Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **505,** 325–338 (2021) doi:10.1093/mnras/stab1286
Advance Access publication 2021 May 6

Rosella: a mock catalogue from the P-Millennium simulation

Sasha Safonova,[1,2]★† Peder Norberg[2,3,4] and Shaun Cole[2]

[1]*Department of Astronomy, Yale University, New Haven, CT 06520, USA*
[2]*Institute for Computational Cosmology, Department of Physics, University of Durham, South Road, Durham DH1 3LE, UK*
[3]*Centre for Extragalactic Astronomy, Department of Physics, University of Durham, South Road, Durham DH1 3LE, UK*
[4]*Institute for Data Science, Department of Physics, University of Durham, South Road, Durham DH1 3LE, UK*

Accepted 2021 April 27. Received 2021 April 22; in original form 2020 August 26

ABSTRACT

The scientific exploitation of the Dark Energy Spectroscopic Instrument Bright Galaxy Survey (DESI BGS) data requires the construction of mocks with galaxy population properties closely mimicking those of the actual DESI BGS targets. We create a high-fidelity mock galaxy catalogue, including information about galaxies and their host dark matter subhaloes. The mock catalogue uses subhalo abundance matching with scatter to populate the P-Millennium *N*-body simulation with galaxies at the median BGS redshift of ∼0.2, using formation redshift information to assign $^{0.1}(g-r)$ rest-frame colours. The mock provides information about *r*-band absolute magnitudes, $^{0.1}(g-r)$ rest-frame colours, and 3D positions and velocities of a complete sample of DESI BGS galaxies in a volume of $(542\ \mathrm{Mpc}\ h^{-1})^3$, as well as the masses of host dark matter haloes. This P-Millennium DESI BGS mock catalogue is ideally suited for the tuning of approximate mocks unable to resolve subhaloes that DESI BGS galaxies reside in, to test for systematics in analysis pipelines and to interpret (non-cosmological focused) DESI BGS analysis.

Key words: methods: analytical – galaxies: abundances – galaxies: haloes – dark energy – dark matter – large-scale structure of Universe.

1 INTRODUCTION

Upcoming cosmological surveys, such as the Dark Energy Spectroscopic Instrument (DESI) survey[1] (DESI Collaboration 2016, 2018), *Euclid*[2] (Laureijs et al. 2011), Large Synoptic Survey Telescope[3] (LSST; Ivezić et al. 2019), the Subaru Prime Focus Spectrograph (PFS),[4] and Wide-Field Infrared Survey Telescope (WFIRST),[5] aim to map the cosmic structures with the goal of measuring the structures' growth, distribution, and the expansion history of the Universe. Cosmological surveys enable measurements of galaxy clustering, redshift-space distortions, and weak lensing, among other qualities of the Universe. These measurements can constrain theories behind cosmic acceleration (Efstathiou, Sutherland & Maddox 1990; Riess et al. 1998; Perlmutter et al. 1999), test general relativity, and give us greater insight into the nature of dark matter.

The often used way of extracting information from such surveys is to compare summary statistics between observed data and mock data generated from theoretical predictions (e.g. DeRose et al. 2019; Smith et al. 2020). In order to compare theoretical predictions to observed quantities, we must create a medium that renders both sides of scientific endeavour – theory and experiment – directly comparable. In the context of cosmology and the large-scale structure of the Universe, that medium is a mock catalogue. Such a catalogue serves as a container of data about the quantities we could feasibly observe with cosmological surveys. These quantities might include the masses of galaxies or their brightnesses (in single or multiple bands), galaxy positions, velocities, redshifts, spectra, object type, and more.

To be a useful connector of theory to observations, mock data must provide quantities that resemble the observations against which it will be compared. The quantities should satisfy two major requirements. First, the mock quantities must be statistically equivalent to real quantities on the level of individual objects. This can be achieved by, for instance, connecting theoretical predictions with empirical measurements from past surveys.

The mock data's large-scale structures, as well as its summary statistics, should closely resemble what we observe in the local Universe. Were our simulations and mock data produced from a model that perfectly represented the Universe, the mock data we create from simulations should be indistinguishable from observed data if we examined both side by side. This level of statistical resemblance enables cosmologists to make comparisons between theory and observations at high levels of accuracy.

Mock catalogues can be used to develop and test the analysis tools intended for completed and upcoming surveys because a mock's cosmology is known a priori. The value of a number of parameters of interest can be measured directly in a mock, without the assumptions that are necessary in analyses of real data. Cosmological surveys also require mocks for testing observational strategies and quantifying biases (e.g. Smith et al. 2017).

★ E-mail: sasha.safonova@yale.edu
† NSF Fellow.
[1]https://www.desi.lbl.gov/
[2]https://www.euclid-ec.org/
[3]https://www.lsst.org/
[4]https://pfs.ipmu.jp/index.html
[5]https://wfirst.gsfc.nasa.gov/index.html

Modern cosmological surveys, such as extended Baryon Oscillation Spectroscopic Survey (Dawson et al. 2013; Blanton et al. 2017), DESI (DESI Collaboration 2016), and LSST (Ivezić et al. 2019), require simulations that cover volumes that exceed 100 $(\mathrm{Gpc}\,h^{-1})^3$ in a multitude of realizations. Such great volumes are motivated by a combination of the scientific questions that the surveys attempt to tackle, as well as the systematics that accompany real-world observations.

For instance, for the analysis of systematics for measurements of baryon acoustic oscillations (BAOs), volumes of the order of $\sim$200 $(\mathrm{Gpc}\,h^{-1})^3$ are necessary (DESI Collaboration 2018). The simulations tailored for such measurements should cover volumes that are at least 10 times greater than the volumes required to carry out the necessary measurements in order to limit the level of theoretical systematics (DESI Collaboration 2018).

Ideally, these simulations would solve equations of the physics of baryons and dark matter across cosmic time. Hydrodynamical simulations that account for the intricate physics that drives the formation of galaxies, however, are computationally expensive. The cost of simulating detailed physics that accounts for baryons in a volume that cosmological surveys require renders such simulations infeasible. Currently available hydrodynamical simulations, e.g. EAGLE (Crain et al. 2015), IllustrisTNG (presented in Naiman et al. 2018; Nelson et al. 2018; Pillepich et al. 2018; and others), and Massive Black II (Khandai et al. 2015), cover volumes that are much smaller than what is required for cosmological surveys' needs.

While insufficient in volume, hydrodynamical simulations offer the potential for direct simulation of physical details behind galaxy formation and evolution. This property makes this class of simulations useful for informing the methods that produce realistic galaxy populations more quickly and at lower computational cost.

One way to circumvent the computational expense of running a full hydrodynamical cosmological simulation is to consider a dark matter-only *N*-body simulation, in which the equations of gravity only are solved, substantially bringing down computational costs. The simulation is then 'populated' with galaxies following some algorithm, resulting in a catalogue of galaxies with properties and distribution that should be expected in a universe like the one that the *N*-body simulation represents. Methods for populating *N*-body simulations with galaxies are able to produce the cosmological-scale mock data that modern surveys require.

These methods can be broadly classified as physical, statistical, and statistical empirical. The physical approach encompasses semi-analytical models (SAMs; e.g. White & Frenk 1991; Kauffmann, White & Guiderdoni 1993; Cole et al. 1994, 2000; Somerville & Primack 1999; Baugh 2006; Gonzalez-Perez et al. 2014; Croton et al. 2016; Lacey et al. 2016; Baugh et al. 2019). Statistical methods include biased dark matter (e.g. Cole et al. 1998; White, Tinker & McBride 2014), halo occupation distributions (HOD; e.g. Benson et al. 2000; Peacock & Smith 2000; Berlind & Weinberg 2002; Berlind et al. 2003), and conditional LFs (e.g. Yang, Mo & van den Bosch 2003, 2008; Cooray 2006). Statistical-empirical approaches include subhalo abundance matching (SHAM; e.g. Kravtsov et al. 2004; Vale & Ostriker 2004; Conroy, Wechsler & Kravtsov 2006) and its modifications (e.g. Skibba & Sheth 2009; Guo et al. 2016).

SHAM is a method of populating dark matter subhaloes with galaxies by matching the cumulative abundance functions of a dark matter halo property (commonly, subhalo dark matter circular velocity or mass) to the luminosity function (LF) or a similar cumulative distribution function (cdf) of a galactic property. A variety of works have proposed that circular velocity, v_{circ}, measured at various times in a subhalo's lifetime, may be an appropriate connector of host subhaloes to galaxies (e.g. Conroy et al. 2006; Masaki et al. 2013a; Reddick et al. 2013; Chaves-Montero et al. 2016).

A number of approaches adding scatter to a SHAM mock have been proposed, such as sampling a probability distribution (Chaves-Montero et al. 2016; Guo et al. 2016), fitting a parametrized model to a hydrodynamical simulation and sampling the resulting likelihood (Chaves-Montero et al. 2016), adding scatter to SHAM-style assignment of galaxy colours (Masaki et al. 2013a; Yamamoto, Masaki & Hikage 2015), deconvolution (Reddick et al. 2013), and shuffling with a fixed scattering magnitude, used in McCullagh et al. (2017), as well as the method described in this work.

SHAM offers the advantage of using a cosmological model's predictive power for the number and properties of subhaloes, as well as their relation to their host haloes while requiring few, if any, parameters (Reddick et al. 2013). Cosmological simulations that resolve subhaloes alleviate the need for assumptions about the occupation number and distribution of halo substructures, which are necessary for statistical models, such as HODs. Implementations of SHAM have been shown to reproduce observed quantities that include the two-point correlation function (e.g. Conroy et al. 2006; Reddick et al. 2013; Lehmann et al. 2017), three-point statistics (e.g. Tasitsiomi et al. 2004; Marín et al. 2008), galaxy–galaxy lensing (e.g. Tasitsiomi et al. 2004), and the Tully–Fisher relation (e.g. Desmond & Wechsler 2015).

The ultimate goal of this research is to produce a mock galaxy catalogue that closely mimics data that will be observed in DESI's Bright Galaxy Survey (DESI Collaboration 2016). The Rosella mock catalogue described here uses SHAM to populate the P-Millennium *N*-body simulation (described in Section 2.1) with galaxies. Our approach provides rest-frame *r*-band absolute magnitudes and ${}^{0.1}(g-r)$[6] colours assigned with algorithms described in Sections 2.2.2 and 2.2.4, as well as positions, velocities, and host dark matter subhalo masses from P-Millennium. This work creates absolute magnitudes and colours *k*-corrected to $z \sim 0.1$ because this is the redshift used in papers that form the basis of our mock, for instance, Zehavi et al. (2011) and Smith et al. (2017).

Our method for galaxy colour and luminosity assignment offers novel developments, namely the magnitude depth, volume, scatter, and the inclusion of subhalo history information. Rosella's method of including scatter in the luminosity data uniquely conserves the target LF, which enables the assignment of galaxy luminosities as faint as $M_r^h \sim -17.5$. We discuss this property in Section 2.2.3.

While we focused the detailed tuning of the mock presented in this paper on the needs of the DESI Bright Galaxy Survey (BGS), the mock can be used for other low-redshift galaxy surveys that might benefit from a $z \sim 0.2$ reference mock (e.g. the WAVES[7] survey in 4MOST). Furthermore, the method behind Rosella can be used to create galaxy mocks at other redshifts and, with some additional steps, extended into a light-cone mock. The method can thus benefit any survey that probes volumes similar to those covered by Rosella (see Section 2.1 for details).

