DEVELOPING AN ACCURATE PROBE OF THE GALAXY-HALO CONNECTION: BARYONIC EFFECTS, SMALL-SCALE GALAXY CLUSTERING, AND HALO MODEL EXTENSIONS

By
Gillian Dora Beltz-Mohrmann

Dissertation

Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in
Astrophysics

May 13, 2022
Nashville, Tennessee

Approved:

Andreas Berlind, Ph.D.

Kelly Holley-Bockelmann, Ph.D.

Robert Scherrer, Ph.D.

Manodeep Sinha, Ph.D.

David Weintraub, Ph.D.

Copyright © 2022 Gillian Dora Beltz-Mohrmann
All Rights Reserved

For my parents, who have given me everything.

ACKNOWLEDGMENTS

None of this would have been possible without the support of many people. First, I would like to thank my advisor, Andreas Berlind, not only for his scientific and academic mentorship, but also his constant support and encouragement over the past six years. I would not be where I am today without you. I would also like to thank my committee for their support throughout this process. In particular, I would like to thank Manodeep Sinha, for lending his expertise from an ocean away, and always with a smile on his face. I would also like to thank Ferah Munshi, for being a great mentor and a wonderful friend, and for all of her support, especially this past year.

I would like to thank my friends at Vanderbilt for their comraderie throughout this entire experience. I specifically thank the large-scale structure group, especially Antonio for being my conference buddy, and Mehnaaz and Abbie for being on this journey with me every step of the way. I would also like to thank Kate, for making my first two years in Nashville so fun, and for her continued friendship even after we no longer shared an apartment. I cannot imagine commiserating over these historic times with anyone else. Finally, I would like to thank Glenna for her constant friendship and support, and Stephanie, who in addition to being a wonderful friend led me to my dog.

Speaking of whom, I would like to thank Gadget, for being a perpetual source of entertainment, and for reminding me that it's important to take breaks from working to go outside and run around (and maybe roll in some grass). I would like to my best friends, Sormeh and Sarah, with whom I can laugh, cry, complain, or celebrate with at any time of day or night. I would like to thank my parents for believing in me and supporting me through all of the highs and lows these past six years. Most of all, I thank Adam for being a true partner in every way. I could not have done this without you.

This work has been supported by the National Science Foundation (NSF) through Award (AST-1909631), and has made use of NASA's Astrophysics Data System; PYTHON (https://www.python.org/); MATPLOTLIB, a Python library for publication quality graphics (Hunter, 2007); SCIPY (Jones et al., 2001; Virtanen et al., 2020); the IPYTHON package (Pérez and Granger, 2007); ASTROPY, a community-developed core Python package for Astronomy (Astropy Collaboration et al., 2018); NUMPY (Van Der Walt et al., 2011; Harris et al., 2020); PANDAS (McKinney, 2010, 2011), and CHAINCONSUMER (Hinton, 2016).

Some of the computational facilities used in this project were provided by the Vanderbilt Advanced Computing Center for Research and Education (ACCRE). Parts of this research were conducted by the Australian Research Council Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), through project number CE170100013. The MCMCs in this work were run on the Texas Advanced Computing Center's Stampede2 supercomputer. The mock catalogs used in this work were produced by the LasDamas project (http://lss.phy.vanderbilt.edu/lasdamas/); we thank NSF XSEDE for providing the computational resources for LasDamas. We acknowledge the Virgo Consortium for making their simulation data available. The EAGLE simulations were performed using the DiRAC-2 facility at Durham, managed by the ICC, and the PRACE facility Curie based in France at TGCC, CEA, Bruyères-le-Châtel.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/. The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

We especially thank our anonymous referees for constructive comments that helped us improve our work. These acknowledgements were compiled using the Astronomy Acknowledgement Generator (http://astrofrog.github.io/acknowledgment-generator/).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

Introduction

The standard cosmological model, ΛCDM, asserts that baryonic matter (i.e., normal matter that makes up gas and stars) makes up only $\sim 16\%$ of the total mass in the Universe. The rest of the mass in the Universe is made of a mysterious substance known as "dark matter." Because baryonic matter interacts through not only gravity but also electromagnetic forces, it is directly observable. Dark matter, however, interacts only through gravity, and therefore cannot be directly observed. However, because dark matter makes up the majority of the mass in the Universe, it is the primary driver of large-scale structure formation.

Over the past 14 billion years, the Universe has evolved from a state of nearly uniform density into a complex web of structure. Slight perturbations in the initial dark matter density field of the Universe shortly after the Big Bang grew over time via gravity, resulting in huge variations in the density field today. What were once small underdense pockets evolved into vast cosmic voids, while slightly dense pockets ultimately became massive overdense regions of dark matter known as "dark matter halos." These dark matter halos are the hosts in which galaxies ultimately form and reside. Thus, despite being unable to directly observe dark matter, we can use observations of the clustering of galaxies (which are directly observable) to build a map of how matter in the Universe is distributed today. By comparing observations of galaxy clustering to our models of structure formation in the Universe, we can gain a better understanding of the physics that governed the initial conditions and subsequent evolution of the Universe.

On very large physical scales, where galaxies are simple biased tracers of the underlying dark matter distribution, our cosmological model is able to accurately predict the galaxy clustering that we observe (Scherrer and Weinberg, 1998). On small physical scales, however, galaxy clustering is affected both by our cosmological model and by the complex physics of galaxy formation and evolution, which is not well understood. Thus, it is difficult to test our cosmological models of structure formation on small scales without first developing an accurate picture of the connection between galaxies and the dark matter halos in which they reside. The goal of this dissertation is to develop an accurate probe of this "galaxy-halo connection" in order to improve both our understanding of galaxy formation and ultimately test our cosmological model.

1.1 ΛCDM Cosmology and the Expanding Universe

The standard cosmological model, ΛCDM, postulates that the Universe is comprised of $\sim 69\%$ dark energy (Ω_{Λ}) and $\sim 31\%$ matter ($\Omega_{\rm M}$), of which $\sim 84\%$ is Cold Dark Matter (CDM) and $\sim 16\%$ is normal (baryonic) matter (Planck Collaboration et al., 2020). In this paradigm, dark energy is the negative vacuum pressure causing the accelerated expansion of the Universe, and cold dark matter is presumed to be a non-relativistic particle which is only affected by the force of gravity. ΛCDM also asserts that the geometry of the Universe is flat, and that the Universe began with the Big Bang, followed by a brief inflationary period of rapid expansion. Observations of the Cosmic Microwave Background radiation have provided very tight constraints on this standard model of cosmology (Penzias and Wilson, 1965; Dicke et al., 1965; Spergel et al., 2003).

The fact that the Universe is expanding was discovered through observations of redshifted galaxies by Edwin Hubble in the 1920s (Hubble, 1929). If the Universe were only composed of normal matter, we would expect gravity to gradually slow the expansion of the Universe over time. Recently, however, observations of distant Type Ia Supernovae confirmed that the expansion of the Universe is in fact accelerating, leading us to discover the presence of dark energy (Riess et al., 1998; Perlmutter et al., 1999). Meanwhile, observations of rotating galaxies led us to discover the existence dark matter (Rubin et al., 1980). While we have not as of yet detected a dark matter particle, a variety of observations indicate that dark matter is cold (moves slowly, i.e. at non-relativistic speeds) and collisionless (i.e., only interacts strongly through the force of gravity) (e.g. Clowe et al., 2006).

1.1.1 Growth of Structure

One of the primary strengths of ΛCDM is its ability to predict the large-scale structure that we observe in the Universe today. Shortly after the Big Bang, the Universe had an almost uniform density. Under such conditions, large-scale structure formation would have been nearly impossible. However, quantum fluctuations in the primordial Universe led to small pockets of over- and underdense regions. At the moment inflation began, these small density perturbations grew rapidly, and over time became the large structures that we see today. Underdense regions became large cosmic voids, while overdense regions collapsed to become gravitationally-bound dark matter halos, with overdensities ~ 200 times the mean density of the Universe. Over time, baryonic matter coalesced within these dark matter halos to form galaxies.

1.2 Probing Cosmology with Galaxy Surveys

There are several observational probes that we can use to constrain our cosmological model, one of which is calculating the positions of nearby galaxies to build a map of the distribution of galaxies in the local Universe. By comparing observations of galaxy clustering in the Universe to predictions of large-scale structure from ΛCDM, we can simultaneously constrain both the parameters of our cosmological model and our understanding of the physics that governs structure formation.

Mapping the distribution of galaxies in the observable Universe requires determining the distances to thousands of galaxies. This is primarily done through galaxy redshift surveys (e.g., York et al., 2000; Colless et al., 2001; Jones et al., 2004; Dawson et al., 2013). These surveys involve measuring the spectra of thousands of galaxies in order to calculate a redshift, z, for each galaxy. As an example, Figure 1.1 depicts the spatial distribution of galaxies observed in the Sloan Digital Sky Survey (SDSS York et al., 2000). Once we have a map of the positions of many galaxies, we can measure a multitude of statistics to

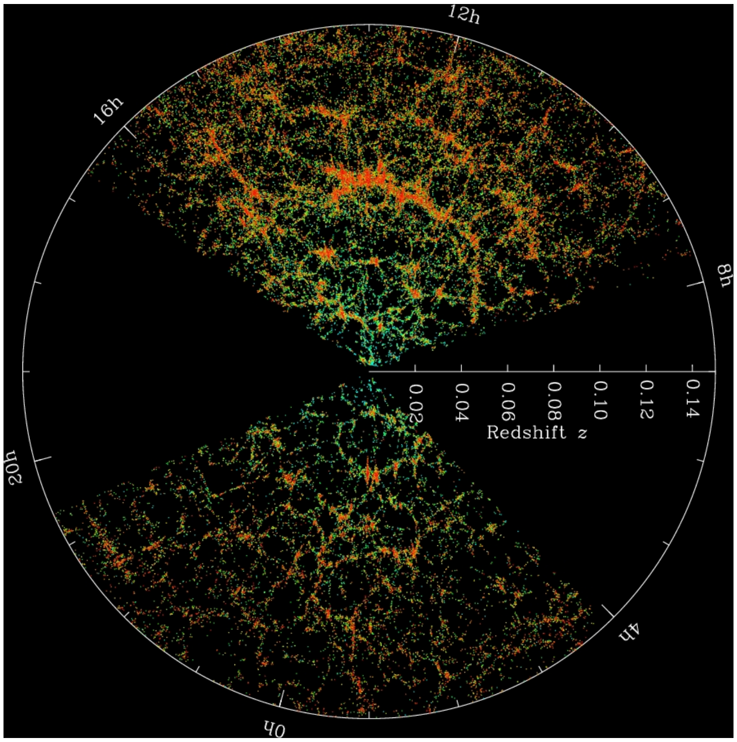

Figure 1.1: Large-scale structure in the Sloan Digital Sky Survey. Each point is a galaxy, and the color of each point corresponds to the (g-r) color of the galaxy. Earth is located at the center. Image Credit: https://www.sdss.org/science/orangepie/; M. Blanton and SDSS.

quantify the clustering of these galaxies. These statistics include things like the number of pairs of galaxies separated by a certain distance, the number of groups of galaxies of a certain size, the number of voids of a given size, etc.

1.2.1 Redshift-space distortions, k-corrections, and volume-limited samples

The redshift of the galaxy is a result of both the expansion of the Universe ($z_{\text{cosmological}}$) and the radial peculiar velocity of the galaxy (z_{doppler}). Thus, the redshift calculated from the galaxy's spectrum is the product of these two effects: $1+z = (1+z_{\text{cosmological}})(1+z_{\text{doppler}})$. At low redshift, the distance to the galaxy is approximated as $d = cz/H_0$, where c is the speed of light and H_0 is the Hubble constant ($\sim 70(\text{km/s})/\text{Mpc}$). The contribution of the galaxy's peculiar velocity to the redshift (and subsequently the distance calculated) is known as a "redshift-space distortion." Redshift-space distortions have a significant impact on the measured distances of galaxies, particularly at very low redshift, where the galaxy's velocity contributes significantly to the observed redshift of the galaxy.

Galaxy absolute magnitudes are typically measured through a single filter, which spans only part of the electromagnetic spectrum. In the Sloan Digital Sky Survey, for example, the filters used are *u,g,r,i,* and *z*, each of which covers a specific wavelength range between 300 and 1100nm. Galaxy magnitudes are typically determined using one particular filter (e.g., the *r* filter). However, because each galaxy has a different redshift, the part of the total spectrum of the galaxy that is observed in the *r* filter differs for each galaxy, which has an effect on the absolute magnitude determination of the galaxy. Thus, galaxy absolute magnitudes at different redshifts cannot be directly compared without first correcting for the effect of the redshift on the observed magnitude. This correction, known as a "k-correction," converts the absolute magnitudes of all galaxies into their rest frame absolute magnitude at a particular redshift (e.g., $z = 0.1$).

Because bright objects can be observed at farther distances than faint objects, our map of the distribution of galaxies is biased toward brighter galaxies at farther distances. Because different types of galaxies cluster in different ways, this selection bias (known as "Malmquist bias") can have a significant impact on galaxy clustering measurements. In order to measure galaxy clustering in a way that is unaffected by our observational limitations, we must construct so-called "volume-limited samples" of galaxies. These samples contain *all* of the galaxies brighter than a certain absolute magnitude threshold (in a particular filter) within a given redshift range. In other words, the sample is "complete" within the redshift range of interest. A lower redshift sample will be complete down to a lower magnitude limit than a high-redshift sample, but within the redshift limits of the sample, there will be no bias toward brighter galaxies at farther distances.

1.3 Modeling Galaxy Clustering

Galaxy redshift surveys provide us with a map of the distribution of galaxies in the nearby Universe. By comparing this map to our predictions of large-scale structure formation, we can use galaxy surveys to test our cosmological model. While we know that structure formation is primarily driven by dark matter, we also know that, in general, galaxies tend to form and reside within dark matter halos. Particularly on very large scales, it is safe to assume that galaxies are simple biased tracers of the underlying dark matter distribution. Thus, for a given set of cosmological parameters, we can model the evolution of dark matter structure in the Universe, and compare this prediction on large scales to the observed galaxy clustering from redshift surveys. This has been used to measure things like the Baryon Acoustic Oscillation (BAO) signature to constrain ΛCDM (Eisenstein et al., 2005).

In general, ΛCDM has been tested very thoroughly on large scales, and seems to produce very accurate predictions of large scale structure. However, ΛCDM has *not* been thoroughly tested on small scales. Small-scale galaxy clustering therefore has the potential to be a powerful new probe of cosmological models. Unfortunately, using small-scale clustering to test ΛCDM is a challenge because on these scales ($\lesssim 10\ h^{-1}$Mpc), the spatial distribution of galaxies is affected not only by our cosmological model but also by all of the complex physics of galaxy formation and evolution. In other words, on these scales, galaxies are no longer simple tracers of the underlying dark matter distribution. Thus, studying the connection between galaxies and the dark matter halos in which they reside is key to using small-scale clustering to constrain cosmological models, as well as understanding galaxy formation and evolution.

