Contreras S., Zennaro M., Pellejero-Ibañez M., Rosas-Guevara Y., 2020b, MNRAS, 495, 4806
Bagla J. S., Khandai N., Datta K. K., 2010, MNRAS, 407, 435
Barnes J., White S. D. M., 1984, MNRAS, 211, 753
Benítez N., et al., 2015, in Highlights of Spanish Astrophysics VIII, pp 148–153
Bernardi M., Meert A., Vikram M., Huertas-Company M., Mei S., Shankar F., Sheth R. K., 2014, MNRAS, 443, 874
Bett P., Eke V., Frenk C. S., Jenkins A., Helly J., Navarro J., 2007, MNRAS, 376, 215
Binney J., Tremaine S., 1987, Galactic dynamics. Princeton University Press, Princeton, NJ
Blanton M. R., et al., 2003, ApJ, 592, 819
Blumenthal G. R., Faber S. M., Flores R., Primack J. R., 1986, ApJ, 301, 27
Blyth S., et al., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14), p. 128 (arXiv:1501.01295)
Castorina E., Villaescusa-Navarro F., 2017, MNRAS, 471, 1788
Chisari N. E., et al., 2018, MNRAS, 480, 3962
Contreras S., Angulo R., Zennaro M., 2020, arXiv e-prints, p. arXiv:2012.06596
Cooray A., Sheth R., 2002, Phys. Rep., 372, 1
DESI Collaboration et al., 2016, arXiv e-prints, p. arXiv:1611.00086
Dempster A. P., Laird N. M., Rubin D. B., 1977, Journal of the Royal Statistical Society. Series B (Methodological), 39, 1
Driver S. P., et al., 2009, Astronomy and Geophysics, 50, 5.12
Dutta S., Khandai N., 2021, MNRAS, 500, L37
Dutta S., Khandai N., Dey B., 2020, MNRAS, 494, 2664
Duffy A., et al., 2008, International Journal of Modern Physics A, 20, 3121
14 http://www.scipy.org
15 https://matplotlib.org/
16 https://jupyter.org
17 https://veusz.github.io
18 http://hpc.iucaa.in
Here we describe the technique for calculating the response of the dark matter profile to presence of baryonic matter through an approximate conservation of angular momentum (Barnes & White 1984; Blumenthal et al. 1986; Gnedin et al. 2004; Abadi et al. 2010; Teyssier et al. 2011; ST15). The discussion below follows section 2.3 of ST15 (see their equations 2.15-2.17).

The basic equation describing this quasi-adiabatic relaxation gives the final radius \( r \) of a spherical dark matter element in terms of its initial radius \( r_{in} \),

\[
\frac{r}{r_{in}} = 1 + q_{dm} \left( \frac{m_{infw}(<r_{in})}{m_{tot}(<r')} \right) - 1 .
\]  

(A1)

Here, \( m_{infw}(<r) = m_{bary}(<r) + m_{dm}(<r) \) is the total mass contained inside the final radius \( r \), with \( m_{bary}(<r) = 4\pi \int_0^r dr' r'^2 f_{bary} \rho_{bary}(r') \) being the baryonic component and \( m_{dm}(<r) \) being the final, relaxed dark matter mass profile which satisfies

\[
m_{tot}(<r) = \int_{r=0}^{r_{in}} dr' m_{infw}(<r_{in}) ,
\]

(A2)

and \( m_{infw}(<r_{in}) = 4\pi \int_0^{r_{in}} dr' r'^2 \rho_{infw}(r') \) is the dark matter mass inside the initial radius as per the original, normalised NFW profile. We remind the reader that all the density profiles are normalised so as to enclose the entire mass \( m_{infw} \) inside \( r = R_{vir} \).

The quantity \( q_{dm} \) is a parameter controlling the level of angular momentum conservation. From equation (A1), we see that \( q_{dm} = 1 \) corresponds to perfect conservation, since \( r m(<r) \) is an adiabatic invariant in this case. On the other hand, \( q_{dm} = 0 \) corresponds to no baryonic backreaction. In this work, we follow ST15 and set \( q_{dm} = 0.68 \), which has been found to accurately describe the cumulative effects of baryonic backreaction effects both in the inner and outer regions of simulated halos, accounting for the fact that the formation of the central galaxy is not instantaneous. The effect of varying \( q_{dm} \) on rotation curves and related statistics will be the focus of a future study.

Defining the ratio

\[
\xi \equiv r/r_{in} ,
\]

(A3)
equation (A1) can be re-written as

\[
\mathcal{L}(\xi|\nu) \equiv \xi - 1 + q_{dm} \int_{r_{in}}^{r_{tot}} \left[ 1 + \frac{m_{bary}(<r)}{f_{infw}m_{infw}(<r/\xi)} \right]^{-1} - 1 .
\]

(A4)

which is conducive to an iterative solution. We employ Newton’s method using an analytical expression for the derivative

\[
\mathcal{L}(\xi|\nu) = \partial \mathcal{L}(\xi|\nu)/\partial \nu \text{ at fixed } \nu \text{ and writing the estimate at the } n^{th} \text{ iteration as}
\]

\[
\xi^{(n)} = \xi^{(n-1)} - \mathcal{L}(\xi^{(n-1)})/\partial \mathcal{L}(\xi^{(n-1)}) .
\]

(A5)

For practically all baryonic configurations and values of \( r \in (10^{-3}, 1) \times R_{vir} \), and for all values \( 0 \leq q_{dm} \leq 1 \), convergence is achieved with a relative tolerance of \( 10^{-5} \) in \( \leq 8 \) iterations using equation (A5). (In contrast, a simple iteration applied directly to equation A4 typically requires several tens to hundreds of iterations for \( q_{dm} \geq 0.9 \).)

Knowing the ratio \( \xi \) at any \( r \) then gives the mass of relaxed dark matter enclosed in radius \( r \) using equation (A2) setting \( r_{in} = r/\xi \) on the right hand side. By construction, \( \xi = 1 \) at \( r = R_{vir} \), so that \( m_{tot}(<R_{vir}) = f_{infw}m_{infw} \) as it should be. The value of \( m_{tot}(<r) \) at any \( r \) is sufficient for calculating the rotation curve \( \nu_{rot}(r) \) using equation (11). For the differential profile shown in Figure 4, we must differentiate equation (A2) with respect to \( r \). The (somewhat cumbersome) result can be written analytically entirely in terms of \( \xi \) and \( r \); we omit it for brevity.

APPENDIX A: QUASI-ADIABATIC RELAXATION

Here we describe the technique for calculating the response of the dark matter profile to presence of baryonic matter through an approximate conservation of angular momentum.

\[
\frac{r}{r_{in}} = 1 + q_{dm} \left( \frac{m_{infw}(<r_{in})}{m_{tot}(<r')} \right) - 1 .
\]

(A1)