We have chosen to create this implementation of Rosella at the simulation snapshot that corresponds to a redshift of 0.203. The choice is motivated by the needs of the BGS. BGS will take the spectra of relatively bright galaxies during bright observing time. Consequently, its selection of target galaxies places the median redshift for future BGS observations at $z \sim 0.2$. Rosella will be

[6]We denote absolute magnitudes and colours *k*-corrected to redshift 0.1 with the superscript 0.1.
[7]https://wavesurvey.org/

useful as a reference mock for BGS, for fulfilling tasks that include analysing survey biases and calibrating approximate mocks that meet the volume and abundance requirements of the experiment (DESI Collaboration 2018).

We evaluate the closeness of the match between our mock and real data by comparing the luminosity- and colour-dependent clustering of our mock's galaxies against previously published clustering of similar galaxy populations in existing observational and mock data.

This paper is organized as follows. Section 2 describes the *N*-body simulation that Rosella is built on and outlines the methodology behind our work. Section 3 described the properties of the Rosella mock, including the LF, the luminosity- and colour-dependent clustering of the galaxies in Rosella, and the colour bimodality of galaxies in Rosella. Section 4 presents our main conclusions.

Throughout this work, *r*-band absolute magnitudes and $(g - r)$ colours are given in *AB* magnitudes, as defined for the *Sloan Digital Sky Survey* (*SDSS*) system (e.g. Blanton et al. 2003).

2 SHAM WITH P-MILLENNIUM FOR THE DESI BRIGHT GALAXY SURVEY

A mock catalogue tailored for the needs of BGS already exists: it is a light-cone mock constructed with an application of HOD to the Millennium-XXL (MXXL) simulation (Smith et al. 2017). However, that mock catalogue has some limitations. The catalogue described in this paper can address these limitations. The simulation we use here, P-Millennium, offers high-mass resolution that enables the tracking of fainter galaxies and the creation of a mock catalogue tailored with the scientific requirements of DESI's BGS in mind.

2.1 Simulation: P-Millennium

The Planck Millennium *N*-body simulation (hereafter P-Millennium) is a high-resolution dark matter-only simulation of a 800 Mpc periodic box (Baugh et al. 2019). It is part of the 'Millennium' series (Springel et al. 2005; Boylan-Kolchin et al. 2009) of dark matter-only simulations of large-scale structure formation in cosmologically representative volumes carried out by the Virgo Consortium.[8]

P-Millennium is run using cosmological parameters given by the best-fitting Lambda cold dark matter (ΛCDM) model to the first-year Planck cosmic microwave background data and measurements of large-scale structure in the spatial distribution of galaxies (Planck Collaboration XVI 2014). The analysis of the final Planck data set has introduced little change to these cosmological parameters (Planck Collaboration X 2020). See Table 1 for a summary of the specifications of the P-Millennium run.

The mass resolution of P-Millennium is 1.06×10^8 $M_\odot\, h^{-1}$ per particle, with 5040^3 particles representing the matter distribution (for a detailed comparison to other simulations in the Millennium suite, see Baugh et al. 2019). The lowest resolved halo mass in P-Millennium is $2.12 \times 10^9\, M_\odot\, h^{-1}$. This makes the simulation appropriate for SHAM, since the simulation's mass resolution lets SUBFIND (Springel et al. 2001) resolve dark matter halo substructures, subhaloes – a central component for creating a mock using SHAM (see Section 2.2 for a discussion).

The low-halo mass limit in P-Millennium allows us to create a mock with a faint absolute magnitude limit that reaches beyond the minimum luminosity cutoffs offered in other mock catalogues. For example, the *Buzzard* catalogue, presented in DeRose et al. (2019),

[8]http://virgo.dur.ac.uk/

Table 1. Selected cosmological parameters of the P-Millennium simulation: (1) Ω_M, present-day matter density in units of the critical energy density of the Universe; (2) Ω_b, the baryon density parameter; (3) Ω_Λ, the energy density parameter of the cosmological constant, Λ; (4) n_{spec}, the spectral index of the primordial density fluctuations; (5) h, the reduced Hubble parameter, $h = H_0/(100\ \mathrm{km\ s^{-1}\ Mpc^{-1}})$; (6) σ_8, the normalization of the density fluctuations at the present day; (7) N_p, the number of particles; (8) L_{box}, the simulation box length; (9) M_p, the mass of individual particles in the simulation; and (10) M_h, the minimum mass of a resolved halo, corresponding to 20 particles. See Baugh et al. (2019) for further details.

Parameter name	Value in P-Millennium
Ω_M	0.307
Ω_b	0.0483
Ω_Λ	0.693
n_{spec}	0.9611
h	0.6777
σ_8	0.8288
N_p	5040^3
L_{box} $(h^{-1}$ Mpc)	542.16
M_p $(h^{-1}\ M_\odot)$	1.06×10^8
M_h $(h^{-1}\ M_\odot)$	2.12×10^9

creates a reference mock that models the galaxy distribution down to roughly[9] $M_r^h = -18.2$, saying that 'the SHAM catalogue is not strictly complete' down to that absolute magnitude. As discussed in Section 3.1, Rosella can be fully complete for galaxies as faint as $M_r^h = -17.5$, depending on the choice of scatter that is implemented.

2.1.1 *Tracing P-Millennium subhalo histories*

v_{peak} is the central quantity that allows us to connect dark matter subhaloes in our *N*-body simulation to the galaxies in our mock catalogue. Its definition is built on the quantity v_{circ}, defined as

$$v\mathrm{circ}(r, z) = \sqrt{\frac{\mathrm{GM}(z, < r)}{r}} \tag{1}$$

r here is the physical distance between the particle and the centre of the subhalo, z is redshift, G is the gravitational constant, and $M(z, <r)$ is the mass enclosed within the radius r, at redshift z. Maximum circular velocity v_{max} is the value of v_{circ} at the radius at which v_{circ} reaches its maximum:

$$v\mathrm{max}(z) = \max[v\mathrm{circ}(r, z)] \tag{2}$$

v_{peak} is the highest v_{max} that a subhalo reaches over the course of its existence in the simulation:

$$v_{\mathrm{peak}} = \max[v\mathrm{max}(z)]. \tag{3}$$

To calculate v_{peak}, as well as a proxy for a subhalo's age, z_{form}, which we describe in Section 2.1.2, we compile the histories of v_{max} values that individual P-Millennium subhaloes reach over the course of the simulation. This non-trivial operation is described and discussed in greater detail in Safonova (2019).[10] Examples of such histories are plotted in Fig. 1. We generated a full dictionary of subhalo histories for subhaloes found at the P-Millennium snapshot

[9]We define *r*-band absolute magnitude dependent on h and with boundaries defined at $z \sim 0.1$ as $M_r^h \equiv {}_{0.1}M_r - 5 \log$ h, where h is the dimensionless constant given as $H_0 = 100\, h$ km s^{-1} Mpc^{-1}.

[10]The code for this procedure, along with the code used to complete the rest of the Rosella methodology, is stored in a private repository at https://github.com/safonova/pmillennium-sham.

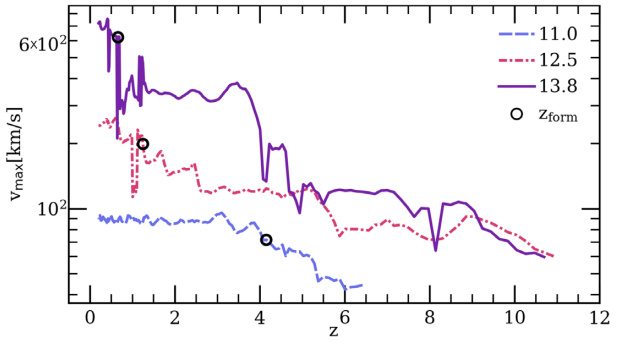

Figure 1. Examples of maximum circular velocities as a function of redshift. The vertical axis shows the v_{max} values of individual subhaloes at given redshifts z. Each line tracks the v_{max} history of a subhalo that has mass M at $z \sim 0.2$, expressed as $\log_{10}[M/(M_\odot\, h^{-1})]$, indicated in the legend. Black circles indicate z_{form} values for these subhaloes, calculated using the method described in Section 2.1.2.

corresponding to $z = 0.203$. The histories show transitory features that appear like short-lived drops in v_{max}, perhaps related to mergers and the difficulties of tracking subhaloes during mergers (e.g. Behroozi et al. 2012). In Section 2.1.2, our definition of z_{form} makes these features inconsequential.

2.1.2 *Definition of formation redshift*

In order to assign colours to Rosella galaxies, we compute each subhalo's 'formation redshift', z_{form}, which serves as a proxy for a subhalo's age. The choice to connect galaxy colours to the ages of their host subhaloes stems from the idea that older subhaloes are likely to have older and, consequently, redder stellar populations (e.g. Mo, van den Bosch & White 2010; Hearin 2015). We compute an individual subhalo's z_{form} based on the criterion that z_{form} corresponds to the maximum output redshift at which a subhalo's v_{max} exceeds v_{form}:

$$v_{form} = f\, v_{peak}. \tag{4}$$

Here, f is a free parameter. We identify the two output redshifts between which v_{form} is located and interpolate between them to get z_{form}. If the pre-v_{peak} history of the subhalo consists only of v_{max} values greater than v_{form}, z_{form} is set to the redshift corresponding to the earliest snapshot at which the subhalo is found.

It is possible to adjust the f parameter, or even the relationship between v_{peak} and v_{form}, to tune the mock data produced with the model presented here. We have considered two values of f, 0.75 and 0.9. We have noted that $f = 0.75$ produces a more favourable match to clustering data (see Section 3.4.2). Expression is inspired by the works of Masaki, Lin & Yoshida (2013b) and Yamamoto et al. (2015); however, those papers work with v_{max} instead of v_{peak}. None the less, the v_{form} in Masaki et al. (2013b) and Yamamoto et al. (2015) has a similar underlying structure to the criterion that serves as a proxy for subhalo age in our methodology.

Choosing lower values of f shifts the distribution of z_{form} to higher redshift. This can be problematic if the true value of z_{form} is then larger than the redshift, z_{max}, of the first P-Millennium snapshot where a subhalo is found. Nevertheless, to create informative galaxy colours based on subhalo v_{max} history, the lowest possible z_{form} values provide the most robust information about subhalo history. Thus, we conduct a test of the strength of subhalo history information with $f = 0.75$ and $f = 0.9$. Fig. 2 shows this test: the fraction of subhaloes for which z_{form} is assigned as the highest z at which the

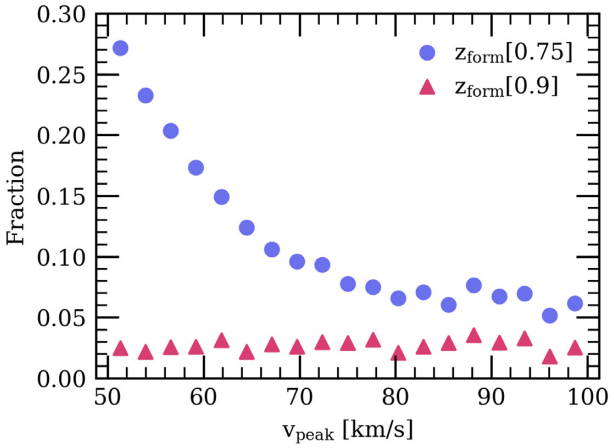

Figure 2. Fraction of subhaloes with subhaloes for which we only have lower limits on the value of z_{form}, plotted as a function of v_{peak}. The value of z_{form} becomes important in our colour assignment scheme. This plot informs us about the completeness of the mock's colour assignments. Subhaloes for which only a lower limit on z_{form} can be set are those whose earliest identified progenitor has $v_{max} > v_{form}$. The blue circles and red triangles correspond, respectively, to using $f = 0.75$ and $f = 0.9$ in equation (4) that defines v_{form}.

subhalo is found in P-Millennium in bins of v_{peak} at $z \sim 0.2$. We consider this condition to describe a 'poorly defined z_{form}', as it does not include information about the subhalo's history when its mass lies below the P-Millennium halo mass resolution. Thus, the blue circles in Fig. 2 trace the fraction of subhaloes that have reached a v_{peak} given on the horizontal axis by $z \sim 0.2$ but have a poorly defined z_{form}.