1.3.1 N-body Simulations

First and foremost, modeling small-scale galaxy clustering requires highly accurate predictions of dark matter structure formation on small scales. At early times, the growth of structure in the Universe can be accurately predicted analytically to first-order using linear theory (i.e. the Zel'dovich approximation, Zel'Dovich, 1970), or to second-order using 2nd Order Lagrangian Perturbation Theory (2LPT Scoccimarro, 1998; Crocce et al., 2006). However, on small scales, structure formation becomes highly nonlinear, and cannot be predicted accurately today using linear theory or 2LPT. Instead, we can use 2LPT to predict the density field at very high redshift (e.g., $z = 99$), and then run an N-body simulation to model the subsequent evolution of structure on all scales.

Because dark matter comprises $\sim 85\%$ of the matter in the Universe, these simulations frequently only involve dark matter. Such "dark matter only" (DMO) simulations are useful because they allow us to predict the large-scale distribution of dark matter as well as the statistical properties of dark matter halos in the Universe (i.e., the number of halos of a given mass, the clustering of halos, etc.) (Springel et al., 2005). These simulations allow us to investigate our cosmological model, our assumptions about the initial conditions of the Universe, and our understanding of the nature of dark matter, dark energy, and gravity, without having to model complex baryonic physics. However, because we cannot directly observe dark matter in the Universe, it is difficult to compare the output of a DMO simulation to our observations of galaxy clustering without having a way to connect galaxies to the dark matter distribution.

1.3.2 Hydrodynamic Simulations

Another type of simulation which can be used to model both dark matter structure evolution *and* galaxy formation is a hydrodynamic simulation (e.g. Vogelsberger et al., 2014a). Hydrodynamic simulations involve dark matter, as well as baryonic matter. Thus, the physics involved includes gravity, as well as complicated baryonic physics to model things like star formation, black hole formation, supernova feedback, and feedback from active galactic nuclei (AGN). Running these simulations involves specifying the values of hundreds of parameters to regulate these various physical processes. Additionally, using these simulations to model galaxy clustering, even on small scales, requires simulating structure formation in large volumes at high-resolution. Running large, high-resolution hydrodynamic simulations is much more computationally expensive than running DMO simulations, owing to the more complex physics calculations involved. Furthermore, currently there is no con-

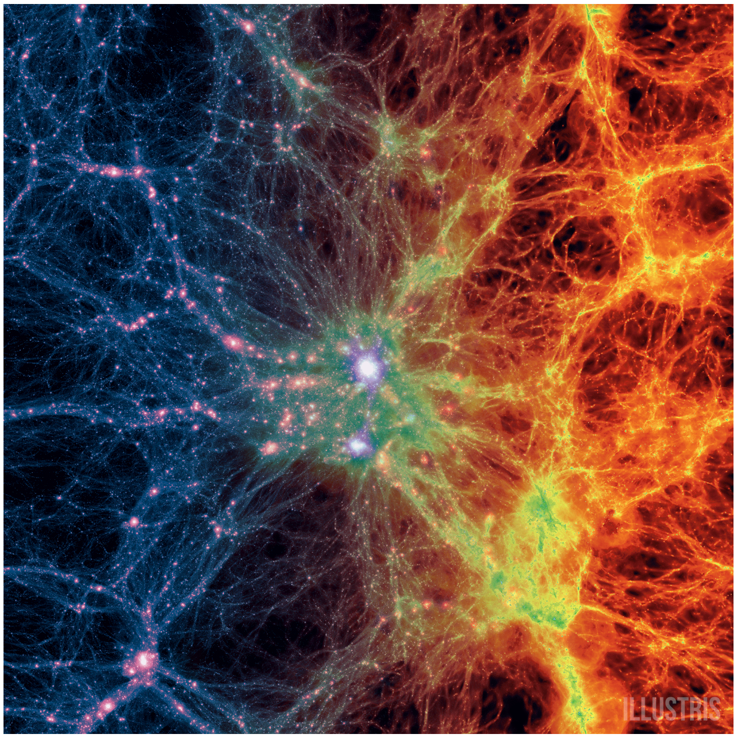

Figure 1.2: A massive cluster in the Illustris hydrodynamic simulation at $z = 0$. The left-hand side shows the dark matter density, and the right-hand side shows the gas density in the simulation. Image Credit: Illustris Collaboration / Illustris Simulation (Vogelsberger et al., 2014a,b).

sensus on the correct physics prescriptions to use. Ideally, we would like to run these simulations many times, while varying cosmological and baryonic physics parameters, in order to explore a large parameter space and constrain our model. However, the expensive and complex nature of hydrodynamic simulations makes them ill-suited for such an exercise. Thus, using hydrodynamic simulations is currently not a viable strategy for modeling small-scale galaxy clustering.

1.3.3 Halo Models

An alternative to modeling galaxy clustering with hydrodynamic simulations is to employ a so-called "halo model" (Neyman and Scott, 1952; Peebles, 1974; McClelland and Silk, 1977; Scherrer and Bertschinger, 1991; Kauffmann et al., 1997, 1999; Baugh et al., 1999; Jing et al., 1998; Benson et al., 2000; Ma and Fry, 2000; Peacock and Smith, 2000; Seljak, 2000; Scoccimarro et al., 2001; Sheth et al., 2001; White et al., 2001; Cooray and Sheth, 2002). These models rely on the assumption that the clustering of galaxies can be fully described by (i) the clustering of their host halos and (ii) the way in which galaxies occupy these halos. By assuming that galaxies form and reside within dark matter halos, this empirical approach allows us to use a few free parameters to connect galaxies to halos in order to match clustering observations. This strategy involves first running a DMO simulation with adequate volume and resolution to model structure formation on small scales, then employing a halofinding algorithm (e.g. Lacey and Cole, 1994; Behroozi et al., 2013) to identify dark matter halos, and finally using a halo model to connect galaxies to the dark matter distribution. Thus, we can quantitatively model galaxy clustering on small scales while bypassing the need for a complete understanding of galaxy formation physics.

One popular flavor of halo model is the Halo Occupation Distribution (HOD) model (Berlind and Weinberg, 2002; Berlind et al., 2003). The standard form of this model assigns a number of galaxies to a halo of mass M using five free parameters which only depend on the halos mass Zheng et al. (2005). This model also contains prescriptions that specify the relative spatial and velocity distributions of galaxies and dark matter within halos. This type of HOD model has become the "standard" in halo modeling studies. Because the standard HOD model contains only a few free parameters, it is simple and computationally feasible to vary these parameters in order to match the clustering of observed galaxies.

The DMO simulation + standard HOD model approach has many advantages for modeling small-scale galaxy clustering, chief among which are its computational feasibility and its ability to successfully reproduce several frequently used galaxy clustering statistics (Sinha et al., 2018). This "forward modeling" approach takes a few free parameters and produces predictions of galaxy clustering that can be directly compared to actual observations of galaxy clustering from redshift surveys. However, this approach has several limitations as well. For example, it relies heavily on the assumption that DMO simulations produce the correct number of halos of different masses (i.e. the correct "halo mass function") as well as the correct clustering of halos. It also assumes that the connection between galaxies and their halos can be described via a few free parameters which only depend on the mass of the dark matter halo.

1.4 Summary

The goal of the work presented in this dissertation is to develop an accurate model of the galaxy-halo connection, through a combination of studies performed on hydrodynamic simulations and analyses of small-scale galaxy clustering in the Sloan Digital Sky Survey. In Chapter 2, I use hydrodynamic simulations of galaxy formation to investigate the extent to which the assumptions of the standard HOD model can affect galaxy clustering statistics. In Chapter 3, I investigate the impact of baryonic physics on the halo population in three different hydrodynamic simulations. In Chapter 4, I add flexibility to the standard Halo Occupation Distribution model to constrain the galaxy-halo connection in the Sloan Digital Sky Survey using a combination of small-scale galaxy clustering statistics. Finally, in Chapter 5, I provide a summary and a discussion of future work.

CHAPTER 2

Testing the Accuracy of Halo Occupation Distribution Modelling using Hydrodynamic Simulations

This chapter was previously published in the February 2020 edition of Monthly Notices of the Royal Astronomical Society[1] *(Beltz-Mohrmann et al., 2020) and is reproduced here, with minor formatting changes, with the permission of the publisher and my co-authors, Andreas A. Berlind and Adam O. Szewciw.*

Halo models provide a simple and computationally inexpensive way to investigate the connection between galaxies and their dark matter haloes. However, these models rely on the assumption that the role of baryons can be easily parametrized in the modelling procedure. We aim to examine the ability of halo occupation distribution (HOD) modelling to reproduce the galaxy clustering found in two different hydrodynamic simulations, Illustris and EAGLE. For each simulation, we measure several galaxy clustering statistics on two different luminosity threshold samples. We then apply a simple five parameter HOD, which was fit to each simulation separately, to the corresponding dark matter only simulations, and measure the same clustering statistics. We find that the halo mass function is shifted to lower masses in the hydrodynamic simulations, resulting in a galaxy number density that is too high when an HOD is applied to the dark matter only simulation. However, the exact way in which baryons alter the mass function is remarkably different in the two simulations. After applying a correction to the halo mass function in each simulation, the HOD is able to accurately reproduce all clustering statistics for the high luminosity sample of galaxies. For the low luminosity sample, we find evidence that in addition to correcting the halo mass function, including spatial, velocity, and assembly bias parameters in the HOD is necessary to accurately reproduce clustering statistics.

[1] Because this chapter was originally published in a journal in the UK, British English spelling conventions are used in this chapter.

2.1 Introduction

Studying the connection between galaxies and the dark matter haloes in which they reside is one of the keys to understanding galaxy formation and evolution, as well as constraining cosmological models. In recent years, using hydrodynamic simulations has become a popular method for investigating this connection (e.g. Vogelsberger et al., 2014a). However, these simulations are computationally expensive, and are thus ill-suited for exploring a large parameter space. Moreover, different hydrodynamic simulations produce different results; we currently lack a consensus on the correct gas physics prescriptions to use.

By contrast, dark matter only simulations are much less computationally expensive, and although the only physics involved is gravity, they still allow us to predict the large-scale distribution of dark matter as well as the statistical properties of dark matter haloes in the Universe. One can then adopt an empirical rather than an *ab-initio* approach and employ a halo model in order to connect galaxies to the dark matter distribution. Halo models are a broad class of models based on the assumption that galaxies form and live inside dark matter haloes. With a few free parameters that can be fit to clustering observations, one can connect galaxies to haloes, thus quantitatively modelling galaxy clustering on small scales while bypassing the need for a complete understanding of galaxy formation physics.

The earliest halo models to describe galaxy clustering were the analytic models of Neyman and Scott (1952), Peebles (1974), and McClelland and Silk (1977). Later, Kauffmann et al. (1997, 1999), and Baugh et al. (1999) showed that semi-analytic models could be used to predict galaxy clustering by combining the results from N-body simulations with theories for the formation and evolution of galaxies within haloes. Soon thereafter, Jing et al. (1998) and Benson et al. (2000) found that galaxy clustering merely depends on halo occupation statistics as a function of halo mass, potentially sidestepping the need to model galaxy formation altogether. Subsequently, several papers (e.g. Ma and Fry, 2000; Peacock and Smith, 2000; Seljak, 2000; Scoccimarro et al., 2001; Sheth et al., 2001; White et al., 2001; Cooray and Sheth, 2002) expanded on the work of Scherrer and Bertschinger (1991) to combine both halo properties and occupation statistics to successfully predict the galaxy correlation function and power spectrum.

A key ingredient of the halo model is the Halo Occupation Distribution (HOD), which defines the bias of a population of galaxies by the conditional probability that a dark matter halo of virial mass M contains N galaxies, together with prescriptions that specify the relative spatial and velocity distributions of galaxies and dark matter within haloes (Berlind and Weinberg, 2002; Berlind et al., 2003). These relations can be parametrized with various degrees of freedom. However, most studies have used simple formulations of the HOD, with at most five free parameters that specify the mean occupation number of galaxies, along with the assumptions that galaxies trace dark matter inside haloes. This type of HOD model, as proposed by Zheng et al. (2005), has become the 'standard' in halo modelling studies.

Halo models have been used to model galaxy clustering in many galaxy redshift surveys, including the Sloan Digital Sky Survey (SDSS; York et al., 2000), the 2dF Galaxy Redshift Survey (2dFGRS; Colless et al., 2001), the 6dF Galaxy Redshift Survey (6dfGRS; Jones et al., 2004), and the SDSS III Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al., 2013). Many studies have used halo models to investigate the two-point correlation function of both low redshift galaxies (e.g. Magliocchetti and Porciani, 2003; Zehavi et al., 2004; Collister and Lahav, 2005; Tinker et al., 2005; Zehavi et al., 2005, 2011; Watson et al., 2012; Beutler et al., 2013; Piscionere et al., 2015) as well as high redshift galaxies (e.g. Bullock et al., 2002; Moustakas and Somerville, 2002; Hamana et al., 2004; Zheng, 2004; Lee et al., 2006; Tinker et al., 2010; Jose et al., 2013; Kim et al., 2014) (as cited in Sinha et al. (2018)).

Some previous works (e.g. Zehavi et al., 2011) have found statistical tension between predictions of the halo model and the real Universe when fitting to galaxy clustering measurements in the SDSS. However, these works rely on analytic halo models that do not adequately control for systematic errors in the modelling procedure, making it difficult to interpret the goodness of fit results. Recently, Sinha et al. (2018) used a “fully numerical mock-based methodology” to test the standard ΛCDM + halo model against the clustering of SDSS DR7 galaxies. Their procedure carefully controlled for systematic errors, allowing them to interpret the goodness of fit of their model. They measured the projected correlation function, group multiplicity function, and galaxy number density, and found that while the model could successfully fit each statistic separately, it was unable to fit them simultaneously. Their best-fitting model was able to reproduce the clustering of low luminosity galaxies, but revealed a 2.3σ tension with the clustering of high luminosity galaxies, indicating a possible problem with the ‘standard’ HOD model.

There are several assumptions built into the standard HOD model that could be incorrect. First, the HOD framework relies on the assumption that cosmology and gravity alone govern the dark matter halo distribution. However, it has been shown that gas physics can also affect the properties of haloes (e.g. Cui et al., 2012; Bocquet et al., 2016). Second, the HOD typically assumes that the occupation of galaxies is solely based on halo mass, and does not depend on secondary halo properties like halo concentration or age. This ignores the possibility that galaxy clustering may be affected by the phenomenon known as assembly bias (Gao et al., 2005; Wechsler et al., 2006; Croton et al., 2007; Padilla et al., 2019; Salcedo et al., 2018; Xu and Zheng, 2018; Zehavi et al., 2018; Contreras et al., 2019). Finally, most HOD modelling assumes that galaxy positions and velocities within haloes trace the underlying distribution of dark matter.

Zentner et al. (2014) examined the extent to which the presence of assembly bias could lead to systematic errors in halo occupation statistics inferred from galaxy clustering. The authors constructed two sets of realistic mock galaxy catalogues with identical HODs: one with assembly bias and one with assembly bias removed. They then fit standard HODs to the galaxy clustering in each catalogue, and found that in the case where assembly bias was removed, the inferred HODs agreed with the true HODs, but when assembly bias was included, the inferred HODs showed significant systematic errors.

Hearin et al. (2016) introduced a new class of HOD models, known as 'decorated HODs', designed to incorporate parameters for assembly bias in halo occupation distribution models. The authors used these new models to characterize the impact of assembly bias on clustering statistics, and found that for SDSS-like samples, assembly bias can affect galaxy clustering by up to a factor of 2 on 200 kpc scales. They also found that on small scales ($r < 1$ Mpc) assembly bias generally enhances clustering, but on large scales it can either increase or decrease clustering. Vakili and Hahn (2019) and Zentner et al. (2019) applied this decorated HOD model to galaxies in the SDSS DR7 and found evidence of galaxy assembly bias for some luminosity samples.

Regarding the spatial distribution of galaxies within haloes, the HOD often uses random dark matter particles to assign positions and velocities to galaxies, or otherwise assumes a dark matter density profile for galaxies (e.g. Navarro et al., 1997, NFW). This does not account for the possibility that galaxies might not move like dark matter due to phenomena such as mergers, tidal stripping, and dynamical friction, leading to effects like spatial and velocity bias. Both Watson et al. (2012) and Piscionere et al. (2015) used halo models to predict the very small-scale clustering of galaxies in the SDSS, and found that more luminous galaxies do not trace underlying dark matter distributions of their haloes, indicating the presence of spatial bias. Guo et al. (2015b) looked at galaxy clustering in SDSS DR11 and found observational evidence for central velocity bias (i.e. that central galaxies on average are not at rest with respect to their host haloes) as well as satellite velocity bias (i.e. in this case, that luminous satellite galaxies move more slowly than the dark matter). In a subsequent paper, Guo et al. (2015a) modelled the projected and redshift-space two-point correlation functions of galaxies in SDSS DR7, and similarly found that luminous central galaxies and faint satellite galaxies exhibit velocity bias. Furthermore, they found that their measurements could be successfully interpreted within an extended HOD framework that includes central and satellite velocity bias parameters to describe the motions of galaxies within haloes.

Pujol and Gaztañaga (2014) investigated how well an HOD model could reproduce the two-point clustering of galaxies in several semi-analytic models, and found that the HOD failed to reconstruct the galaxy bias for low mass haloes, indicating the presence of assembly bias. They also found that clustering shows some dependence on the substructure of the host halo. Subsequently, Pujol et al. (2017) further compared the HOD model to semi-analytic models, and found that using local density rather than halo mass in the HOD model was a better predictor of galaxy bias.

In this paper we use hydrodynamic simulations of galaxy formation to investigate the extent to which all these built-in assumptions to the standard HOD model can affect galaxy clustering statistics. Although previous works (e.g. Artale et al., 2018; Bose et al., 2019) have used hydrodynamic simulations to investigate variations in halo occupancy with environment, concentration, and formation time, none have looked at the impact of the assumptions of the HOD on galaxy clustering statistics compared to clustering in hydrodynamic simulations. Additionally, previous works have not looked at a wide variety of clustering statistics, nor have they compared bias effects across multiple different hydrodynamic simulations.

In this work, we focus on two different hydrodynamic simulations, as well as two different luminosity threshold samples of galaxies. We measure several different galaxy clustering statistics on each of our samples. We then fit a five parameter HOD model to each simulation and sample, and apply these models to the corresponding dark matter only simulations. We then measure the same galaxy clustering statistics on our HOD galaxies as we did on our hydrodynamic galaxies. We examine the accuracy with which we can predict galaxy clustering using our HOD modelling framework, as compared to the full hydrodynamic simulations. Finally, we investigate how we might expand the HOD model to include effects like assembly, spatial, and velocity bias in order to increase the accuracy of the model. We note that our analysis strictly compares HOD modeling to hydrodynamic simulations and not to real galaxy surveys. Therefore, conclusions should not be drawn about the accuracy of the clustering produced either by the simulations or the HOD models as compared to real observations. However, the conclusions that we draw about the need to add freedom to HOD models are still valid.

We discuss our simulations in Section 2.2, and our halo model in Section 2.3. In Section 2.4 we discuss our clustering statistics, and in Section 2.5 we discuss the accuracy of our model. In Section 2.6 we discuss our halo populations, and in Section 2.7 we discuss possible extensions to our HOD model. Finally, in Section 2.8 we summarize our results and conclusions.

2.2 Simulations

We use two cosmological N-body simulations for our analysis: Illustris (Nelson et al., 2015; Vogelsberger et al., 2014b,a; Genel et al., 2014) and EAGLE (Schaye et al., 2015; McAlpine et al., 2016; The EAGLE team, 2017; Springel, 2005; Crain et al., 2015). The Illustris-2 simulation has a volume of $75^3(h^{-3}\text{Mpc}^3)$ and a dark matter particle mass of $3.5\times10^7(h^{-1}M_\odot)$. The EAGLE simulation (RefL100N1504) has a volume of $67.77^3(h^{-3}\text{Mpc}^3)$ and a dark matter particle mass of $6.6\times10^6(h^{-1}M_\odot)$. A summary of the simulation parameters can be found in Table 2.1.

Each of these hydrodynamic simulations has a corresponding dark matter only (DMO) counterpart, derived from the same cosmology and initial conditions. These two simulations are ideal for our analysis because they have high enough resolutions for the galaxies we are interested in, as well as large enough volumes to accurately measure clustering statistics out to $10h^{-1}$Mpc scales. We specifically choose to use Illustris-2 because the resolution of Illustris-3 is not quite high enough for our purposes, but the resolution of Illustris-1 is not necessary for the halo mass range that we are interested in. This is because in this work, the smallest haloes that we will ever populate with galaxies using our HOD model are on the order of $10^{11}(h^{-1}M_\odot)$. In Illustris-2-Dark, a halo of this size has about 2400 particles, so it is well-resolved. Additionally, such a small halo will only ever

Table 2.1: Simulation parameters. The columns show (from left to right): simulation name, box size in h^{-1}Mpc, number of dark matter particles, dark matter particle mass (for the hydrodynamical run) in $h^{-1}M_{\odot}$, redshift used, and cosmological parameters. The dark matter particle mass for Illustris-2-Dark is $4.2 \times 10^7 (h^{-1}M_{\odot})$, and for EAGLE Dark it is $7.5 \times 10^6 (h^{-1}M_{\odot})$.

Sim.	$L_{\rm box}$	$N_{\rm DM}$	$m_{\rm DM}$	z	h	Ω_m	Ω_Λ	Ω_b	σ_8	n_s
Illustris	75	910^3	$3.5 \cdot 10^7$	0.13	0.704	0.273	0.727	0.0456	0.809	0.963
EAGLE	67.77	1504^3	$6.6 \cdot 10^6$	0.101	0.6777	0.307	0.693	0.0483	0.829	0.961

be assigned a central galaxy (if it is assigned a galaxy at all), and thus the only halo properties that we need to know are the position and velocity of the halo, which should be well-established with 2400 particles.

The Illustris simulation was performed with the moving-mesh code AREPO, while the EAGLE simulation was performed with the GADGET-3 tree-SPH code, a modified version of the public GADGET-2 simulation code. Both simulations employ models for star formation, stellar evolution, gas cooling and heating, supernovae feedback, black hole formation, and AGN feedback. According to Scannapieco et al. (2012), while GADGET-3 and AREPO share the same sub-grid physics, their different numerical hydrodynamical techniques can lead to large discrepancies in their galaxies. In their tests, GADGET-3 formed only about half as many stars as AREPO, and AREPO has a much higher gas and stellar mass fraction than GADGET-3. The benefit of using two simulations with different physics for our analysis is that we can compare our results from the two different simulations, providing us with some theoretical uncertainty on our results.

We are interested in two different samples of galaxies: a "high" luminosity sample, similar to that of the volume-limited SDSS DR7 (Abazajian et al., 2009) $M_r < -21$ sample, and a "low" luminosity sample, similar to that of the SDSS DR7 $M_r < -19$ sample. (We will refer to these samples as M_r^{-21} and M_r^{-19} henceforth.) We choose to use the $z = 0.13$ snapshot of the Illustris simulation because it is the closest available redshift to the median redshift of the SDSS M_r^{-21} sample ($z_{\rm med} = 0.132$). We choose the $z = 0.101$ snapshot of the EAGLE simulation because it is also the closest available redshift to that of the SDSS

DR7 M_r^{-21} sample. The M_r^{-19} luminosity threshold sample has a median redshift of 0.054. For the EAGLE simulation, the closest available redshift is still the $z = 0.101$ snapshot. Therefore, because the snapshot does not change for our analysis on the EAGLE simulation, we likewise chose not to change the snapshot for the Illustris simulation. However, there is little evolution between $z = 0.13$ and $z = 0.054$, and we do not compare our clustering statistics to those measured on SDSS data, so our choice of snapshot should not impact our results.

To create our galaxy samples, for each simulation we find the luminosity threshold that results in a galaxy number density equivalent to that of the SDSS datasets of interest (either M_r^{-21} or M_r^{-19}). The luminosity threshold for each simulation and sample is given in Table 2.2. We note that the luminosity thresholds are not exactly -21 or -19, which indicates that the luminosity functions in Illustris and EAGLE are not the same as that in the SDSS, nor are they the same as each other. (This discrepancy emphasizes the lack of consensus among hydrodynamic simulations, and thus the advantage of using HOD modeling with plenty of freedom to model galaxy clustering in the real Universe.) Thus, if we create our samples based on luminosity, our number density will be different than that of the SDSS samples. Therefore, we choose to use a different luminosity threshold to do an accurate number density comparison. We will still refer to the samples as the M_r^{-21} and M_r^{-19} samples.

After setting the luminosity threshold, we then determine the number of remaining galaxies in each halo, and average in bins of halo mass. For the M_r^{-21} samples we use 14 evenly spaced logarithmic bins between 11.9 and 14.52. For the M_r^{-19} samples we use 20 evenly spaced logarithmic bins between 11.0 and 14.52. Our halo occupation distributions for each galaxy sample are shown in Figure 2.1. The Illustris samples are plotted in red, and the EAGLE samples are plotted in blue.

2.3 Halo Occupation Modelling

2.3.1 The Halo Occupation Distribution

The Halo Occupation Distribution framework governs the number, positions, and velocities of galaxies within a dark matter halo based on a few free parameters, which depend only on the mass of the halo. The version of the HOD that we utilize in this work is the five parameter 'vanilla' HOD model of Zheng et al. (2007) (as cited in Sinha et al. (2018)). Within their haloes, galaxies are split into centrals and satellites (Kravtsov et al., 2004; Zheng et al., 2005).

The mean number of central galaxies in a halo of mass M is described by[2]

$$\langle N_{\mathrm{cen}} \rangle = \frac{1}{2}\left[1+\mathrm{erf}\left(\frac{\log M - \log M_{\mathrm{min}}}{\sigma_{\log M}}\right)\right], \tag{2.1}$$

where M_{min} is the mass at which half of halos host a central galaxy, $\sigma_{\log M}$ is the scatter around this halo mass, and $\mathrm{erf}(x)$ is the error function, $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-y^2)dy$. The central galaxy is always placed at the centre of the halo, and given the mean velocity of the halo (i.e. we assume that the central galaxy is at rest with respect to the halo).

We determine the number of satellite galaxies to place in each halo by drawing from a Poisson distribution with a mean given by

$$\langle N_{\mathrm{sat}} \rangle = \langle N_{\mathrm{cen}} \rangle \times \left(\frac{M - M_0}{M_1}\right)^{\alpha}, \tag{2.2}$$

where M_0 is the halo mass below which there are no satellite galaxies, M_1 is the mass where haloes contain on average one satellite galaxy, and α is the slope of the power-law occupation function at high masses. Each satellite galaxy is assigned the position and velocity of a randomly chosen dark matter particle within the halo, i.e. we assume that satellite galaxies trace the spatial and velocity distribution of dark matter within the halo.

In summary, our HOD model contains five free parameters that control the number of

[2] Throughout this paper, log refers to $\log_{10}$.

galaxies in each halo as a function of halo mass. Our model assumes that all galaxies live inside dark matter haloes, and that the number of galaxies in a halo depends only on the mass of the halo and not on any other halo properties, such as age or concentration (i.e. there is no galaxy assembly bias). However, recent work (e.g. Zentner et al., 2014; Vakili and Hahn, 2019; Zentner et al., 2019) indicates that galaxy assembly bias is probably present in luminosity threshold samples, so this assumption is likely incorrect.

Additionally, our model assumes that the number of satellite galaxies in each halo is governed by a Poisson distribution. However, results from simulations indicates that the scatter in the number of satellite galaxies at fixed halo mass is probably non-Poissonian (Boylan-Kolchin et al., 2010; Mao et al., 2015). In fact, Jiménez et al. (2019) found that the HOD was best able to reproduce the spatial distribution of galaxies in a semi-analytical model when they used a negative binomial distribution to govern the number of satellite galaxies in a halo.

Finally, our model assumes that the central galaxy in each halo lives at the centre of the halo and moves with the mean velocity of the halo (i.e. there is no central spatial or velocity bias), and that the satellite galaxies in each halo follow the spatial and velocity distribution of dark matter within the halo (i.e. there is no satellite spatial or velocity bias). However, observations suggest that both central and satellite galaxies probably do exhibit spatial bias (e.g. Watson et al., 2012; Piscionere et al., 2015) as well as velocity bias (e.g. Van den Bosch et al., 2005; Guo et al., 2015b,a).

While we do use this standard ‘vanilla’ HOD in our initial analysis, we will discuss variations and extensions of this model in Section 2.7.

2.3.2 Fitting the HOD

Next, we need to determine the five parameters that best describe the HOD in each simulation and sample. We do this in the following way. We start with an initial guess for each parameter. Using this fiducial HOD model, we assign a number of central and satellite

Table 2.2: HOD parameters for each sample. The columns show (from left to right): the simulation name, the absolute magnitude limit for the SDSS sample whose number density we are matching, the absolute magnitude limit used in the case of the given simulation, the galaxy number density in $h^3\mathrm{Mpc}^{-3}$, the five best-fitting HOD parameters for that sample, and the corresponding reduced chi-square value.