Fig. 2 illustrates the impact of the P-Millennium resolution on the choice of f: the smaller f is, the larger the limit on v_{peak} has to be to ensure that subhalo progenitor trees are sufficiently complete. Typically with $f = 0.75$, we can consider P-Millennium to be complete for subhaloes with $v_{peak} \geq 75$ km s^{-1}. We discuss this further in Section 3.1.

2.1.3 *Definition of central galaxies*

In Rosella, every central galaxy is located at the position of the most gravitationally bound particle in its host friends-of-friends halo. Galaxies in subhaloes outside the central gravitational well of a friends-of-friends halo are considered satellites.

2.2 SHAM with P-Millennium

There are several advantages to SHAM as a method for populating P-Millennium with galaxies.

Implementing SHAM is relatively quick compared to a physical method, such as a full semi-analytical galaxy formation model. Additionally, it can be arbitrarily tuned to reproduce certain statistics, as it includes empirical components in its methodology through its free parametrization via both functional models and numerical values.

SHAM is ideal for the analysis of groups and clusters for which BGS data may be used in the future and for which HOD models might not be complex enough. For example, it is not clear whether the mitigation techniques planned for DESI can recover statistics affected by assembly bias. BGS will benefit from mock data that includes halo assembly bias.

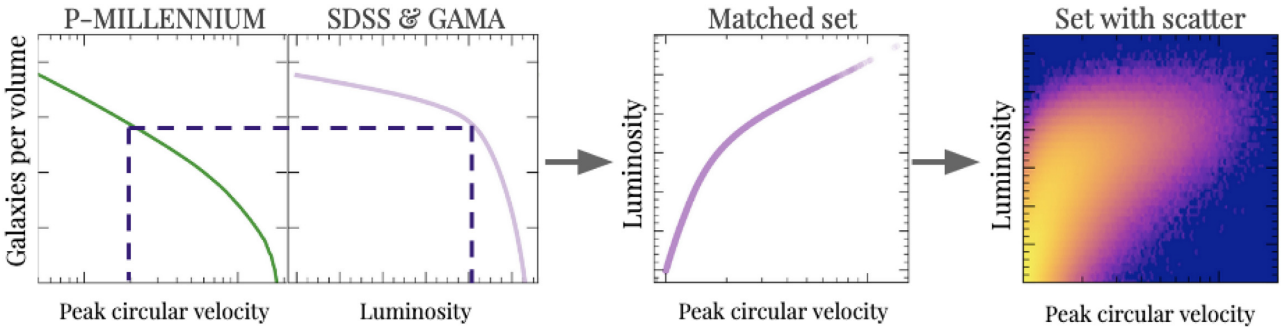

Figure 3. Illustration of assigning luminosities to galaxies using SHAM with the addition of scatter. The first two panels show abundance relations: left-hand panel shows the abundance of subhaloes as a function of their $v_{\rm peak}$ in our *N*-body simulation, and the second panel from the left shows the LF. For a given subhalo with a known $v_{\rm peak}$, we follow the dashed line to match its abundance in a simulation to a luminosity value that has the same abundance in observations. Repeating this matching for a set of subhaloes produces a set of points that form a tight line, as seen on the third panel from the left. Beginning with the third panel, we add scatter to the luminosity–$v_{\rm peak}$ data set. The data with the added scatter no longer follow a line in luminosity–$v_{\rm peak}$ space. The rightmost panel shows the logarithmic density of this illustrative set of data points in luminosity–$v_{\rm peak}$ space after the addition of scatter. The scatter method used here preserves the LF of the no-scatter counterpart of this SHAM data set, as discussed in Section 3.1.

2.2.1 *Assembly bias with SHAM*

Halo assembly bias describes the phenomenon that dark matter halo clustering depends on properties besides halo mass, including but not limited to formation time, concentration and spin (e.g. Gao, Springel & White 2005; Wechsler et al. 2006; Gao & White 2007). For a given halo mass, clustering is stronger in dark matter haloes that form at earlier times. The dependence of clustering on halo formation time increases with decreasing halo mass (Gao et al. 2005).

$v_{\rm max}$ characterizes the depth of gravitational potential. At fixed halo mass, $v_{\rm max}$ is directly related to halo concentration (e.g. Conroy et al. 2006; Zehavi et al. 2019). As halo concentration has been suggested to be a quantity that can track galaxy assembly bias, it offers the possibility of lifting the systematic effects of galaxy assembly bias in mock data. However, $v_{\rm max}$ describes the present state of a subhalo, which may miss some of the historical information contained in, for example, $v_{\rm peak}$. Chaves-Montero et al. (2016) offers one comparison of the qualities that $v_{\rm circ}$-related SHAM proxies impart on mock data.

This presents a problem for halo occupation models that assume the independence of the distribution and properties of galaxies from their environment beyond halo mass (Gao et al. 2005). Abundance matching on subhalo quantities that include information about their history, such as peak circular velocity $v_{\rm peak}$ or satellite subhalo accretion mass $M_{\rm acc}$, may lift part of this assumption of distribution-environment independence in the galaxy–halo occupation relation.

By incorporating a proxy that implicitly accounts for subhalo assembly history, $v_{\rm peak}$, a SHAM catalogue can be more informative when investigating the effects of assembly bias on observational data and computing statistics that may be affected by it, compared to a traditional HOD mock. Some work, however, has been done that allows tunable assembly bias to be included in modified HOD methods (e.g. Hearin et al. 2016) and SHAM methods (e.g. Contreras, Angulo & Zennaro 2020).

2.2.2 *Algorithm for luminosity assignment*

We assign luminosity values to galaxies in our mock catalogue by assuming that a galaxy occupies every dark matter subhalo that satisfies a minimum value of $v_{\rm peak}$. We assume that galaxy luminosities correlate with $v_{\rm peak}$.

$v_{\rm peak}$, by construction, includes information about a subhalo's formation history. When we populate satellite subhaloes with galaxies, $v_{\rm peak}$ allows us to account for the historical values of that subhalo's $v_{\rm max}$, thus mitigating the influence of effects like dark matter mass stripping as a consequence of mergers. There has been evidence of subhaloes with higher $v_{\rm peak}$ values tending to have higher concentration and earlier formation times, which are some of the properties associated with assembly bias (see Xu & Zheng 2020, and reference therein). $v_{\rm peak}$ can have some downsides, such as some post-merger transient features which may not correlate with changes in galaxy properties (Chaves-Montero et al. 2016).

Initially, we assume that subhalo $v_{\rm peak}$ follows a monotonic relation with the galaxy absolute magnitude in the *r* band, M_r^h. In the first step of luminosity assignment, we operate under the assumption that the relation between magnitude and $v_{\rm peak}$ are one to one, but that assumption is no longer applicable once we add scatter to the mock data. For the first, no-scatter, stage of our algorithm, the assumed relation between M_r^h and $v_{\rm peak}$ can be expressed as

$$n_{\rm g}(< M_r^h) = n_{\rm h}(> v_{\rm peak}). \tag{5}$$

Here, $n_{\rm g}$ is the number density of galaxies of a given M_r^h or brighter, and $n_{\rm h}$ is the number density of subhaloes of a given $v_{\rm peak}$ or higher. In other words, the magnitude M_r^h that we assign to a galaxy in a subhalo with $v_{\rm peak,i}$ is set by matching the abundance of subhaloes with $v_{\rm peak} > v_{\rm peak,i}$ to the abundance of galaxies with $M_r^h < M_{r,\rm i}^h$.

We follow a number of specific steps to assign magnitude values to the galaxies in our sample:

(i) Get the evolving *r*-band galaxy LF using the *SDSS r*-band LF (Blanton et al. 2003) and the GAMA *r*-band LF (Loveday et al. 2012). The combined smooth LF used here is the one that Smith et al. (2017) used for the development of a DESI BGS light-cone mock catalogue. We call this set of reference data the 'target LF', as it is the LF that we aim to replicate in our mock.

(ii) Perform SHAM with zero scatter using the target LF with the monotonic relation between luminosity and $v_{\rm peak}$ in equation (5). Chaves-Montero et al. (2016) and McCullagh et al. (2017) also used this relation as the basis of their SHAM assignments. Fig. 3 offers an illustration of the process.

(iii) Add luminosity-dependent scatter, following McCullagh et al. (2017),[11] using a magnitude-dependent scatter $\sigma(M_r^h)$ to produce results that are illustrated in Fig. 3. See below for more details on the scatter algorithm.

[11]This method effectively shuffles the ranks while maintaining the originally assigned set of luminosities. Hence, it does not perturb the cumulative LF,

Before scatter, we use the galaxy cumulative LF down to $M_r^h = -10$ for the purposes of fully utilizing our LF-preserving scatter method. After scatter, we keep galaxies that are $M_r^h = -17.5$ or brighter, which corresponds to a minimum $v_{\rm peak}$ of $\sim$75 km s^{-1}. Our choice to limit the analysis to subhaloes with $v_{\rm peak} \geq 75$ km s^{-1} is motivated by the P-Millennium resolution (see Sections 2.1.2 and 3.1).

2.2.3 Adding scatter to SHAM

The approach described here uses a magnitude-dependent scatter magnitude $\sigma(M_r^h)$ (called ΔM_r^h in McCullagh et al. 2017) to produce results that are illustrated in Fig. 3. We execute the following four steps to add luminosity-dependent scatter to the magnitude values of the galaxies in our sample:

(i) Assign a magnitude without scatter, M_r^h, to every galaxy using the method described above;

(ii) For every galaxy, draw a new magnitude, $M_r^{h\prime}$, from a Gaussian distribution clipped at 2.5 $\sigma(M_r^h)$, with the mean equal to the galaxy's M_r^h value and the standard deviation $\sigma(M_r^h)$ computed as a function of the galaxy's absolute magnitude. In this work, $\sigma(M_r^h)$ is given by a smooth step function of the form

$$\sigma\left(\mathrm{M}_r^h\right) = \alpha + \beta \tanh\left(\mathrm{M}_r^h - M_{r,\mathrm{ref}}^h\right), \quad (6)$$

where α, β, and $M_{r,\mathrm{ref}}^h$ are free parameters that we can tune to match clustering (Section 2.3 discusses tuning the parameters in this method). A variable, luminosity-dependent $\sigma(\mathrm{M}_r^h)$ was chosen to create luminosity-threshold clustering of the data that matches the clustering of galaxies in observations and existing mock data, (for the clustering analysis, see Section 3.4.1). To create the Rosella catalogue presented here, we use the following parameter values for the model in equation (6): $\alpha = 0.8$; $\beta = 0.4$; $M_{r,\mathrm{ref}}^h = -20$

(iii) Rank galaxies in order of the new magnitude, $M_r^{h\prime}$;

(iv) Rank subhaloes in order of their $v_{\rm peak}$ values;

(v) Place galaxies in subhaloes so that the galaxies with the brightest $M_r^{h\prime}$ are located in the subhaloes with the largest $v_{\rm peak}$ values;

(vi) Assign each galaxy's original magnitude, M_r^h, to the galaxy's final location computed in the step above.