Simulation	M_r	M_r^{lim}	n_{g}	$\log M_{\mathrm{min}}$	$\sigma_{\log M}$	$\log M_0$	$\log M_1$	α	χ^2/dof
Illustris	-21	-22.840	0.0012	12.681	0.532	12.296	13.635	0.994	0.908
Illustris	-19	-20.354	0.0149	11.500	0.180	11.659	12.590	0.979	8.560
EAGLE	-21	-21.852	0.0012	12.767	0.504	12.467	13.799	1.000	1.498
EAGLE	-19	-19.695	0.0149	11.555	0.237	11.717	12.566	0.938	3.635

galaxies to the haloes in the hydrodynamic run of the simulation. (The halo mass that we use for this is the total friends-of-friends group mass, i.e. including dark matter as well as baryonic particles.) Because there is some random variation in the HOD modelling framework, we repeat this process 300 times in order to generate 300 different realizations of our fiducial HOD. We then determine the number of galaxies in each halo (averaged in bins of halo mass), in the same way that we did for the original galaxies in the simulation. We can then calculate a χ^2 to assess how well our fiducial HOD model fits the simulation:

$$\chi^2 = \sum_i \frac{(D_i - M_i)^2}{\sigma_i^2}, \tag{2.3}$$

where D_i is the number of galaxies in one halo mass bin from the simulation, M_i is the number of galaxies in the same halo mass bin averaged over 300 realizations of our fiducial HOD model, and σ_i is the standard deviation among the 300 different realizations of our fiducial HOD. We do this separately for centrals and satellites, and then sum over all of our halo mass bins. Based on this χ^2, we adjust our fiducial HOD parameters and repeat this process. We use a Nelder-Mead optimization algorithm (Nelder and Mead, 1965; Gao and Han, 2012; Jones et al., 2001) to minimize χ^2.

In Table 2.2, we list the luminosity thresholds for each sample, as well as the best-fitting HOD parameters for each simulation. Shown in Figure 2.1 are the best fit HODs for each of our simulations and density samples. While the M_r^{-21} samples in both simulations

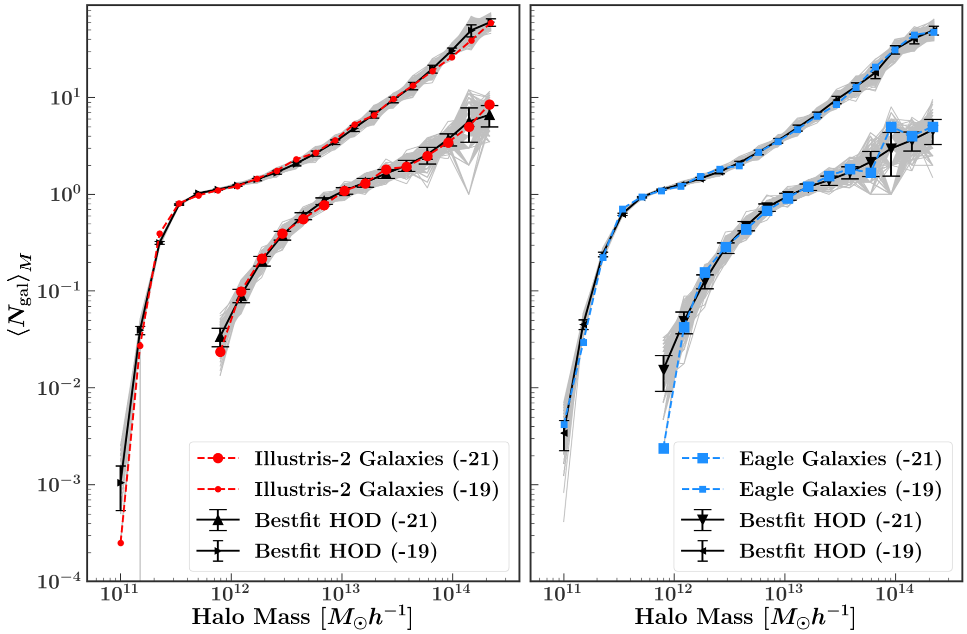

Figure 2.1: Best-fitting HOD for Illustris-2 (left) and EAGLE (right) galaxies. The Illustris-2 high luminosity (M_r^{-21}) galaxy sample is plotted with a dashed red line and large circles, and the low luminosity (M_r^{-19}) sample is plotted with a dashed red line and small circles, while the EAGLE high luminosity sample is plotted with a dashed blue line and large squares, and the low luminosity sample is plotted with a dashed blue line and small squares. The gray lines in each case show 300 realizations of the best-fitting HOD model for that sample. The black line and error bars represent the mean and standard deviation among these 300 realizations.

each achieved a χ^2/DOF of close to 1, the M_r^{-19} samples are not fit as well by the HOD, particularly in Illustris. This could be an indication that the form of the HOD is not optimal for describing a low-luminosity galaxy sample, but it can easily describe a high-luminosity sample.

One of the assumptions made in our modelling procedure is that the probability distribution governing the number of satellite galaxies in a halo is Poissonian. To investigate this assumption we examine the average number of satellite-satellite pairs per halo in bins of halo mass, $\langle N(N-1)\rangle_M$, or $\langle N^2\rangle_M - \langle N\rangle_M$. A Poisson distribution of mean $\langle N\rangle$ has variance $\langle N^2\rangle = \langle N\rangle^2 + \langle N\rangle$. Thus, if the number of satellite galaxies comes from a Poisson distribution, then $\langle N(N-1)\rangle_M/\langle N\rangle^2$ should be equal to 1 (Berlind et al., 2003). In Figure 2.2 we have plotted this quantity for the Illustris (left, red) and EAGLE (right, blue) M_r^{-19} samples as a function of halo mass. We have also plotted percentiles for our 300

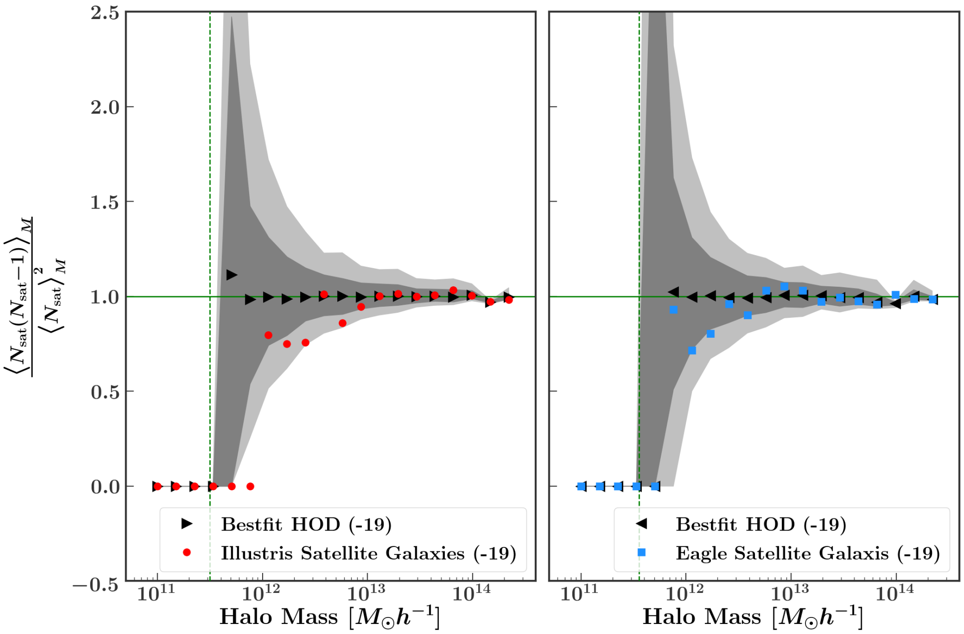

Figure 2.2: The second moment of the HOD for Illustris-2 M_r^{-19} galaxies (red points, left) and EAGLE M_r^{-19} galaxies (blue points, right). The dark and light gray shaded regions show the inner 68 and 95% of the realizations of the best-fitting HOD model for that sample, and the black points are the median of the 300 realizations.

HOD realizations for each sample (shown in gray), as well as the median of the 300 realizations. In our HOD model, the number of satellite galaxies is drawn from a Poisson distribution by design, so the median of these realizations should be 1 for all halo mass bins above $M_{\rm min}$ (indicated by the vertical green dashed line; below $M_{\rm min}$ it is extremely unlikely that there will be any satellites, so this quantity should be 0.) Both the Illustris and EAGLE samples are Poissonian at higher halo masses, but appear slightly sub-Poissonian at lower halo masses. However, neither sample is incompatible with its corresponding distribution of HOD realizations, so it is reasonable to conclude that the satellite numbers in Illustris and EAGLE are consistent with our HOD model. (The M_r^{-21} samples have very few satellites, and thus are very noisy, which is why they are not shown here. They do not exhibit any non-Poissonian trends.)

2.3.3 Building mock galaxy catalogs

Once we have determined the best-fitting HOD parameters for our sample, we then need to actually place galaxies in haloes. We do this on the dark matter only versions of the simulations. As stated earlier, the halo mass of interest is the total mass of the Friends-of-Friends group (i.e. parent halo). We assign the central galaxy the position of the group, which is defined as the spatial position within the periodic box of the particle with the minimum gravitational potential energy (in comoving coordinates). Additionally, we assign the central galaxy the velocity of the group, which is the sum of the mass weighted velocities of all particles/cells in the group. The peculiar velocity is obtained by multiplying this value by $1/a$, where a is the scale factor. (In the EAGLE simulation, the velocity of the parent halo is not provided, so we instead assign the central galaxy the velocity of the central subhalo.) To place satellite galaxies, we randomly select dark matter particles from the parent halo and assign galaxies the positions and velocities of these randomly chosen particles. The only stipulation we make is that we never choose the same random dark matter particle twice; i.e. we will never place two galaxies on the same particle, but we can place them on very nearby particles. We repeat this process 1000 times, so that we ultimately have 1000 different realizations of our best-fitting HOD model applied to our dark matter only simulation. We will refer to these 1000 realizations as mock galaxy catalogues.

2.4 Galaxy Clustering Measurements

Once we have populated the dark matter haloes in each simulation with galaxies, the next step is to measure a series of clustering statistics on both the galaxies from the original simulation and the galaxies from our mock catalogues. We measure these statistics in the same way on the simulation galaxies as we do on our mocks, in order to assess how well our HOD model can reproduce galaxy clustering properties as compared to a full hydrodynamic simulation.

The first property that we measure is the number density of galaxies. By comparing the number densities of galaxies in our simulations and in our mocks, we can test how well the HOD fits the simulation, as well as how similar the halo mass functions are in the hydrodynamic and dark matter only simulations. Figures 2.3-2.6 show results for the Illustris M_r^{-21}, EAGLE M_r^{-21}, Illustris M_r^{-19}, and EAGLE M_r^{-19} samples, respectively. The top left panel of each figure shows the distribution of number densities among the 1000 mocks for that sample (together with the mean and standard deviation), as well as the number density for the corresponding hydrodynamic sample. The shaded region in each figure shows cosmic variance errors (one standard deviation) calculated from 400 mock galaxy catalogues of the corresponding SDSS sample (Sinha et al., 2018). The spread among our 1000 HOD mocks indicates how well we can measure galaxy number density in a box given the scatter in our HOD model. The spread among 400 SDSS mocks indicates how accurately a difference in number density could be detected by the SDSS.

In every case, applying the HOD to the dark matter only simulation results in a significantly overestimated galaxy number density (by up to 20% for the Illustris M_r^{-21} sample). For both M_r^{-21} samples (Figures 2.3 and 2.4), this difference in number density is larger than the cosmic variance error from the SDSS M_r^{-21} sample (shown in green); in other words, an SDSS-like survey would easily notice this discrepancy. For the M_r^{-19} samples (Figures 2.5 and 2.6), although the difference between the simulation and the HOD number density is quite significant, the cosmic variance error (shown in yellow) is larger, indicating that an SDSS-like survey would not pick up on this difference. None the less, it is shocking that in every case the HOD (which was fit to the simulation) systematically significantly overestimates the galaxy number density. This points to a major issue with applying HOD to a dark matter only simulation: the halo mass function is different in hydrodynamic and dark matter only simulations. This will be discussed further in Section 2.6.

Next, we measure five additional clustering statistics. Before we can do this, we must introduce redshift-space distortions into both our simulation galaxies as well as our mock galaxies. We do this by placing an observer infinitely far away from our box and taking the

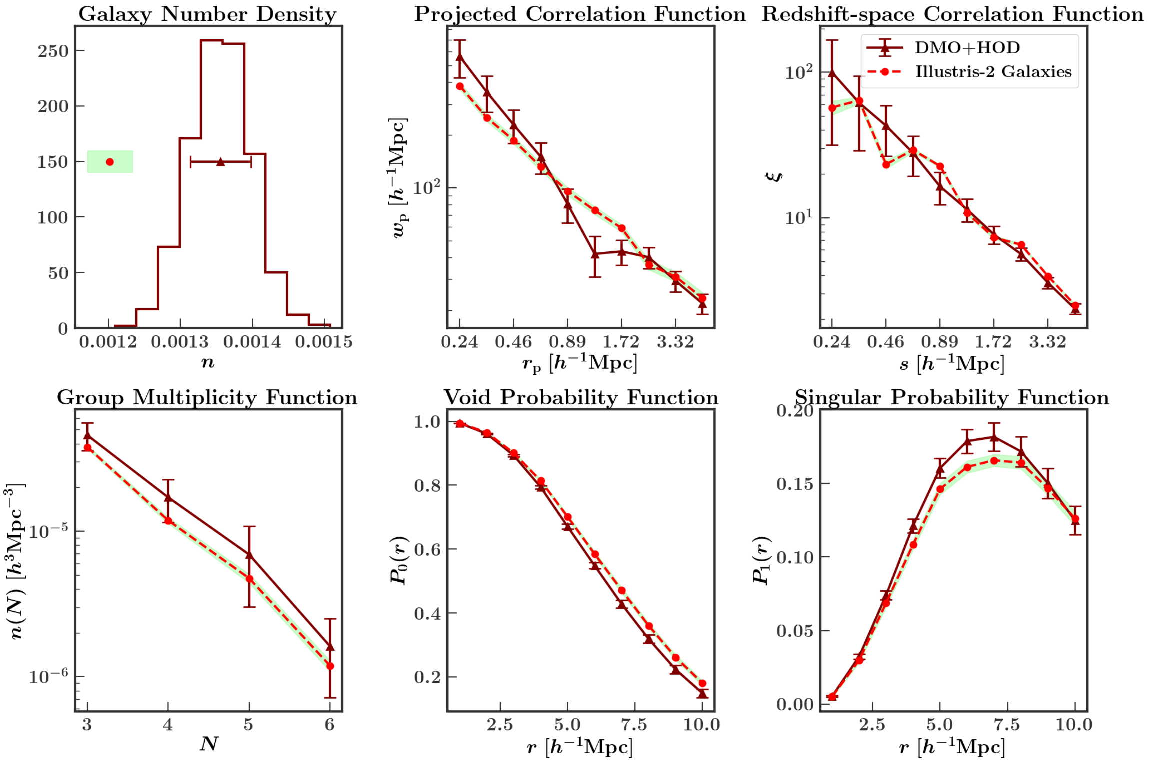

Figure 2.3: All clustering measurements for the M_r^{-21} sample of Illustris-2 galaxies. The red lines are measured on galaxies from the original hydrodynamic simulation, while the dark red lines show the average of 1000 realizations of the best-fitting HOD model applied to the dark matter only simulation. The error bars represent the standard deviation among the 1000 realizations. The shaded regions around the red lines show cosmic variance errors (one standard deviation) calculated from 400 mock galaxy catalogues of the SDSS M_r^{-21} sample, and thus illustrate the size of deviations that could be detected by the SDSS.

z-axis as the line of sight coordinate (using periodic boundary conditions). Including these distortions allows us to probe how well our model reproduces the velocities of the galaxies.

Berlind and Weinberg (2002) investigated galaxy bias in an HOD framework by measuring several clustering statistics. They found that the galaxy correlation function is affected by different parts of the HOD on different scales, and that other clustering statistics (such as the void probability function and the group multiplicity function) are also sensitive to different combinations of HOD parameters. Sinha et al. (2018) similarly found that analyses involving several different galaxy clustering statistics have the most power to constrain galaxy bias. Because of this, the five additional clustering statistics that we measure in this work are the redshift-space correlation function, the projected correlation

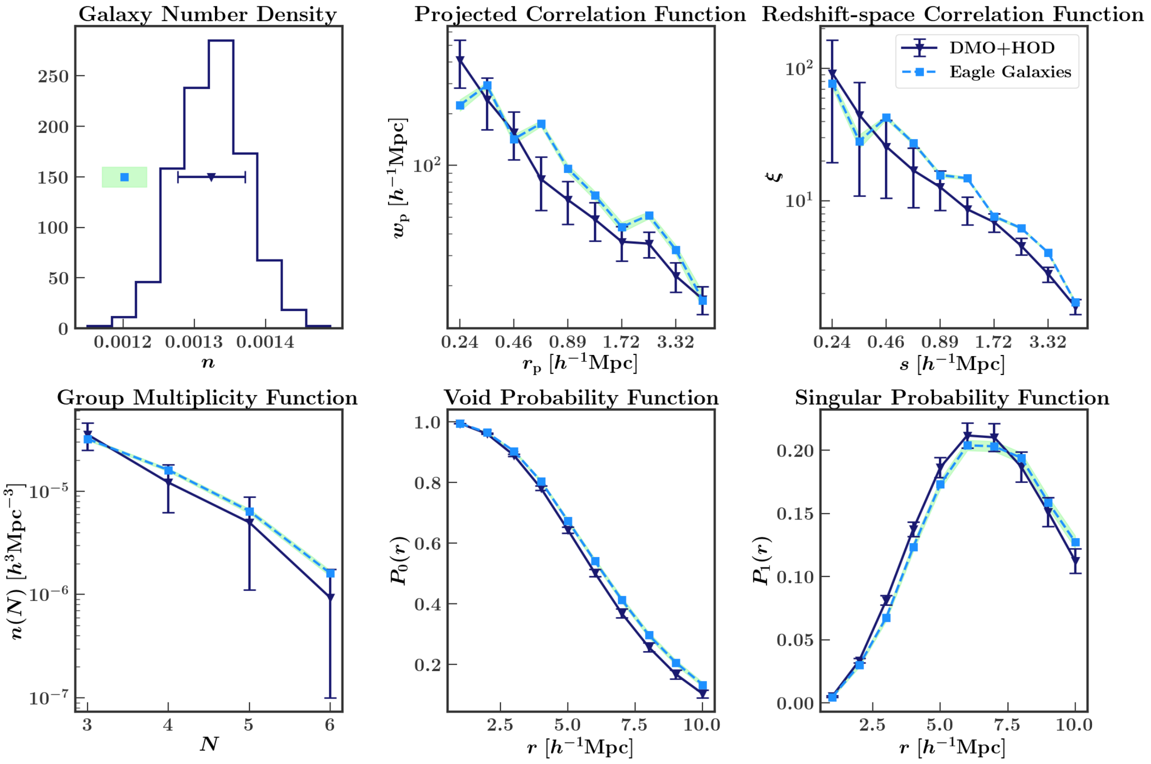

Figure 2.4: Same as Fig. 2.3 for the M_r^{-21} sample of EAGLE galaxies.

function, the group multiplicity function, the void probability function, and what we call the "singular probability function" (i.e. the probability of having exactly one galaxy in a region). These five different clustering statistics are described in detail below.

2.4.1 The projected correlation function

The most commonly used galaxy clustering statistic, the projected correlation function, removes the effect of redshift-space distortions by first counting pairs of galaxies in bins of their line-of-sight and projected components, π and r_p, and then integrating over π:

$$w_\mathrm{p}(r_\mathrm{p}) = 2\int_0^{\pi_\mathrm{max}} \xi(r_\mathrm{p},\pi)d\pi. \tag{2.4}$$

We count pairs of galaxies in 10 evenly spaced logarithmic bins of projected separation r_p between 0.2 and 5.37h^{-1}Mpc. We then integrate out to π_max of 20h^{-1}Mpc for each sample. (For computational reasons, π_max must be $< \frac{1}{3}L_{box}$.) We use the blazing fast

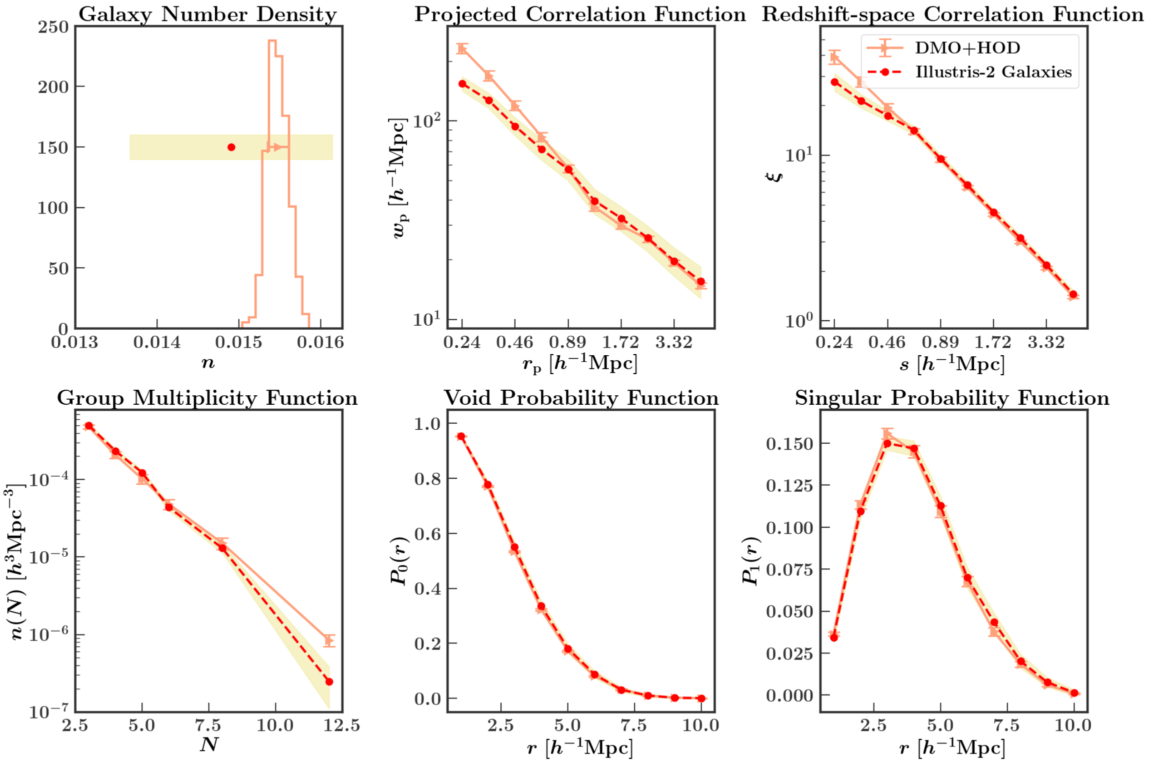

Figure 2.5: Same as Fig. 2.3 for the M_r^{-19} sample of Illustris-2 galaxies.

code CORRFUNC (Sinha and Garrison, 2017, 2019) to compute our projected correlation function.

The projected correlation function has been used as the workhorse of HOD modelling (e.g., Zehavi et al., 2011; Sinha et al., 2018). Recently, Zentner et al. (2019) used measurements of the projected correlation function to constrain assembly bias of SDSS DR7 galaxies within the decorated HOD model of Hearin et al. (2016). The authors found highly significant central galaxy assembly bias in the M_r^{-20} and $M_r^{-20.5}$ samples, as well as significant satellite galaxy assembly bias for the M_r^{-19} sample. They did not find any assembly bias in the M_r^{-21} sample. Meanwhile, Vakili and Hahn (2019) also looked at clustering measurements of SDSS DR7 galaxies and found that at fixed halo mass, satellite galaxies show no correlation with halo concentration, and central galaxies shows little correlation with halo concentration for the M_r^{-21} and $M_r^{-21.5}$ samples, and slight correlation with halo concentration in the $M_r^{-20.5}$, M_r^{-20}, and M_r^{-19} samples.

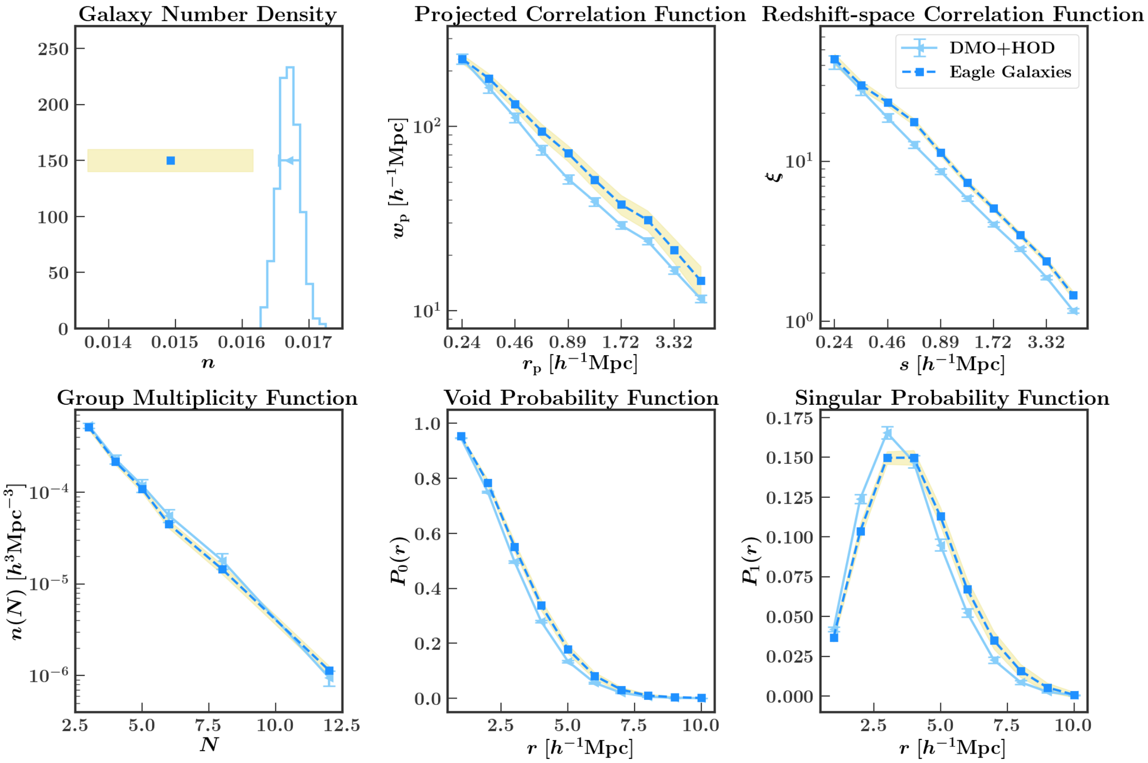

Figure 2.6: Same as Fig. 2.3 for the M_r^{-19} sample of EAGLE galaxies.

In the top middle panels of Figures 2.3–2.6 we have plotted the projected correlation function from the hydrodynamic simulations, as well as the average projected correlation function of our 1000 dark matter only mocks, for each of our samples. For the M_r^{-21} samples (Figures 2.3 and 2.4) the HOD does reasonably well at recovering the projected correlation function from the simulations. Though there are visible discrepancies, these are not highly significant given the plotted uncertainties. However, for the Illustris M_r^{-19} sample (Fig. 2.5), the HOD significantly overestimates the projected correlation function at small scales. In contrast, for the EAGLE M_r^{-19} sample (Fig. 2.6), the HOD significantly underestimates the projected correlation function at all but the smallest scales. This indicates that although the clustering is correct for high luminosity galaxies, there is a possible problem with the spatial assumptions made in the HOD, which specifically impacts the clustering of low luminosity galaxies. The Illustris M_r^{-19} sample is most likely affected by spatial bias, which impacts small scales, while the EAGLE M_r^{-19} is likely more affected by assembly bias, which impacts large scales. We note that the projected correlation function is not sensitive to velocity information, so any discrepancies must be due to spatial and/or assembly bias, and not velocity bias. These biases will be discussed further in Section 2.7.

2.4.2 The redshift-space correlation function

The three-dimensional redshift-space two-point correlation function $\xi(s)$ is the excess number of galaxy pairs above that which is expected for a random distribution of points, as a function of redshift-space pair separation s (in contrast to the projected separation r_p described above). In this work, we count pairs in 10 bins of separation s between 0.2 and $5.37h^{-1}$Mpc (the same bins as those used for the projected correlation function). We also use CORRFUNC to compute our redshift-space correlation function. Measuring the redshift-space correlation function allows us to access not only spatial information about our galaxies, but also velocity information, because the redshift-space distortions of our galaxies depend on their velocities. Thus, with this measurement, we can examine the validity of the assumption in the HOD that galaxies trace the velocity distribution of dark matter within the halo (in addition to examining our assumptions about the spatial distribution of galaxies).