2.2.4 Luminosity-dependent colour assignment

A number of methods that have built upon original abundance matching assign colours to galaxies in gravity-only simulations based on (sub-)halo age or environment (e.g. Hearin & Watson 2013; Masaki et al. 2013b; Hearin et al. 2014; Yamamoto et al. 2015).

A common approach to assigning galaxy colours in a SHAM-like paradigm matches subhaloes' directly simulated (sub-)halo property, such as $v_{\rm max}$ or $v_{\rm peak}$, and a secondary (sub-)halo property that serves as a proxy for its age (see Masaki et al. 2013b; Kulier & Ostriker 2015; Yamamoto et al. 2015). This is the so-called 'age model' of the dark matter halo-based prediction of galaxy colour. The approach is based on the notion that older galaxies should contain older, and, consequently, redder, stellar populations. Thus, if galaxy colour can be used as a measure of stellar population age when we analyse observations, we should be able to reverse the process and assign colours to simulated galaxies based on the ages of their subhaloes.

The procedure for the assignment of $^{0.1}(g-r)$ colours to Rosella galaxies comprises three steps, illustrated in Fig. 4, and is built around

and no deconvolution is necessary, unlike other methods of adding scatter to SHAM data.

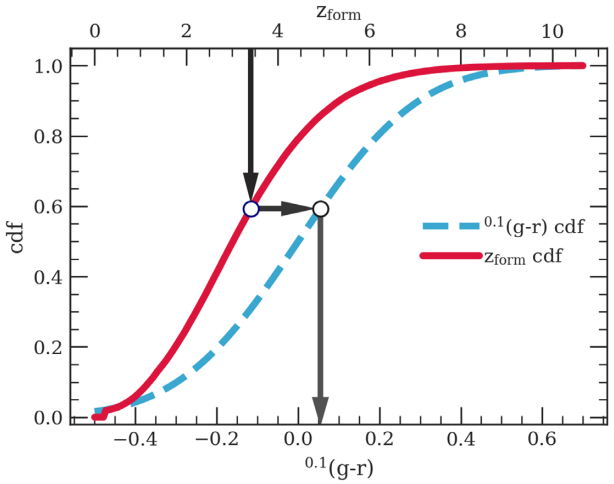

Figure 4. Illustration of colour assignment for a single subhalo. The top horizontal axis shows formation redshift, $z_{\rm form}$; the bottom horizontal axis shows $^{0.1}(g-r)$. The vertical axis shows cdf values for the red (solid) and blue (dashed) curves. The red (solid) curve is the cdf of subhalo $z_{\rm form}$ values for a given $v_{\rm peak}$. 0 on this curve means that no subhaloes of the given $v_{\rm peak}$ should be expected to have that or lower $z_{\rm form}$. 1 on the red (solid) curve signifies that all subhaloes of the given $v_{\rm peak}$ should be expected to have lower $z_{\rm form}$ values. The red (solid) curve is computed by interpolating between $z_{\rm form}$ cdf curves calculated in bins of $v_{\rm peak}$ (see the left-hand panel of Fig. 5 for examples). The blue (dashed) curve is given by equation (7) and is the cdf of $^{0.1}(g-r)$ for this subhalo's M_r^h.

two notions. Galaxy colour bimodality analyses (e.g. Baldry et al. 2004) show that brighter galaxies tend to be redder across both blue and red populations of galaxies. Thus, we begin colour assignment by calculating a cdf of $^{0.1}(g-r)$, conditional on M_r^h, individually for each galaxy. We describe this procedure in Section 2.2.5.

The second part of our colour assignment procedure builds upon the correlation between galaxy colour and age (e.g. Mo et al. 2010; Hearin 2015; Chaves-Montero & Hearin 2020). Our colour assignment method requires relating the cdf of $z_{\rm form}$ at a given $v_{\rm peak}$ value to the cdf of M_r^h-dependent $^{0.1}(g-r)$ colour distributions. However, we do not know the distribution of $z_{\rm form}$ for any individual galaxy with its unique $v_{\rm peak}$ a priori. We therefore construct cdfs of $z_{\rm form}$ from subsample populations of galaxies in narrow bins of $v_{\rm peak}$. Examples of such $z_{\rm form}$ distribution functions are provided in the left-hand panel of Fig. 5.

Note that median $z_{\rm form}$ values decrease with increasing median $v_{\rm peak}$ values in the left-hand panel of Fig. 5, which is a consequence of the hierarchical formation of structure in the Universe. During colour assignment, we interpolate between the full set of curves covering the full range of $v_{\rm peak}$ values to find an appropriate $z_{\rm form}$ cdf for each subhalo.

As the final step in $^{0.1}(g-r)$ assignment, we find each subhalo's position on the $v_{\rm peak}$-dependent distribution of $z_{\rm form}$ and translate it to a $^{0.1}(g-r)$ value for the galaxy residing in it, as illustrated in Fig. 4 and discussed in more detail below.

2.2.5 Luminosity-dependent galaxy colour distributions

The colour–magnitude diagram of observed galaxies has a bimodal distribution (e.g. Baldry et al. 2004) that can be described as a sum of components that correspond to red and blue galaxy populations. To assign a colour to a galaxy in Rosella, we start with the empirical

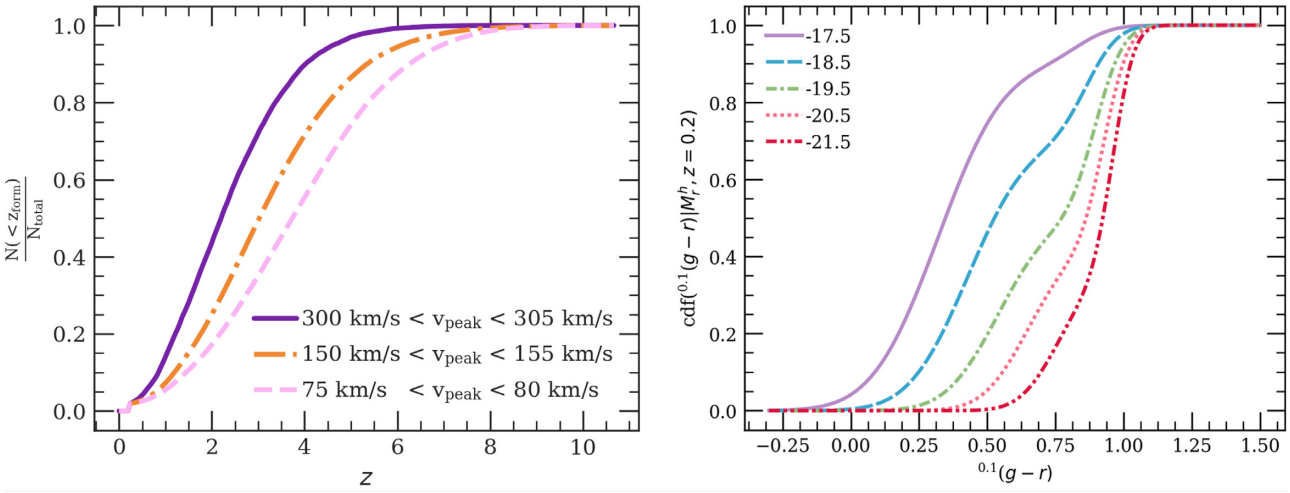

Figure 5. Left-hand panel: Example cdfs of formation redshift in bins of $v_{\rm peak}$. The vertical axis displays the fraction of subhaloes with a given formation redshift $z_{\rm form}$ or lower. The horizontal axis is the formation redshift $z_{\rm form}$. Curves are colour coded by bins of $v_{\rm peak}$. Right-hand panel: cdfs of $^{0.1}(g-r)$ at $z = 0.2$ for a selection of M_r^h values, as indicated in the legend. The functional form of these distributions is given in equation (7).

model for the observed bimodal luminosity-dependent distribution of $^{0.1}(g-r)$ colours given in Smith et al. (2017).

At a given magnitude M_r^h, it is assumed that the blue and red components of the $^{0.1}(g-r)$ distribution functions each have Gaussian forms and that their combined cdf is given by

$$\begin{aligned}\mathrm{cdf}\left(M_r^h\right) = {} & f_{\rm blue}\left(M_r^h\right)\,\mathrm{G}\left(M_r^h\right)_{\rm blue} \\ & + \left(1 - f_{\rm blue}\left(M_r^h\right)\right)\,\mathrm{G}\left(M_r^h\right)_{\rm red},\end{aligned} \tag{7}$$

where $f_{\rm blue}$ is the fraction of blue galaxies. This fraction is a function of magnitude

$$f_{\rm blue} = \begin{cases} 0 & \text{if } M_r^h < -26.571 \\ 0.46 + 0.07(M_r^h + 20.) & \text{if } -26.571 \le M_r^h < -19.539 \\ 0.4 + 0.2(M_r^h + 20.) & \text{if } -19.539 \le M_r^h > -17.173 \\ \frac{1}{1+\exp(-(M_r^h+20.5))} & \text{if } M_r^h > -17.173, \end{cases} \tag{8}$$

while the mean and scatter of each of the Gaussian components are magnitude- and redshift-dependent, given in equation (10). The sigmoid expression for the faintest galaxies ensures that the fraction of red galaxies slowly tapers off instead of meeting a sharp cutoff at a fixed magnitude, which makes our model slightly different from the prescription in Smith et al. (2017).

We adopt relations from Smith et al. (2017) evaluated at $z = 0.2$ as the mean and scatter of each of the Gaussian components in equation (7)[12]

$$\begin{aligned}\left\langle {}^{0.1}(g-r)|M_r^h\right\rangle_{\rm blue} &= 0.595 - 0.11\left(M_r^h + 20\right) \\ \mathrm{rms}\left({}^{0.1}(g-r)|M_r^h\right)_{\rm blue} &= 0.14 + 0.02\left(M_r^h + 20\right) \\ \left\langle {}^{0.1}(g-r)|M_r^h\right\rangle_{\rm red} &= 0.914 - 0.032\left(M_r^h + 20\right) \\ \mathrm{rms}\left({}^{0.1}(g-r)|M_r^h\right)_{\rm red} &= 0.076 + 0.01\left(M_r^h + 20\right).\end{aligned} \tag{10}$$

[12]The z term in the expression for $\mathrm{rms}({}^{0.1}(g-r)|M_r^h)_{\rm red}$ contained a typo in equation (33) in Smith et al. (2017). The correct formulation is

$$\begin{aligned}\mathrm{rms}\,(g-r \mid M_r)_{\rm red}\,(z) = {} & \mathrm{rms}\,(g-r \mid M_r)_{\rm red} \\ & + 0.05(z-0.1) + 0.1(z-0.1)^2.\end{aligned} \tag{9}$$

The right-hand panel in Fig. 5 shows examples of $^{0.1}(g-r)$ colour cdfs for a selection of M_r^h values.

We connect the cdfs of $z_{\rm form}$ to those of $^{0.1}(g-r)$, as illustrated in Fig. 4, where the $z_{\rm form}$ cdf for a single subhalo is given by the red curve and is conditional on its $v_{\rm peak}$. Colour assignment consists of four steps:

(i) Compute the $^{0.1}(g-r)$ cdf by applying a galaxy's M_r^h magnitude to equation (7);
(ii) Compute the $z_{\rm form}$ cdf from the host subhalo's $v_{\rm peak}$ value, as described in Section 2.2.3;
(iii) Find the cdf value corresponding to the host subhalo's $z_{\rm form}$, as shown by the top vertical arrow in Fig. 4;
(iv) Determine the $^{0.1}(g-r)$ value that matches the aforementioned cdf value, as demonstrated by the horizontal arrow in Fig. 4. Assign this $^{0.1}(g-r)$ value to the galaxy.

2.3 Tuning the catalogue

The method has the following freedoms and free parameters:

(i) The subhalo attribute connecting its present state to its history for age-matching colour assignment – in the current method, this attribute is the distribution of $z_{\rm form}$ conditional on $v_{\rm peak}$ and M_r^h (see Section 2.1.2 and Fig. 4);
(ii) The functional form of $\sigma(M_r^h)$ in equation (6);
(iii) The parameters α, β, and $M_{r\,\rm ref}^h$ in equation (6);
(iv) The specific definition of $z_{\rm form}$ (see Section 2.1.2), including f in equation (4).

By construction, our $z \sim 0.2$ mock is tuned to reproduce the galaxy LF, following the parametrization proposed by Smith et al. (2017), which agrees with observational constraints provided by *SDSS* and GAMA. We also match, by construction, the luminosity-dependent colour distribution.

We tune the free parameters of the Rosella mock to match the observed luminosity- and colour-dependent clustering by comparing our results to the Smith et al. (2017) Millennium-XXL mock, as it represents observational data well. The Millennium-XXL mock fits the observational data at a range of redshifts, but can be estimated

at $z \sim 0.2$, the reference redshift of Rosella. In this work, we compare Rosella to the clustering of galaxies in the Millennium-XXL mock as it is presented in Smith et al. (2017). The authors of Smith et al. (2017) use redshift ranges that correspond to *SDSS* volume-limited luminosity threshold samples in Zehavi et al. (2011). We also compare our results to *SDSS* results presented in Zehavi et al. (2011). For luminosity-dependent clustering, we examine redshift-space results in the context of existing mock data.

Full optimization of these parameters is beyond the scope of this paper, as that depends on the science goals for which Rosella and its methods are to be used.

To choose the value of f in equation (4), we also consider the resolution of subhalo progenitors, as shown in Fig. 2. To choose the subhalo attribute for age-matching colour assignment, we also considered the galaxy colour bimodality discussed in Section 3.2.

We examined the effect of the choice of f on the colour-dependent clustering of Rosella galaxies. The difference in the colour-dependent clustering between $f = 0.9$ and $f = 0.75$ (the default value) is small. Qualitatively, changing the value of f from 0.75 to 0.9 slightly increases the gap between the clustering of red and blue galaxies. On small scales, 0.75 provides a better match, and we do not see the reason to increase the gap between red and blue galaxy clustering by setting f to 0.9 on the larger scales.

The motivation for a non-zero β in equation (6) is the observation that scatter driven by a constant $\sigma(M_r^h)$ produces unsatisfactory clustering results, generating a data set with clustering that was consistently higher than that measured in observations, as shown in Zehavi et al. (2011), particularly in the brightest samples. A luminosity-dependent formulation for $\sigma(M_r^h)$ brought the clustering of the mock closer to that of observations.

$\sigma(M_r^h)$ ranging from $\sim$0.4 for the brightest galaxies and $\sim$1.2 for the faint end produced favourable clustering results in our analysis. The sigmoid shape of tanh(x) and the fact that tanh(x) is bound to $(-1, +1)$ naturally brought us to the values $\alpha \sim 0.8$ and $\beta \sim 0.4$.

When we first implemented the scatter using a standard, non-clipped, Gaussian, objects that started out with low luminosities overwhelmed the brightest population because of the comparatively large abundance of the low-luminosity objects. This leads central galaxies to form a bimodal distribution at masses of friends-of-friends haloes with $\log_{10}(M_{200,\mathrm{mean}}/h^{-1}M_\odot) > 13$. This is a result of the fact that our method for adding scatter to M_r^h does not distinguish between satellite and central subhaloes. We tried a variety of modifications to our scatter method and found that clipping the Gaussian in our scatter at 2.5σ solved the problem of false central galaxy M_r^h bimodality. We discuss this further in Section 3.3.

Our definition of z_{form} as the subhalo attribute that connects a subhalo's colour to its history was inspired by Masaki et al. (2013b) and Yamamoto et al. (2015); with a modification that our z_{form} is defined in terms of v_{peak}, as opposed to v_{max}.

To calculate our clustering results, we use the publicly available code CORRFUNC[13] (Sinha & Garrison 2017; Sinha & Garrison 2019).

3 PROPERTIES OF THE ROSELLA CATALOGUE

In this section, we examine the Rosella mock produced using the methodology introduced in Section 2. In Section 3.1, we open with a discussion of the properties of the LF of the galaxies in Rosella and discuss the brightness limits that it can potentially reach, followed by, in Section 3.2, the galaxy colour bimodality. Section 3.3 describes the impact that our model of luminosity scatter has on the distribution of central and satellite galaxies in the mock. Section 3.4 considers the clustering in our mock, with a comparison to previously published observational and simulated data.

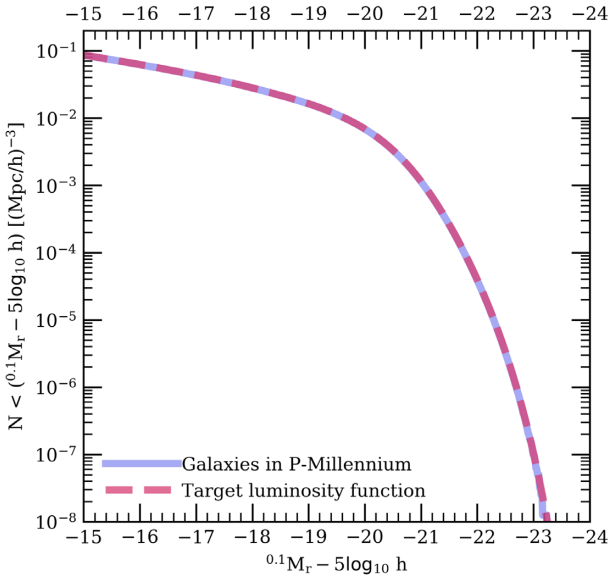

Figure 6. The *r*-band cumulative LF. The function for galaxies in the mock catalogue is plotted as a solid violet line. The pink dashed line is the target LF based on *SDSS* and GAMA observations, taken from the fit provided in Smith et al. (2017).

3.1 Rosella luminosity function and resolution

By construction, the implementation of SHAM used here reproduces its target LF. Fig. 6 demonstrates that the LF produced in our mock exactly matches the cumulative galaxy LF based on *SDSS* (Blanton et al. 2003) and GAMA (Loveday et al. 2012) data provided in Smith et al. (2017) to magnitudes at least as faint as $M_r^h = -15$. This, however, does not mean that the properties of the Rosella catalogue are converged at such faint magnitudes: these faint galaxies may reside in haloes of such low v_{peak} as to be where the P-Millennium catalogue is incomplete. Moreover, to assign a colour to a Rosella galaxy, we need to have a reliable z_{form} for such haloes. Earlier in Fig. 2, we saw that we require $v_{\mathrm{peak}} > 75$ km s^{-1} for z_{form} to be well defined. Hence, to determine the magnitude limit down to which Rosella is complete and produces reliable colours, we need to identify the magnitude at which the mock galaxies reside only in haloes with $v_{\mathrm{peak}} > 75$ km s^{-1}. This is revealed in Figs 7 and 8.

Fig. 7 shows the SHAM absolute magnitudes as a function of v_{peak} before (white curve) and after (hexbin colour map) scatter has been added. Histograms through this distribution are shown for three magnitude bins in Fig. 8. From these, we see that the magnitude bin extending as faint as $M_r^h = -17.5$ tapers smoothly to zero above $v_{\mathrm{peak}} = 75$ km s^{-1}, indicating that our catalogue is complete to this magnitude limit.

The lower limit on the absolute magnitude that produces a complete sample of galaxies may vary if one were to add scatter that follows a functional form different from equation (6) or utilize a different set of parameters, or apply this method to a simulation other than P-Millennium.

[13] https://github.com/manodeep/Corrfunc

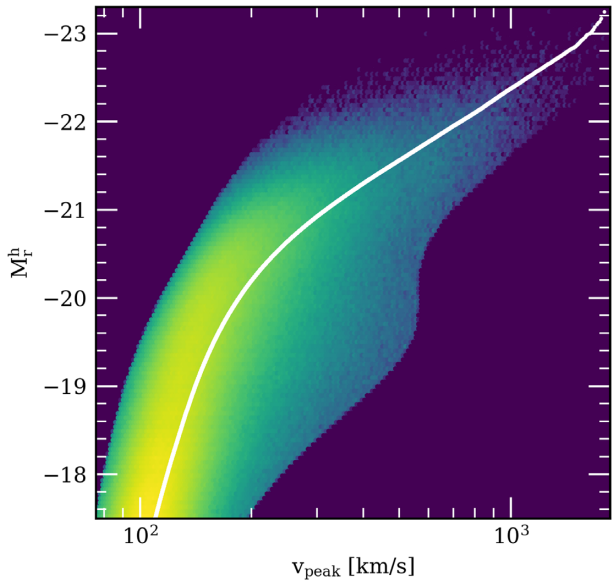

Figure 7. Hexbin map of SHAM absolute magnitudes with scatter. The colour indicates the number of galaxies per hexagonal bin of given M_r^h and v_{peak} values, plotted on a logarithmic scale, with purple indicating bins with zero galaxies and lime-green indicating bins with the most galaxies. The white line plotted on top of the hexbin map shows the M_r^h values assigned to P-Millennium subhaloes before the addition of scatter. The density of galaxies in the brightest yellow regions is about 5 orders of magnitude higher than the faintest non-zero blue regions.

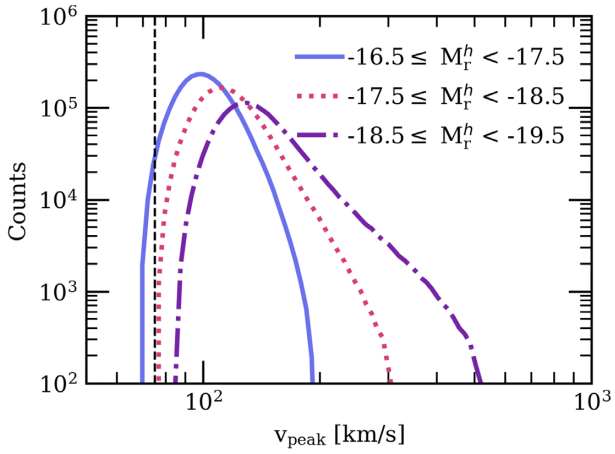

Figure 8. Histograms of v_{peak} in bins of M_r^h, created by drawing M_r^h-limited samples, as indicated in the legend, from the full set of galaxies depicted in Fig. 7. The histograms cover the same set of 125 bins that cover the range 45 km s^{-1} < v_{peak} <2500 km s^{-1}. The dashed black vertical line indicates the v_{peak} = 75 km s^{-1} boundary.