In the top right panels of Figures 2.3–2.6 we have plotted the redshift-space correlation function from our simulations, as well as the average redshift-space correlation function of our 1000 mocks, for each of our samples. Results are qualitatively similar to those using the projected correlation function. For the M_r^{-21} samples (Figs 2.3 and 2.4) the HOD successfully recovers the redshift-space correlation function from the simulations. However, for the Illustris M_r^{-19} sample (Fig. 2.5), the HOD once again significantly overestimates the correlation function at small scales, while for the EAGLE M_r^{-19} sample (Fig. 2.6), the HOD significantly underestimates the correlation function at all but the smallest scales. This again suggests a problem with the spatial assumptions made in the HOD, as well as the velocity assumptions, which specifically impact the clustering of low luminosity galaxies.

This will be discussed further in Section 2.7.

2.4.3 The group multiplicity function

The group multiplicity function is the abundance of galaxy groups as a function of the number of galaxies in the group, $n(N)$ (e.g., Berlind and Weinberg, 2002). We use the Berlind et al. (2006a) friends-of-friends algorithm for identifying groups. Galaxies are linked together if their projected and line-of-sight separations are both less than a corresponding linking length. We adopt the Berlind et al. (2006a) linking lengths of $b_{\perp} = 0.14$ and $b_{\|} = 0.75$, which are given in units of the mean inter-galaxy separation ${n_{\rm g}}^{-1/3}$, where $n_{\rm g}$ is the sample number density. For our low luminosity samples, we measure groups with the following numbers of galaxies: $3, 4, 5, 6-7, 8-11, > 12$. For our high luminosity samples, we measure groups of 3, 4, 5, and 6 or more galaxies.

In the lower left panels of Figures 2.3–2.6 we have plotted the group multiplicity function from our simulations, as well as the average group multiplicity function of our 1000 mocks, for each of our samples. For the M_r^{-21} samples (Figures 2.3 and 2.4) the HOD successfully recovers the group multiplicity function from the simulations. The HOD also successfully reproduces the group multiplicity function for the EAGLE M_r^{-19} sample (Fig. 2.6). However, for the Illustris M_r^{-19} sample (Fig. 2.5), the HOD significantly overestimates the group multiplicity function for the largest groups. This further points to a problem with the spatial and velocity assumptions made in the HOD, particularly as they affect the clustering of low luminosity galaxies in Illustris. This will be discussed further in Section 2.7.

2.4.4 Counts-in-cells statistics

Counts-in-cells statistics measure the probability of finding a given number of galaxies within a randomly placed finite region (e.g. a sphere) as a function of region size (e.g. radius). One special case of this is the void probability function (VPF), which measures the probability of finding no galaxies in a random region of space. Tinker et al. (2006a) attempted to constrain galaxy bias using void statistics within an HOD framework, and found that the VPF, in contrast to the projected correlation function, is quite sensitive to environmental variations of the HOD. Later, McCullagh et al. (2017) showed that catalogues created using SHAM and the semi-analytic model GALFORM, which were designed to have the same large-scale 2-point clustering, have different VPFs due to their different HOD shapes, suggesting that the VPF could be used to rule out certain HOD models. Recently, Walsh and Tinker (2019) fit the standard HOD model to the two-point correlation function of BOSS galaxies and found that it was able to accurately predict the void probability function, indicating that galaxy assembly bias does not affect the clustering of massive galaxies.

Wang et al. (2019) studied the power of the VPF, counts-in-cylinders, and counts-in-annuli, as well as the projected two-point correlation function and the galaxy-galaxy lensing signal to constrain galaxy assembly bias from redshift survey data using the decorated HOD, and found that the counts-in-cells statistics are more efficient at constraining galaxy assembly bias when combined with the projected correlation function than galaxy-galaxy lensing is.

Another variation of counts in cells that we use is what we will refer to as the "singular probability function," (SPF) or the probability of finding exactly one galaxy in a randomly placed region. We measure both the VPF and the SPF in spheres of evenly spaced bins of radius r, beginning with $1h^{-1}$Mpc and ending with $10h^{-1}$Mpc.

In the lower middle (right) panels of Figures 2.3–2.6 we have plotted the VPF (SPF) of our simulations, as well as the average of our 1000 mocks, for each of our samples. For the Illustris M_r^{-21} sample (Fig. 2.3) the HOD struggles to recover the VPF at intermediate and large scales, and likewise struggles to recover the SPF at intermediate scales. For the EAGLE M_r^{-21} sample (Fig. 2.4) the HOD shows similar tension in the VPF and the SPF. For the Illustris M_r^{-19} sample the agreement looks better, but the error bars are very small so it is difficult to surmise based on looking at Figure 2.5 alone. For the EAGLE M_r^{-19} sample

(Fig. 2.6) the HOD struggles to reproduce both the VPF and the SPF at most scales. These problems could indicate issues with the assumptions made in the HOD. They could also be compounded by the inability of the HOD to reproduce the correct number density, since counts-in-cells statistics, and the VPF in particular, are very sensitive to number density. This will be discussed further in Section 2.7.

2.5 Assessing the Accuracy of the HOD Model

In Figures 2.3–2.6 we saw that for some statistics (like number density) the HOD applied to dark matter only simulations does not provide a good fit to the hydrodynamic simulations for any of our samples, while for other statistics (like the correlation functions) the HOD appeared to provide a good fit to the simulations for the high luminosity samples and not the low luminosity samples. In general, however, the success of the HOD model is difficult to ascertain visually because error-bars are often small and are likely correlated. In order to quantify the accuracy with which our HOD model can reproduce the clustering statistics measured on a hydrodynamic simulation, we calculate χ^2 for each clustering statistic

$$\chi^2 = \sum_{ij} \chi_i R_{ij}^{-1} \chi_j, \tag{2.5}$$

where

$$\chi_i = \frac{D_i - M_i}{\sigma_i}, \tag{2.6}$$

D_i is the value of one bin of a clustering measurement on the hydrodynamic simulation galaxies (either Illustris or EAGLE, and either M_r^{-19} or M_r^{-21}), M_i is that same measurement averaged over our 1000 mock galaxy catalogues for that sample, and σ_i is the standard deviation of that measurement among the 1000 mock galaxy catalogues. R_{ij} is the correlation matrix for each clustering statistic

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}, \tag{2.7}$$

which is the covariance matrix normalized by its diagonal elements. The covariance matrix is calculated as

$$C_{ij} = \frac{1}{N-1} \sum_{1}^{N} (y_i - \overline{y_i})(y_j - \overline{y_j}), \tag{2.8}$$

where the sum is over the $N = 1000$ mock galaxy catalogues, and y_i and y_j are two bins of a clustering statistic, and $\overline{y_i}$ and $\overline{y_j}$ are the mean measurements over the 1000 mocks. We note that since the hydrodynamic simulation and the HOD mocks come from initial conditions with the same phases, cosmic variance errors do not apply to this comparison.

From this χ^2, we can calculate the corresponding p-value, which represents the probability that a sample randomly drawn from the best-fitting HOD model could have a χ^2 value greater than the one exhibited by the simulation. In other words, the p-value represents the probability that the hydrodynamic simulation is consistent with the DMO+HOD model. The p-value for each clustering measurement uses all the spatial bins of the measurement, as well as the full covariance matrix for that statistic. These p-values are listed in Table 2.3 (in the rows labeled as "No Correction").

Looking at Figures 2.3–2.6 or the p-values in Table 2.3, it is immediately clear that the vanilla HOD model, when applied to haloes from a dark matter only simulation, does not provide a good fit to the corresponding hydrodynamic simulation for all of the clustering statistics in question. However, the success of the HOD model is highly dependent on the simulation and luminosity sample in question. For example, the model generally performs better for high luminosity galaxies than for low luminosity galaxies. Specifically, for the Illustris M_r^{-21} sample, all of the clustering statistics are well fit by the HOD model, at least within a 3σ tolerance, except for number density. For the EAGLE M_r^{-21} sample, even the number density works well. However, for the low luminosity samples, almost none of the clustering statistics are well fit by the DMO+HOD model, and in most cases exhibit discrepancies far greater than $> 3\sigma$.

The green shaded regions in Figures 2.3 and 2.4 represent one standard deviation of cosmic variance errors calculated from 400 mock galaxy catalogues of the SDSS M_r^{-21}

Table 2.3: p-values from comparing the clustering statistics of hydrodynamic galaxies to those of DMO+HOD mock galaxies, for different simulations (Illustris or EAGLE) and samples (-21 or -19), with no correction (first), after correcting the halo mass function (second), additionally removing satellite spatial bias (third), additionally removing all spatial and velocity bias (fourth), and additionally removing assembly bias (fifth). The columns show (from left to right): simulation name, magnitude limit for the SDSS sample with the same galaxy number density, which model was used, and the p-values for each of our six measurements.

Sim.	M_r	Corr.	n	$w_p(r_p)$	$\xi(s)$	$n(N)$	$P_0(r)$	$P_1(r)$
I	-21	None	2.84×10^{-4}	4.61×10^{-2}	6.62×10^{-1}	5.36×10^{-1}	1.86×10^{-2}	4.04×10^{-1}
I	-21	HMF	4.54×10^{-1}	1.43×10^{-1}	9.14×10^{-1}	9.77×10^{-1}	4.79×10^{-1}	6.28×10^{-1}
I	-21	+SB	4.54×10^{-1}	6.46×10^{-1}	7.51×10^{-1}	6.95×10^{-1}	4.84×10^{-1}	6.35×10^{-1}
I	-21	+VB	4.54×10^{-1}	6.66×10^{-1}	5.98×10^{-1}	7.09×10^{-1}	3.97×10^{-1}	6.13×10^{-1}
I	-21	+AB	4.54×10^{-1}	6.15×10^{-1}	5.25×10^{-1}	6.11×10^{-1}	5.23×10^{-1}	6.87×10^{-1}
I	-19	None	8.35×10^{-6}	1.13×10^{-7}	1.61×10^{-4}	5.36×10^{-4}	4.44×10^{-6}	5.99×10^{-2}
I	-19	HMF	6.66×10^{-2}	5.23×10^{-6}	2.43×10^{-3}	3.48×10^{-4}	1.05×10^{-3}	5.25×10^{-2}
I	-19	+SB	6.66×10^{-2}	2.58×10^{-2}	1.14×10^{-1}	1.87×10^{-2}	2.11×10^{-3}	5.69×10^{-2}
I	-19	+VB	6.66×10^{-2}	2.89×10^{-2}	1.94×10^{-1}	8.76×10^{-2}	9.68×10^{-2}	4.42×10^{-1}
I	-19	+AB	6.66×10^{-2}	7.64×10^{-2}	4.81×10^{-1}	1.65×10^{-1}	3.93×10^{-1}	7.82×10^{-1}
E	-21	None	9.84×10^{-3}	5.89×10^{-3}	3.69×10^{-3}	8.18×10^{-1}	5.32×10^{-2}	1.91×10^{-2}
E	-21	HMF	8.56×10^{-1}	3.64×10^{-2}	4.07×10^{-2}	7.02×10^{-1}	5.55×10^{-1}	2.01×10^{-1}
E	-21	+SB	8.56×10^{-1}	4.05×10^{-1}	1.53×10^{-1}	9.18×10^{-2}	6.99×10^{-1}	2.92×10^{-1}
E	-21	+VB	8.56×10^{-1}	4.06×10^{-1}	2.53×10^{-1}	1.98×10^{-1}	6.61×10^{-1}	2.69×10^{-1}
E	-21	+AB	8.56×10^{-1}	3.08×10^{-1}	5.55×10^{-1}	4.84×10^{-1}	3.55×10^{-1}	4.06×10^{-1}
E	-19	None	6.4×10^{-29}	1.1×10^{-13}	1.6×10^{-24}	4.50×10^{-1}	7.1×10^{-54}	3.4×10^{-22}
E	-19	HMF	8.25×10^{-1}	1.06×10^{-8}	3.4×10^{-10}	6.31×10^{-1}	4.8×10^{-13}	1.42×10^{-7}
E	-19	+SB	8.25×10^{-1}	3.90×10^{-5}	2.22×10^{-8}	1.13×10^{-1}	8.6×10^{-13}	1.87×10^{-7}
E	-19	+VB	8.25×10^{-1}	6.80×10^{-5}	2.40×10^{-5}	2.24×10^{-1}	7.9×10^{-10}	6.50×10^{-5}
E	-19	+AB	8.25×10^{-1}	1.49×10^{-1}	3.10×10^{-1}	4.92×10^{-1}	4.97×10^{-1}	6.07×10^{-1}

sample. These mocks were created as part of the Large Suite of Dark Matter Simulations project (LasDamas; McBride et al., 2009) and used in Sinha et al. (2018). In our M_r^{-21} Illustris and EAGLE samples, the errors among our 1000 mock galaxy catalogues (which are different HOD realizations) are much larger than the cosmic variance errors from the 400 SDSS-like mocks. Consequently, though the HOD model appears to be a good fit to the simulations for high luminosity galaxies, an SDSS size M_r^{-21} survey (which has small errors due to its large volume) could be sensitive to clustering differences that we are unable to detect in our analysis due to our smaller volume.

Similarly, the yellow shaded regions in Figures 2.5 and 2.6 represent one standard deviation of cosmic variance errors calculated from 400 mock galaxy catalogues of the SDSS M_r^{-19} sample, constructed in a similar way as those in Sinha et al. (2018). In our M_r^{-19} Illustris and EAGLE samples, the errors among our 1000 mock galaxy catalogues are smaller than the cosmic variance errors from the 400 SDSS-like mocks. For some statistics (such as the number density), a survey with the precision of SDSS would not necessarily be able to detect the differences we have found between the HOD model and the hydrodynamic simulation. For other clustering statistics (particularly the correlation functions) it is clear that, although the cosmic variance errors are somewhat broad, there is still an obvious difference between the HOD model and the simulation, to which even an SDSS-like survey would be sensitive.

2.6 The Effect of Baryons on the Halo Mass Function

Figures 2.3 – 2.6 revealed that the galaxy number density is not well predicted in any sample. Recall that, in our vanilla HOD, the number of galaxies in a halo is solely dependent on the mass of the halo. Thus, the fact that our HOD systematically over-predicts the galaxy abundance indicates either that the functional form of our HOD is incorrect, or that the halo mass functions (HMFs) are different in the hydrodynamic simulations compared to their dark matter only (DMO) counterparts.