3.2 Galaxy colour bimodality

Fig. 9 shows histograms of $^{0.1}(g-r)$ colour values in Rosella, along with curves produced by the analytical expressions generated at specific M_r^h values with equation (7). By construction, we match the colour distributions in Smith et al. (2017), which were designed to match *SDSS* and GAMA data. The resulting colour distributions are compared to observational GAMA data in fig. 14 of Smith et al. (2017). The histograms in Fig. 9 reveal a good match with the target colour distributions examined in fig. 14 in Smith et al. (2017), which, in turn, provide a good match to those of the *SDSS* and GAMA

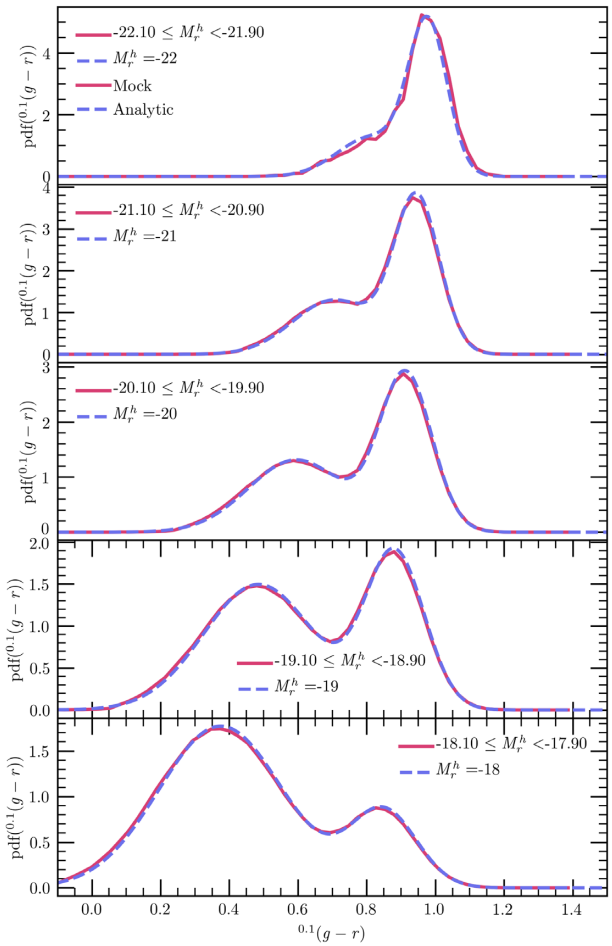

Figure 9. Distribution of $^{0.1}(g-r)$ values among Rosella galaxies. The red (solid) shows a normalized histogram of Rosella galaxies that fall in the range of M_r^h values indicated in the legend. The blue (dashed) line is the input function 7, calculated at values of M_r^h indicated in the legend of each panel.

surveys (e.g. Baldry et al. 2004). Each histogram generally shows two major peaks with the blue being dominant for low luminosity and the red for high luminosity. The location of both peaks moves redward with increasing luminosity. These distributions combine to produce the colour–magnitude diagram shown in Fig. 10, whose morphology is akin to those shown in the literature (e.g. Baldry et al. 2004).

3.3 Distribution of central and satellite galaxies

The number of satellite galaxies as a fraction of the total galaxy population in Rosella varies with halo mass. In the top panel in Fig. 11, one can see that galaxies that are assigned brighter absolute magnitudes with SHAM before scatter are preferentially central galaxies. In both cases of SHAM samples with and without scatter, the trend in the ratio of central galaxies to the total galaxy population tapers off to an almost constant rate of about 60 per cent between $M_r^h = -20$ and $M_r^h = -17.5$. The scattered sample of SHAM, however, exhibits a lower fraction of satellites compared to the no-scatter sample at the bright end of the catalogue. This is the result of the scattering process moving the magnitudes of galaxies that start out in central subhaloes to satellite subhaloes.

Fig. 11 shows the fractions of blue and red galaxy populations that are central, given the galaxies' M_r^h. The nominal separation between

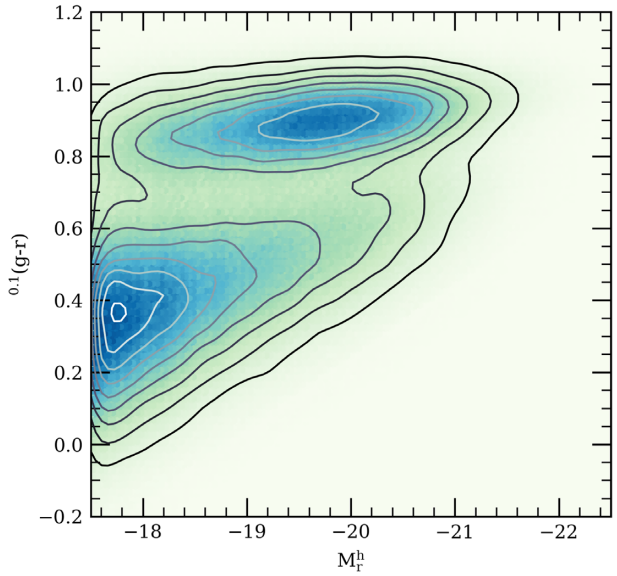

Figure 10. Colour–magnitude diagram of Rosella galaxies as a hexbin map with contours. The map shows the density of galaxies in hexagonal bins of ${}^{0.1}(g-r)$ and M_r^h values. Fig. 9 shows slices through this distribution.

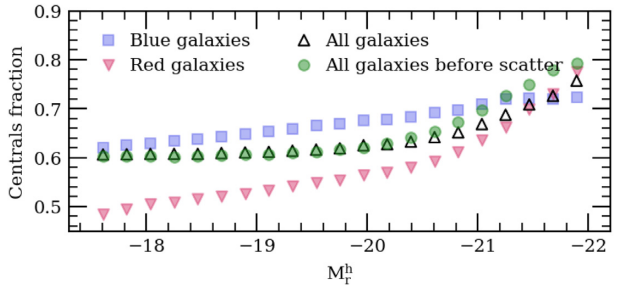

Figure 11. Fraction of red (filled red triangles) or blue (blue squares) galaxies that are centrals given the galaxies' M_r^h; fraction of all galaxies that are central in Rosella with (empty triangles) and without (green circles) scatter as a function of the galaxies' M_r^h. Blue and red galaxy populations are defined in equation (11).

'red' and 'blue' galaxies is given by an expression introduced in Zehavi et al. (2005)

$$ {}^{0.1}(g-r)_{\rm cut} = 0.21 - 0.03 M_r^h. \quad (11) $$

Galaxies whose ${}^{0.1}(g-r)$ values are greater than this ${}^{0.1}(g-r)_{\rm cut}$ are classified as 'red', while the others are 'blue'.

The trend in Fig. 11 demonstrates a steady increase in the fraction of central galaxies across the range of absolute magnitudes among Rosella galaxies in both red and blue galaxies. The blue population has a higher central galaxy fraction compared to the red population across all magnitudes, except for the brightest bins, with $M_r^h < \sim -21.5$.

Fig. 12 shows the normalized distributions of central and satellite galaxies in bins of host halo mass of 0.5 dex width. We see that the no-scatter SHAM sample (bottom panel of Fig. 12) exhibits a clear and expected trend of the peak of the distribution of centrals in the catalogue moving to a brighter magnitude with increasing halo mass.

The top panel in Fig. 12 shows that when we add scatter using the formulation in Section 2.2.2, the distinct population of central galaxies is preserved. This result comes from trying different prescriptions for adding scatter to M_r^h, and was achieved when

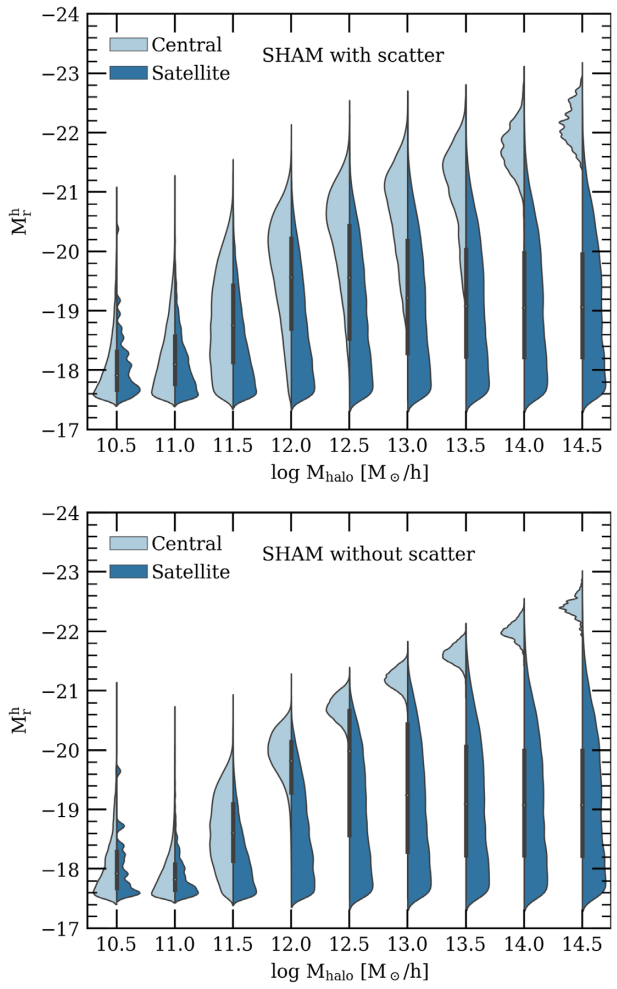

Figure 12. Distribution of central and satellite galaxies in halo mass bins for a sample of Rosella galaxies with $M_r^h < -17.5$. The vertical axis shows M_r^h, and the horizontal axis represents the bins of host halo mass ($M_{200,\rm mean}{}^{14}$). The top panel shows the distributions of satellite (light blue) and central (dark blue) galaxies with respect to their M_r^h values in bins of halo mass in Rosella with scatter described in Section 2.2.2. The bottom panel shows analogous distributions for a SHAM catalogue with no scatter. Kernel smoothing has been applied to these violin histograms, which creates the false illusion of data stretching to magnitudes fainter than M_r^h of −17.5. The plots are normalized in a way that lets all histograms have the same width to draw our attention to the distribution of galaxies along the M_r^h axis, and not to the relative sizes of these populations.

we combined the luminosity-dependent scatter (equation 6) with a Gaussian distribution clipped at 2.5 σ.

3.4 Real- and redshift-space clustering in Rosella

Studies of the clustering of early- and late-type galaxies, classified by spectral type, offer observational evidence of the dependence of the strength of galaxy clustering on morphology and luminosity. Observational evidence points to a trend in the spatial correlation function, where brighter galaxies are more clustered than their fainter counterparts (e.g. Norberg et al. 2001; Zehavi et al. 2005; and references therein). Early studies of this phenomenon considered red and blue galaxies classified by spectral type, and observed that galaxies that belong to the 'early-type' population, which has been shown to be dominated by red and quenched galaxies, is more

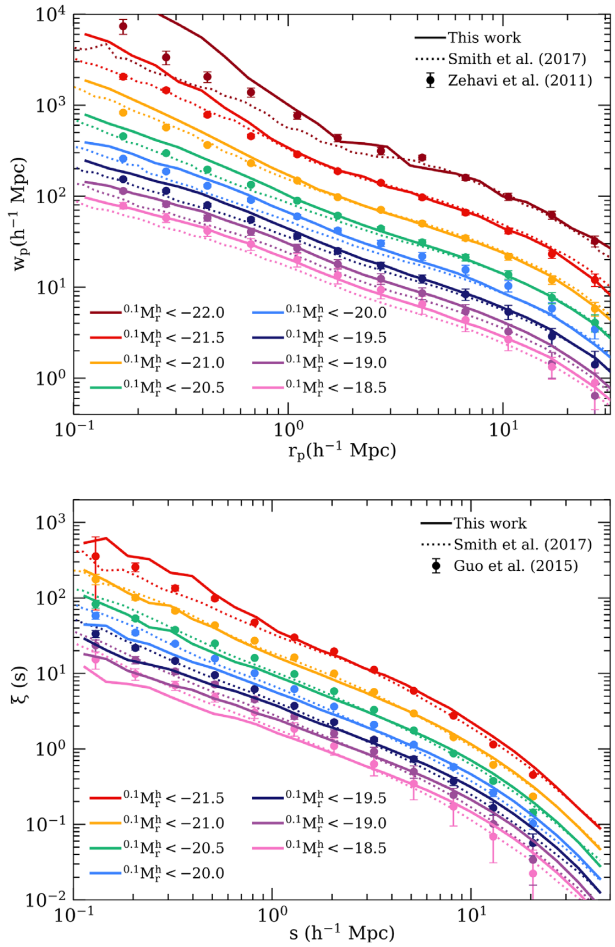

Figure 13. Projected correlation function (upper panel) and redshift-space correlation function (lower panel) for luminosity threshold galaxy samples. The solid lines show the clustering results of Rosella. The solid points with error bars represent clustering measurements using *SDSS* data from Zehavi et al. (2011) (upper panel) and Guo et al. (2015) (lower panel). The dotted lines show the corresponding correlation functions from the Millennium-XXL mock catalogue in Smith et al. (2017). The results for each sample have been offset by successive intervals of 0.15 dex, starting at the $M_r^h < -20.5$ sample, with the faintest sample corresponding to the lowest curve in the graph, for clarity.

clustered than the 'late-type' population (e.g. Norberg et al. 2001; Norberg & 2dFGRS Team 2002; Zehavi et al. 2005; and references therein). The relatively high clustering of more luminous, redder galaxies, has led the luminous red galaxy (LRG) population to be a popular target sample for galaxy surveys that aim to study the large-scale structure of the Universe (e.g. Eisenstein et al. 2005a, b).

The luminosities and colours assigned to our high-fidelity mock offer a possibility of comparing the colour- and luminosity-dependent correlation functions to the trends observed in past surveys.

3.4.1 *Clustering as a function of luminosity*

Projected correlation functions of galaxies in Rosella are shown by the bold curves in the upper panel of Fig. 13 for different luminosity threshold samples at $z \sim 0.2$. In the figure, we show the projected two-point correlation functions (2PCF) calculated using the publicly available code corrfunc (Sinha & Garrison 2017; Sinha & Garrison 2019).

The samples presented in the upper panel of Fig. 13 show the projected 2PCF in samples of galaxies with a faint limit on absolute magnitude (luminosity threshold). The sample cut-off limits have been chosen to make it possible to compare the clustering results of Rosella data to those of the HOD mock presented in Smith et al. (2017) and of the *SDSS* data presented in Zehavi et al. (2011). It should be noted, however, that in addition to luminosity thresholds, the observed clustering of galaxies in Zehavi et al. (2011) and Smith et al. (2017) was measured for volume-limited samples. Each volume-limited sample covers a specific range of redshifts, and the range is wider for the bright samples. Rosella, on the other hand, is a single snapshot at $z \sim 0.2$, which may result in slight differences in the clustering of galaxies in Rosella and the *SDSS* data (Zehavi et al. 2011) and MXXL mock (Smith et al. 2017). Considering that the *SDSS* and MXXL mock data do not cover the same redshift sample as Rosella, a more robust comparison of Rosella clustering to data would require detailed tuning of Rosella using data that is centred on $z \sim 0.2$, which will be available from DESI.

While Rosella's projected 2PCF fits the *SDSS* data quite well on scales greater than 1 h^{-1} Mpc, all but the two faintest samples exhibit clustering that appears to be slightly too high on small scales. We suspect that this might be a result of our SHAM model treating satellite and central galaxies in the same manner. Whether this feature of the model is compatible with quenched fraction estimates in Mandelbaum et al. (2016) is worth investigating in further work.

Additionally, the MXXL mock included galaxies assigned to haloes which were given random positions, corresponding to haloes below the mass resolution of the MXXL simulation. This random position assignment dilutes the clustering of MXXL galaxies slightly for faint galaxy samples, which explains why the clustering of the MXXL mock is low compared to Rosella for the $M_r^h < -18.5$ and $M_r^h < -19$ samples.

We have conducted the luminosity-dependent clustering analysis for a variety of models of scatter during the process of tuning our mock, presented in Section 2.3.

The lower panel in Fig. 13 shows the redshift-space correlation function monopole for Rosella, compared to the redshift-space clustering of the mock presented in Smith et al. (2017), as well as clustering of *SDSS* data from Guo et al. (2015). We note a slight difference in shape and amplitude of the redshift-space 2PCF monopole between Rosella and *SDSS*. Addressing this difference would require further in-depth analysis that we leave for a future work.

3.4.2 *Clustering as a function of colour*

Fig. 14 shows the projected correlation function of Rosella galaxies separately for red and blue galaxy populations in bins of absolute magnitude. The same figure shows a comparison of our data to those presented in Smith et al. (2017) and Zehavi et al. (2011), where red and blue samples are defined using the same colour cut as in this work's, given by equation (11).

For the purposes of analysis, the nominal separation between 'red' and 'blue' galaxies is given by equation (11). It should be noted that this expression, first introduced in Zehavi et al. (2005), does not account for the fact that there is a continuum in galaxy colours, and instead serves as a tool for comparing colour-dependent clustering among different samples.

For *SDSS*, the clustering of the red galaxy population is stronger than that of the blue galaxy population. This effect is likely associated with the presence of red elliptical galaxies, which are more likely

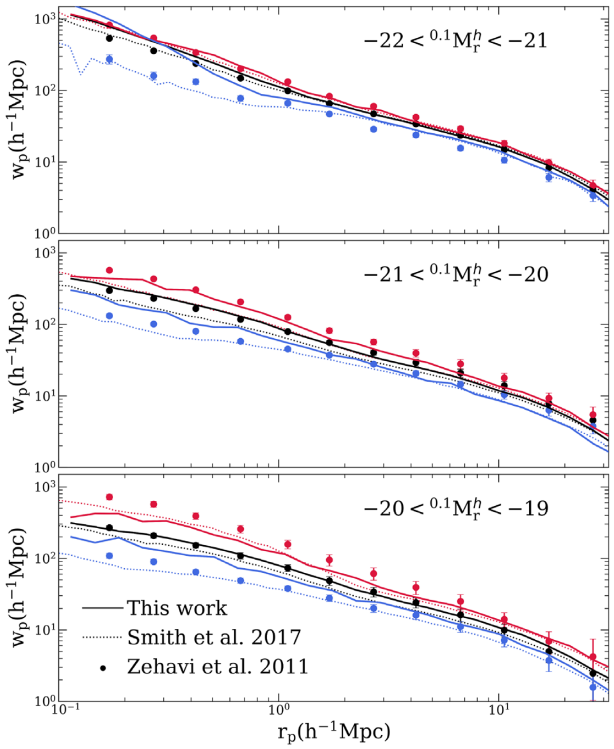

Figure 14. Projected correlation function for red and blue galaxies in bins defined by absolute magnitude. The clustering of Rosella galaxies is presented in bold lines. Clustering of low-redshift galaxies from the Millennium-XXL catalogue in Smith et al. (2017) is plotted in faint lines. The lines in each box correspond to bins defined by ranges of absolute magnitude, as indicated in the legend. Points with error bars correspond to the analysis of volume-limited samples of *SDSS* data in Zehavi et al. (2011). The clustering of all galaxies in a sample is shown in black. Red and blue galaxy populations, defined by equation (11), are plotted in red and blue colours, respectively.

to reside in the more strongly biased massive haloes (e.g. Eisenstein et al. 2005b). As the samples get fainter, the strength of the colour dependence evidently increases for both the observational data and the galaxies presented in Rosella.

In summary, the comparison of Rosella galaxy clustering shows a favourable match to *SDSS* data, considering the differences that may arise from comparing Rosella's fixed-redshift sample to the volume-limited samples that cover ranges of redshifts in *SDSS*. This is therefore useful for developing analyses of DESI BGS.

4 CONCLUSION

Modern galaxy surveys require realistic mock catalogues in order to test analysis tools, assess completeness, and determine error covariances in observed data. The mock catalogues can serve as the connector of quantities that the surveys observe, such as galaxy luminosities and positions, to quantities that are only available in simulations, including but not limited to host halo and subhalo masses, velocities, and halo assembly histories.

We have outlined a method for creating a mock galaxy catalogue that closely mimics data that will be observed in DESI's BGS (DESI Collaboration 2016; Ruiz-Macias et al. 2020). The resulting mock, Rosella, provides the rest-frame r-band absolute magnitudes, rest-frame ${}^{0.1}(g-r)$ colours, and 3D positions and velocities for galaxies inhabiting a volume of approximately $(542\ \mathrm{Mpc\,h^{-1}})^3$, as well as the masses of their host haloes.

The approach described here relies on SHAM with luminosity-dependent scatter to populate the P-Millennium N-body simulation with galaxies and assign them rest-frame absolute magnitudes in r band, M_r^h. Due to our approach of adding scatter to the mock, Rosella preserves the target LF by construction. Our method also faithfully reproduces a specified redshift-dependent target distribution of ${}^{0.1}(g-r)$ colours. The colours it assigns are linked to the formation redshifts we determine by tracking the formation history of each individual subhalo. In correlating colour with formation time, we are following an approach similar to Hearin & Watson (2013).

As a reference mock, Rosella will be useful for fulfilling tasks that include analysing galaxy survey biases and calibrating approximate mocks that can scale up the galaxy population data in Rosella to meet volume and abundance requirements.

The mock presented here may be useful for low-redshift galaxy surveys that could benefit from a $z \sim 0.2$ reference mock. The method behind Rosella can further be used to generate galaxy catalogues at other redshifts. Should one need a light-cone catalogue with galaxies populated with the Rosella method, one could populate other snapshots in the P-Millennium simulation and produce a light-cone from the resulting suite of reference mocks that correspond to fully populated boxes of the P-Millennium volume. The method used here can thus benefit any survey that probes volumes similar to those covered by P-Millennium.