Figure 2.7 compares the abundance of haloes in the hydrodynamic and DMO versions of the same simulation. The comparison reveals sizeable discrepancies between the halo mass functions. In Illustris (red), the hydrodynamic HMF is consistently lower than the DMO HMF above $10^{12}h^{-1}M_{\odot}$, and higher than the DMO HMF at smaller masses. In EAGLE (blue), the hydrodynamic HMF is below the DMO HMF at all halo masses below $10^{14}h^{-1}M_{\odot}$. In other words, the hydrodynamic HMFs are shifted to lower masses in both simulations, but the detailed effects of baryons on the HMF are different in the two simulations.

This result is consistent with both Desmond et al. (2017) and Schaller et al. (2015), who examined the differences between the halo masses in the EAGLE dark matter only and hydrodynamic runs, and found the haloes to be less massive on average in the hydrodynamic run. Desmond et al. (2017) found that, at low halo masses, stellar feedback in EAGLE removes baryons from the halo, which in turn reduces the growth rate of the halo. At slightly higher halo masses, stellar feedback becomes less effective, but AGN feedback is still capable of expelling baryons. For the most massive haloes, AGN feedback too becomes less effective, and thus there is little discrepancy between the hydrodynamic and DMO halo mass functions.

Our results for the Illustris haloes are consistent with the findings of Vogelsberger et al. (2014b), who found that the halo mass function in Illustris is most affected at low ($< 10^{10}h^{-1}M_{\odot}$) and high ($> 10^{12}h^{-1}M_{\odot}$) halo masses, where baryonic feedback processes (e.g. reionization, SN feedback, and AGN feedback) are strongest, leading to a reduction in halo mass compared to their DMO counterparts. They found that removing AGN feedback boosts the massive end of the halo mass function (e.g. Cui et al., 2012). They also found that haloes around $10^{11}h^{-1}M_{\odot}$, where star formation is most efficient, tend to be more massive than their DMO counterparts.

In Figure 2.8 we show the ratio of halo masses in the hydrodynamic simulation over the masses in the DMO simulation as a function of halo mass in the DMO simulation, for

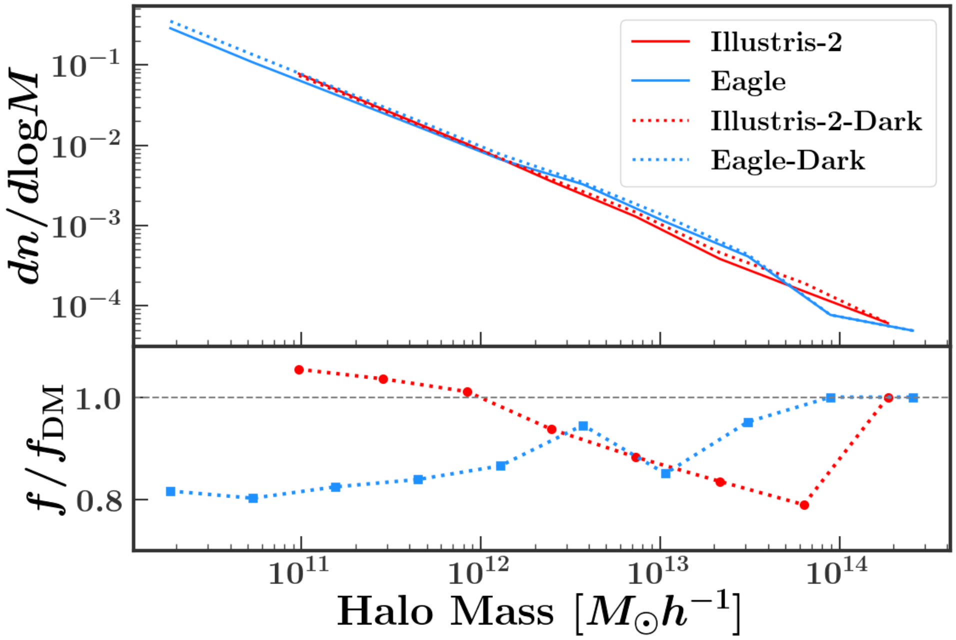

Figure 2.7: Halo mass functions of hydrodynamic compared to dark matter only simulations in the case of Illustris-2 (red) and EAGLE (blue). The hydrodynamic versions are plotted with solid lines, while the dark matter only versions are plotted with dotted lines. The bottom panel shows the ratio of the hydrodynamic to dark matter only mass functions for the two simulations.

both the Illustris-2 (red) and the EAGLE (blue) simulations. The hydrodynamic and DMO haloes are matched based on their ranked masses, rather than spatial positions, so that the point furthest to the right in the figure corresponds to the highest mass DMO halo, paired with the highest mass hydrodynamic halo. In other words, we essentially abundance match the haloes in the hydrodynamic and DMO simulations. As a result, the figure shows the mass correction one would need to apply to the DMO masses in order to recover the global hydrodynamic HMF. However, applying this correction would not necessarily result in the correct dependence of the HMF on environment.

Our result is consistent with the results of Vogelsberger et al. (2014b) and Schaller et al. (2015), who looked at matched haloes in Illustris and EAGLE, respectively. Additionally, Springel et al. (2018) looked at this same quantity for the IllustrisTNG simulations and found a trend that is different from both Illustris and EAGLE. Baryons in the IllustrisTNG seem to have a larger impact on low mass haloes and a smaller impact on high mass haloes

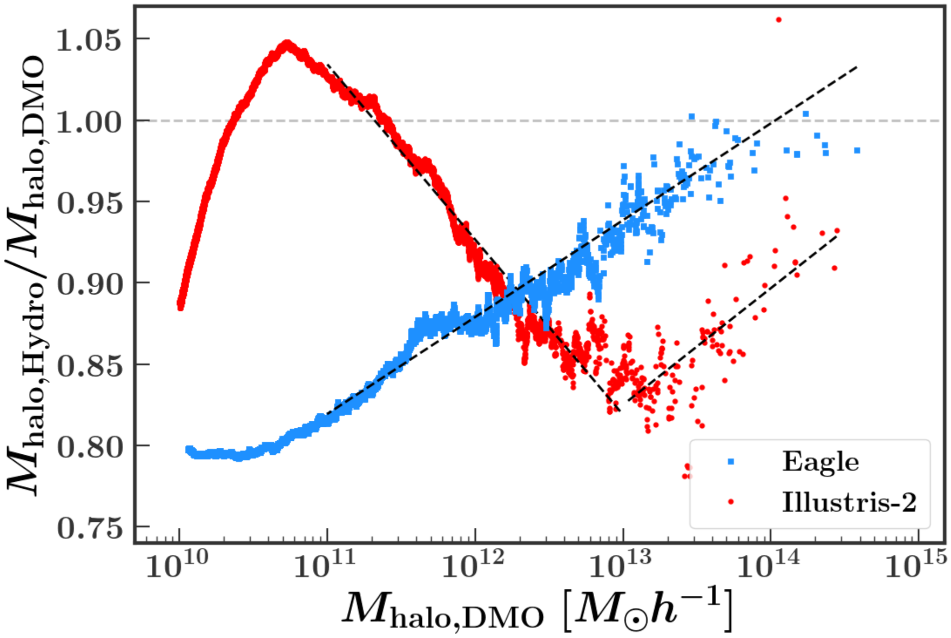

Figure 2.8: The ratio of halo masses from the hydrodynamic simulations to halo masses from the dark matter only simulations, as a function of dark matter only halo mass. Illustris-2 haloes are plotted in red and EAGLE haloes are plotted in blue. The halo mass is the total FoF mass from all particles, which in the hydrodynamic versions includes baryons. Hydrodynamic and dark matter only haloes are matched by their mass rank, rather than by position. The displayed ratio thus represents the correction factor needed to apply to the dark matter only haloes in order to recover the hydrodynamic mass function. The dashed black lines show simple fits to these relationships, down to $10^{11}h^{-1}M_{\odot}$, which we discuss in Section 2.8.

compared to Illustris. This is to be expected, since IllustrisTNG has weaker AGN feedback than the original Illustris simulation, which affects more massive haloes. The effect of feedback on lower mass haloes in TNG is stronger than that in Illustris due to the wind model used in TNG.

Figure 2.8 emphasizes the fact that the effect of baryons on the halo mass function is to decrease the HMF to lower masses. However, it is clear that this effect is very different in these two different simulations. The effect of baryons on the HMF in the EAGLE simulation is more prominent at lower masses, and the ratio of hydrodynamic halo mass to DMO halo mass increases almost linearly with log halo mass. In Illustris, the effect of baryons on the HMF is more prominent at higher masses, and the relationship is more complex than it is in EAGLE. In other words, the halo mass function is significantly affected by baryonic feedback processes, but there is no consensus among hydrodynamic simulations on what the correct feedback model is.

This halo mass function discrepancy presents a challenge when using an HOD framework to populate haloes from a dark matter only simulation with galaxies. The HOD parameters only describe how many galaxies to put in a halo of a given mass, but do not take into account how many haloes there are in a given mass bin. Therefore, because the dark matter only versions of Illustris and EAGLE have mass functions that are shifted to higher masses, there are more high mass haloes, so more galaxies are placed overall. Thus, even when applying the correct HOD parameters as extracted from the hydrodynamic simulation, the overall galaxy number density will be too high when this HOD is applied to the dark matter only simulation.

One possible solution to this is to adjust the HMF in the dark matter only simulation so that it is consistent with the HMF in the hydrodynamic version. We do this by identifying the most massive halo in the dark matter only simulation and assigning it the mass of the most massive halo in the hydrodynamic version, and then we do the same for the next most massive halo, and so on. In other words, we multiply the DMO halo masses in each simulation by their y-axis value in Figure 2.8. This process serves to isolate the effect of baryons on the halo mass function, allowing us to correct the DMO HMFs so that they agree with the HMFs from the hydrodynamic simulations. We note that this technique does not involve matching haloes based on position or particle-IDs. Because of this, we are not explicitly taking environment into account, so we are not correcting the conditional HMF. We have examined the conditional HMF in Illustris, however, and have found that the effect of baryons on the HMF only depends on environment at very high halo masses. Additionally, we have examined the effect on our clustering statistics if we use an environment-dependent HMF correction and find that the difference is negligible. We have also examined the halo correlation functions in Illustris and EAGLE in two different halo mass bins for the hydrodynamic simulations, the DMO simulations, and the corrected

DMO simulations, and have found that the corrected DMO halo correlation functions are in better agreement with the hydrodynamic halo correlation functions.

We now explore to what extent applying mass corrections to DMO halo masses improves the agreement between the clustering statistics of hydrodynamic and DMO+HOD galaxies. We first multiply each DMO halo mass by the correction shown in Figure 2.8 (i.e. we use our abundance matching technique for each halo as described above, and not the dashed-black fits shown in the figure). We then make new mock galaxy catalogues by applying the same best-fitting HOD (from Table 2.2) to our new mass-adjusted dark matter haloes. We thus have 1000 new mock catalogues for each sample. We then repeat the same procedure outlined in Sections 2.4 and 2.5 to get new clustering statistics and new p-values, which we list in Table 2.3 (in the rows labeled “Halo Mass Function”).

Figure 2.9 presents our p-values for the four samples (two simulations and two luminosity samples) for all six statistics we consider. The left-most point in each panel shows the original p-value we obtained and discussed in Section 2.5. The second point in each panel shows the new p-value we get after first applying a correction factor to the DMO halo masses. Horizontal dashed lines show the 1σ, 2σ, 3σ, 4σ, and 5σ tolerance levels. As we can see in Figure 2.9, after correcting the masses of haloes, our ability to accurately predict galaxy number density (top left panel) with our vanilla HOD model shows a drastic improvement for all samples. Thus, the vanilla form of HOD that we have adopted is sufficient for accurately (better than 2σ tolerance) predicting galaxy number density if it is applied to the correct population of haloes.

In addition to the improvement in our galaxy number density predictions for all samples, correcting the halo mass function yields a slight improvement to the other clustering statistics across all samples. For the M_r^{-21} samples, after correcting the halo mass function, all clustering statistics are at or better than the 2σ level. Thus, when applied to the correct halo population, the 5 parameter HOD model is able to accurately predict all clustering statistics for our high luminosity samples of galaxies. For the low luminosity samples, although the other clustering statistics do improve, most are still below the 3σ level, with the exception of the group multiplicity function in the EAGLE M_r^{-19} sample and the singular probability function in the Illustris M_r^{-19} sample. It is worth noting that the VPF does improve in all samples after correcting the halo mass function, indicating that part of the original VPF discrepancy was due to the incorrect number density. However, for the Illustris M_r^{-19} sample the VPF is still below the 3σ level, and for the EAGLE M_r^{-19} sample it is still well below 5σ, so we can conclude that not all of the issues with reproducing the VPF can be attributed to the number density.

These results indicate that although the HOD model for the brightest galaxies is successful when applied to the correct halo population, the HOD model for fainter galaxies is less successful, even when applied to the correct halo population. Thus, there must be some other assumptions in our HOD that are incorrect when applied to a low luminosity sample of galaxies. In the next section, we investigate possible extensions to our vanilla HOD.