Compared to the HOD-based mock presented in Smith et al. (2017), the Rosella mock includes a greater degree of assembly bias by construction from the v_{peak}-based SHAM method for luminosity assignment and a colour assignment method that relies on each galaxy's individual subhalo history. Rosella connects the simulation-only properties that are not directly observable, such as halo and subhalo mass, to directly observable quantities, M_r^h brightness and ${}^{0.1}(g-r)$ colour. This opens the possibility of using Rosella and the method behind it to search for evidence of assembly bias in galaxy surveys that probe volumes similar to Rosella's.

We evaluate the closeness of the match between our mock and real data by comparing the luminosity- and colour-dependent clustering of our mock's galaxies against previously published clustering of similar galaxy populations in existing observational and mock data. Users of Rosella and its method may be interested in other summary statistics, e.g. redshift-space distortions.

The tuning of the mock for specific scientific goals may adjust the choice of free parameters in the creation of our data, such as the functional form and parameters in the luminosity-dependent scatter added to the M_r^h data, as well as the definition of z_{form}. While we have considered two values of f in relating subhalo v_{max} histories to z_{form} and found that the two options did not have a significant effect on colour-dependent clustering, other formulations of z_{form} might be possible and may suit specific scentific goals.

To put constraints on cosmological parameters using Rosella's linking of observable and simulation-based galaxy qualities, such as luminosity and halo mass, error covariances need to be determined. This requires the use of many mock catalogues, of the order of up to 10^4 and greater (e.g. White et al. 2014; Kitaura et al. 2016; Villaescusa-Navarro et al. 2020). This can be achieved by calibrating fast mock generation methods using the reference mock presented here and, potentially, doing so at a variety of redshifts by applying the Rosella method to a variety of P-Millennium snapshots.

ACKNOWLEDGEMENTS

SS has received funding from the U.S.–U.K. Fulbright Commission and the Gruber Science Fellowship. PN and SMC acknowledge

the support of the Science and Technology Facilities Council (ST/L00075X/1 and ST/P000541/1).

The authors acknowledge Alex Smith, Jeremy Tinker, and Eduardo Rozo for insightful comments. The authors thank the groups behind Guo et al. (2015), Smith et al. (2017), and Zehavi et al. (2011) for providing data points from their publications.

This research was supported by the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy (DOE) under Contract No. DE-AC02-05CH1123, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract; additional support for Dark Energy Spectroscopic Instrument (DESI) is provided by the U.S. National Science Foundation, Division of Astronomical Sciences under Contract No. AST-0950945 to the National Science Foundation's (NSF's) National Optical-Infrared Astronomy Research Laboratory; the Science and Technologies Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Science and Technology of Mexico; the Ministry of Economy of Spain; and by the DESI Member Institutions. The authors are honored to be permitted to conduct astronomical research on Iolkam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation.

This work used the COSMA Data Centric system at Durham University, operated by the Institute for Computational Cosmology on behalf of the Science and Technology Facilities Countil Distributed Research utilising Advanced Computing (STFC DiRAC) High Performance Computing Facility (www.dirac.ac.uk).

DATA AVAILABILITY

The data underlying this article were accessed from the COSMA Data Centric system at Durham University (www.dirac.ac.uk). The derived data generated in this research will be shared on reasonable request to the corresponding author.

REFERENCES

Baldry I. K., Glazebrook K., Brinkmann J., Ivezić Ž., Lupton R. H., Nichol R. C., Szalay A. S., 2004, ApJ, 600, 681
Baugh C. M., 2006, Rep. Prog. Phys., 69, 3101
Baugh C. M. et al., 2019, MNRAS, 483, 4922
Behroozi P. S., Wechsler R. H., Wu H.-Y., Busha M. T., Klypin A. A., Primack J. R., 2012, Astrophysics Source Code Library, record ascl:1210.011
Benson A. J., Cole S., Frenk C. S., Baugh C. M., Lacey C. G., 2000, MNRAS, 311, 793
Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
Berlind A. A. et al., 2003, ApJ, 593, 1
Blanton M. R. et al., 2003, ApJ, 592, 819
Blanton M. R. et al., 2017, AJ, 154, 28
Boylan-Kolchin M., Springel V., White S. D. M., Jenkins A., Lemson G., 2009, MNRAS, 398, 1150
Chaves-Montero J., Hearin A., 2020, MNRAS, 495, 2088
Chaves-Montero J., Angulo R. E., Schaye J., Schaller M., Crain R. A., Furlong M., Theuns T., 2016, MNRAS, 460, 3100
Cole S., Aragon-Salamanca A., Frenk C. S., Navarro J. F., Zepf S. E., 1994, MNRAS, 271, 781
Cole S., Hatton S., Weinberg D. H., Frenk C. S., 1998, MNRAS, 300, 945
Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, MNRAS, 319, 168
Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, 201
Contreras S., Angulo R., Zennaro M., 2021, MNRAS, preprint (arXiv:2005.03672)
Cooray A., 2006, MNRAS, 365, 842
Crain R. A. et al., 2015, MNRAS, 450, 1937
Croton D. J. et al., 2016, ApJS, 222, 22
Dawson K. S. et al., 2013, AJ, 145, 10
DeRose J. et al., 2019, preprint (arXiv:1901.02401)
DESI Collaboration, 2016, DESI Final Design Report Part I: Science,Targeting, and Survey Design, preprint (arXiv:1611.00036)
DESI Collaboration, 2018, Technical Report v1.0, DESI Cosmological Simulations Requirements Document. Dark Energy Spectroscopic Instrument
Desmond H., Wechsler R. H., 2015, MNRAS, 454, 322
Efstathiou G., Sutherland W. J., Maddox S. J., 1990, Nature, 348, 705
Eisenstein D. J., Blanton M., Zehavi I., Bahcall N., Brinkmann J., Loveday J., Meiksin A., Schneider D., 2005a, ApJ, 619, 178
Eisenstein D. J. et al., 2005b, ApJ, 633, 560
Gao L., White S. D. M., 2007, MNRAS, 377, L5
Gao L., Springel V., White S. D. M., 2005, MNRAS, 363, L66
Gonzalez-Perez V., Lacey C. G., Baugh C. M., Lagos C. D. P., Helly J., Campbell D. J. R., Mitchell P. D., 2014, MNRAS, 439, 264
Guo H. et al., 2015, MNRAS, 453, 4368
Guo H. et al., 2016, MNRAS, 502, 3599
Hearin A. P., 2015, MNRAS, 451, L45
Hearin A. P., Watson D. F., 2013, MNRAS, 435, 1313
Hearin A. P., Watson D. F., Becker M. R., Reyes R., Berlind A. A., Zentner A. R., 2014, MNRAS, 444, 729
Hearin A. P., Zentner A. R., van den Bosch F. C., Campbell D., Tollerud E., 2016, MNRAS, 460, 2552
Ivezić Ž. et al., 2019, ApJ, 873, 111
Kauffmann G., White S. D. M., Guiderdoni B., 1993, MNRAS, 264, 201
Khandai N., Di Matteo T., Croft R., Wilkins S., Feng Y., Tucker E., DeGraf C., Liu M.-S., 2015, MNRAS, 450, 1349
Kitaura F.-S. et al., 2016, MNRAS, 456, 4156
Kravtsov A. V., Berlind A. A., Wechsler R. H., Klypin A. A., Gottlo S., Allgood B., Primack J. R., 2004, ApJ, 601, 35
Kulier A., Ostriker J. P., 2015, MNRAS, 452, 4013
Lacey C. G. et al., 2016, MNRAS, 462, 3854
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Lehmann B. V., Mao Y.-Y., Becker M. R., Skillman S. W., Wechsler R. H., 2017, ApJ, 834, 37
Loveday J. et al., 2012, MNRAS, 420, 1239
Mandelbaum R., Wang W., Zu Y., White S., Henriques B., More S., 2016, MNRAS, 457, 3200
Marín F. A., Wechsler R. H., Frieman J. A., Nichol R. C., 2008, ApJ, 672, 849
Masaki S., Hikage C., Takada M., Spergel D. N., Sugiyama N., 2013a, MNRAS, 433, 3506
Masaki S., Lin Y.-T., Yoshida N., 2013b, MNRAS, 436, 2286
McCullagh N., Norberg P., Cole S., Gonzalez-Perez V., Baugh C., Helly J., 2017, preprint (arXiv:1705.01988)
Mo H., van den Bosch F. C., White S., 2010, Galaxy Formation and Evolution. Cambridge Univ. Press, Cambridge
Naiman J. P. et al., 2018, MNRAS, 477, 1206
Nelson D. et al., 2018, MNRAS, 475, 624
Norberg P. et al., 2001, MNRAS, 328, 64
Norberg P., 2dFGRS Team, 2002, in Metcalfe N., Shanks T., eds, ASP Conf. Ser. Vol. 283, A New Era in Cosmology. Astron. Soc. Pac., San Francisco, p. 47
Peacock J. A., Smith R. E., 2000, MNRAS, 318, 1144
Perlmutter S. et al., 1999, ApJ, 517, 565
Pillepich A. et al., 2018, MNRAS, 475, 648
Planck Collaboration XVI, 2014, A&A, 571, A16
Planck Collaboration X, 2020, A&A, 641, A10
Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, ApJ, 771, 30
Riess A. G. et al., 1998, AJ, 116, 1009
Ruiz-Macias O. et al., 2020, Res. Notes Am. Astron. Soc., 4, 187
Safonova A., 2019, Master's thesis, Durham University
Sinha M., Garrison L., 2017, Astrophysics Source Code Library, record ascl:1703.003

Sinha M., Garrison L., 2019, in Majumdar A., Arora R., eds, Software Challenges to Exascale Computing. Springer, Singapore, p. 3
Skibba R. A., Sheth R. K., 2009, MNRAS, 392, 1080
Smith A., Cole S., Baugh C., Zheng Z., Angulo R., Norberg P., Zehavi I., 2017, MNRAS, 470, 4646
Smith A. et al., 2020, MNRAS, 499, 269
Somerville R. S., Primack J. R., 1999, MNRAS, 310, 1087
Springel V. et al., 2005, Nature, 435, 629
Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726
Tasitsiomi A., Kravtsov A. V., Wechsler R. H., Primack J. R., 2004, ApJ, 614, 533
Vale A., Ostriker J. P., 2004, MNRAS, 353, 189
Villaescusa-Navarro F. et al., 2020, ApJS, 250, 2
Wechsler R. H., Zentner A. R., Bullock J. S., Kravtsov A. V., Allgood B., 2006, ApJ, 652, 71
White M., Tinker J. L., McBride C. K., 2014, MNRAS, 437, 2594
White S. D. M., Frenk C. S., 1991, ApJ, 379, 52
Xu X., Zheng Z., 2020, MNRAS, 492, 2739
Yamamoto M., Masaki S., Hikage C., 2015, preprint (arXiv:1503.03973)
Yang X., Mo H. J., van den Bosch F. C., 2003, MNRAS, 339, 1057
Yang X., Mo H. J., van den Bosch F. C., 2008, ApJ, 676, 248
Zehavi I. et al., 2005, ApJ, 630, 1
Zehavi I. et al., 2011, ApJ, 736, 59
Zehavi I., Kerby S. E., Contreras S., Jiménez E., Padilla N., Baugh C. M., 2019, ApJ, 887, 17

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