2.7 Extensions of the HOD

2.7.1 Spatial bias

In our vanilla HOD model, we assume that each central galaxy lives at the centre of its halo, and that satellite galaxies trace the spatial distribution of dark matter within the halo. However, it is possible that these assumptions are incorrect, i.e. that galaxies exhibit spatial bias. More specifically, central spatial bias occurs when the central galaxy is not located at the centre of its halo, and satellite spatial bias occurs when the satellite galaxies do not trace the distribution of dark matter particles within their halo. To test for the presence of spatial bias, one option is to add spatial bias parameters to our HOD model and find a new best-fitting model that includes spatial bias. However, a simpler alternative is to remove the potential effects of spatial bias from the hydrodynamic simulation. If doing this yields better agreement between clustering statistics from our DMO+HOD mocks and

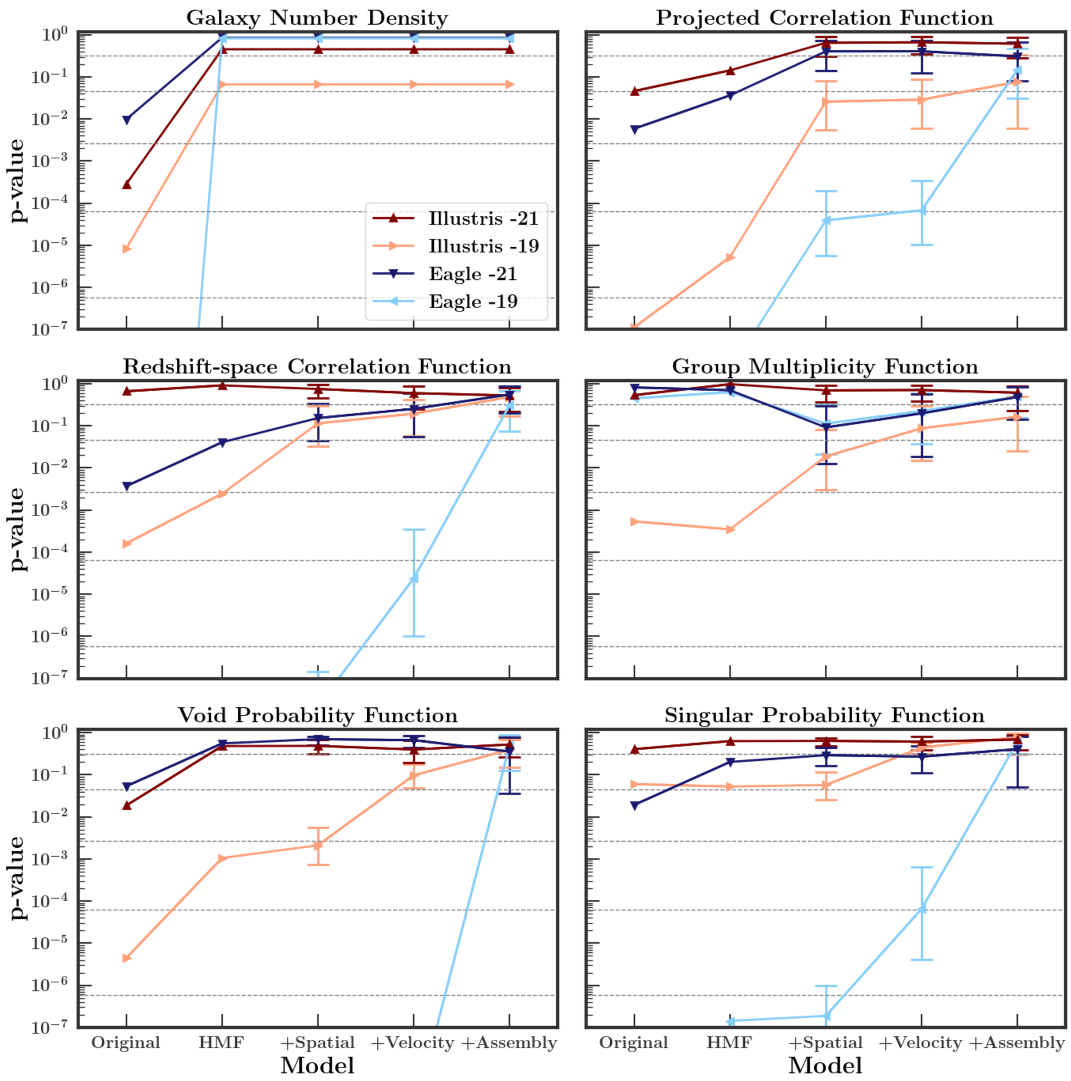

Figure 2.9: p-values from comparing the clustering of galaxies in hydrodynamic simulations to the clustering of mock galaxies in their dark matter only (DMO) counterparts. Each panel shows results for a different clustering statistic, as listed at the top of each panel. The dark red diamonds and dark blue squares represent the high luminosity samples of Illustris-2 and EAGLE, respectively, while the light red inverted triangles and the light blue triangles represent the low luminosity samples of Illustris-2 and EAGLE, respectively. The horizontal dashed gray lines denote the 1σ, 2σ, 3σ, 4σ, and 5σ confidence levels. The x-axis in each panel corresponds to different modifications to the haloes or to the galaxies in the simulations. From left to right, p-values are shown for (i) the original DMO+HOD model; (ii) the same DMO+HOD model after adjusting the DMO halo mass function to match the mass function in the hydrodynamic simulation; (iii) additionally removing satellite spatial bias from the hydrodynamic simulation galaxies; (iv) additionally removing central and satellite velocity bias from the hydrodynamic simulation galaxies; (v) additionally removing assembly bias from the hydrodynamic simulation galaxies. The last three p-values in each panel (with the exception of number density) are the median of many realizations (1000, 1000, and 4000), with error bars showing the 16th and 84th percentiles. For the low luminosity sample of EAGLE (light blue), several points are not shown because they fall below 10^{-7}. The values of these points are given in Tables 2.3.

the simulation galaxies, this would indicate that there is spatial bias in the hydrodynamic simulation, and therefore spatial bias parameters will need to be included in any future HOD modelling work to account for the possibility that there is spatial bias present in survey data.

We first test for the presence of central spatial bias. We do this by taking the Illustris and EAGLE galaxies identified as centrals and give them the position of their host halo, which is the position of the particle with the minimum gravitational potential energy. We do this without changing any central velocity information or any satellite galaxy information, in order to isolate the effect of central spatial bias. Thus, if there is any central spatial bias present in the original simulation, this procedure would remove it, yielding better agreement with our HOD model. The results of this show no change for either simulation or sample, indicating that any central spatial bias has a negligible impact on clustering statistics.

We next test for the presence of satellite spatial bias. We do this by taking the galaxies identified as satellites in the hydrodynamic simulations and assigning them the positions of random dark matter particles in their host halo (also in the hydrodynamic simulations). We do this without changing any satellite velocity information or any central galaxy information, in order to isolate the effect of satellite spatial bias. We repeat this process 1000 times, in order to generate 1000 different realizations of our simulation with satellite spatial bias removed. We can therefore generate 1000 different p-values for each clustering statistic. Table 2.3 (rows labeled “Satellite Spatial Bias”) lists the median p-values from these 1000 realizations of our simulation with satellite spatial bias removed. We note that it is possible that placing satellite galaxies on dark matter subhaloes rather than particles would alleviate some of the tension that we see between our HOD and the hydrodynamic simulations. However, traditional HOD models do not use subhaloes, in part because the DMO simulations to which they are applied often do not have high enough resolution to resolve small subhaloes. Therefore, we do not explore the option of placing satellite galaxies on dark matter subhaloes in this analysis, but note that it would be worth investigating in future work.

The third point in each panel of Figure 2.9 shows these median p-values that result from both correcting the DMO halo masses and removing satellite spatial bias from the hydrodynamic simulations. Error bars show the range of p-values that correspond to the middle 68% of our 1000 realizations with satellite spatial bias removed. We can see that the M_r^{-21} samples show either slight improvement or no change after removing satellite spatial bias, while the M_r^{-19} samples show significant improvement. In particular, the projected and redshift-space correlation functions are much improved in the M_r^{-19} samples of both EAGLE and Illustris. From these results, we can conclude that the galaxies in EAGLE and Illustris do exhibit satellite spatial bias, the effects of which are more prominent when considering low luminosity galaxies. We can also conclude that the effects shown are definitively the results of spatial bias and not a difference in halo profile due to the presence of baryons; if the clustering differences were due to a difference in halo density profile when baryons are included versus when they are not, then giving the satellite galaxies the positions of random dark matter particles in the halo (in the hydrodynamic simulation) would not have a significant effect on clustering.

The extent and nature of the satellite spatial bias is similar in the two different simulations. In Figure 2.5, it is clear from looking at both the projected and redshift-space correlation functions that Illustris M_r^{-19} galaxies are less clustered on small scales than the DMO+HOD mock galaxies, or in other words, Illustris galaxies are less concentrated than the dark matter. When satellite spatial bias is removed, the satellite galaxies become more concentrated, and are thus a better fit to the HOD on small scales. The picture looks a bit different in Figure 2.6, where EAGLE M_r^{-19} galaxies are less clustered than DMO+HOD mock galaxies on small scales. However, this amplitude difference in the correlation functions extends to large scales and is thus not caused by satellite spatial bias (it is caused by assembly bias, as we will see later). If we examine the slopes of the correlation functions at small scales in Figure 2.6, we see that EAGLE M_r^{-19} galaxies have a shallower slope than DMO+HOD, which means that they are less concentrated within their haloes (Berlind and Weinberg, 2002), similar to Illustris M_r^{-19} galaxies.

Despite the improvement that we see in Figure 2.9 when removing spatial bias, many clustering statistics for the M_r^{-19} samples are still not well predicted by our HOD model, even after correcting the halo mass function and removing satellite spatial bias from the simulations. This is especially true for EAGLE M_r^{-19} galaxies, where all statistics except number density and group multiplicity function still show a significant discrepancy between hydrodynamic and DMO+HOD galaxies.

2.7.2 Velocity bias

The vanilla HOD model also assumes that each central galaxy moves with the mean velocity of its halo (i.e. there is no central velocity bias), and that satellite galaxies trace the velocity distribution of dark matter within their halo (i.e. there is no satellite velocity bias). Once again, it is possible that these assumptions are incorrect, due to the effects of phenomena such as mergers, dynamical friction, and tidal stripping.

To test for the presence of central velocity bias, we take the Illustris and EAGLE galaxies identified as centrals and assign them the velocity of their host halo. By doing this, we are removing the possibility that the central galaxy might not be at rest with respect to its host halo. In Illustris, this is the sum of the mass weighted velocities of all particles/cells in the group, multiplied by $1/a$. (In EAGLE, the velocity of the parent halo is not provided, so this test is not possible. Central galaxies already have the velocity of the central subhalo.) As in the case of central spatial bias, removing central velocity bias has a negligible effect on the clustering statistics we consider.

To remove satellite velocity bias, we take the hydrodynamic simulation galaxies identified as satellites and assign them the velocities of random dark matter particles in the halo. We do this in combination with other effects (e.g. central velocity bias, central spatial bias, satellite spatial bias). In other words, we take the central galaxy and give it the position and velocity of its host halo, and we take satellite galaxies and give them the positions and velocities of randomly chosen dark matter particles in the halo, so that all spatial and velocity bias has been removed from the simulation galaxies. We repeat the random selection of dark matter particles 1000 times, so that we ultimately generate 1000 different realizations of the simulation galaxies after removing all spatial and velocity bias. The results of this are shown in Table 2.3, where the p-values given are the median of 1000.

The fourth point in each panel of Figure 2.9 shows these median p-values that result from correcting DMO halo masses and removing spatial and velocity bias from the hydrodynamic simulations. Once again, error bars show the range of p-values that correspond to the middle 68% of our 1000 realizations with satellite spatial and velocity bias removed. The figure shows that removing velocity bias provides an additional improvement for our clustering statistics for the M_r^{-19} samples. In particular, the Illustris M_r^{-19} sample shows significant improvement in the void probability function and slight improvement in all other clustering statistics. All statistics now show no significant discrepancy between the hydrodynamic galaxies and our DMO+HOD model. The EAGLE M_r^{-19} sample shows improvement in the redshift-space correlation function, as well as both counts-in-cells statistics. It is to be expected that number density does not change when spatial and velocity bias are removed, because the number of galaxies is not affected. Additionally the projected correlation function is by design not affected by velocity, so it is not surprising that there is no change after removing velocity bias. Despite these improvements, the differences between the statistics of EAGLE M_r^{-19} galaxies and the DMO+HOD model are still highly significant.

At this point, after removing all spatial and velocity bias from our simulations, all statistics are well predicted ($< 2\sigma$ tension) by our HOD model for the Illustris M_r^{-19} sample, while the number density and group multiplicity function are well predicted ($< 2\sigma$ tension) for the EAGLE M_r^{-19} sample. However, the correlation functions and counts-in-cells statistics are still not well predicted for EAGLE M_r^{-19}. This indicates the possibility that the number of galaxies in a halo may depend on a halo property other than mass, such as age or concentration. This will be discussed in the next section.

2.7.3 Assembly/secondary bias

Halo assembly/secondary bias is the phenomenon whereby halo clustering depends on a secondary parameter, such as age or concentration, at fixed halo mass (e.g., Gao et al., 2005; Wechsler et al., 2006; Salcedo et al., 2018). If the number of galaxies in a halo depends on this secondary parameter, the clustering of galaxies will inherit this additional halo clustering, a phenomenon known as galaxy assembly bias (e.g., Croton et al., 2007; Zentner et al., 2014). Galaxy assembly bias could be present in Illustris or EAGLE, but it is explicitly not present in our DMO+HOD model. We now remove any effects of assembly bias from our hydrodynamic simulation galaxies, with the understanding that if this procedure improves our ability to predict clustering statistics with our DMO+HOD model, this is an indication that future HOD modelling should incorporate parameters that deal with assembly bias.

To remove the presence of assembly bias from our simulation galaxies, we identify pairs of haloes with similar masses, and swap the positions and velocities of their galaxies. This is done after already removing all spatial and velocity bias. In other words, we first generate 1000 realizations of the simulation galaxies after removing spatial and velocity bias (as described above), and then exchange galaxies in haloes of similar mass. When we exchange galaxies in pairs of haloes, we shift the galaxy positions by the difference in halo centre positions, so that a galaxy is in the same position relative to the halo centre, but the halo centre has been switched. For the velocities, we take the peculiar velocity of a galaxy and subtract the mean halo velocity, thus putting the galaxy in the frame of the halo. We then add this velocity to the velocity of the new halo to get the new velocity of the galaxy. In other words, we keep the velocity of the galaxy in the frame of its halo the same, and simply give it a new halo velocity. We use four different combinations of halo pairs, ultimately resulting in 4000 realizations of our simulation galaxies after removing all spatial, velocity, and assembly bias.

This procedure of exchanging galaxies in haloes of similar mass effectively removes assembly bias from our data because it nullifies any environmental effects on the number of galaxies in each halo. If the number of galaxies in each halo was already only dependent on halo mass, then this procedure should not produce any change in clustering statistics. However, if the number of galaxies in a halo had a dependence on a property other than halo mass, then swapping galaxies in haloes with similar masses would remove the effect of this phenomenon on our clustering statistics. The results of this are detailed in Table 2.3. Once again, the p-values given are the median of many realizations (in this case 4000).

The last point in each panel of Figure 2.9 shows these median p-values that result from removing assembly bias (in addition to correcting the HMF and removing all spatial and velocity bias). Once again, error bars show the range of p-values that correspond to the middle 68% of our 4000 realizations. Removing assembly bias results in all clustering statistics being well predicted by our HOD for both simulations and luminosity samples. In the M_r^{-21} samples, all clustering statistics were already well predicted, so there is very little change. More importantly, in the M_r^{-19} samples, there are slight improvements in all clustering statistics for the Illustris galaxies, and there are major improvements for the correlation functions and counts-in-cells statistics for the EAGLE galaxies. Of particular note is the void probability function for the EAGLE M_r^{-19} sample, which remained below 5σ until assembly bias was removed, at which point it reached 1σ confidence that the HOD model is a good fit to the simulation. This agrees with the results of Chaves-Montero et al. (2016), who detected galaxy assembly bias in the EAGLE simulation, and found that the signature of assembly bias was stronger for low mass galaxies. This is also consistent with the results of Tinker et al. (2006a), which suggested that VPF is sensitive to the presence of assembly bias. More recently, Walsh and Tinker (2019) also showed that counts-in-cells statistics can be powerful probes of assembly bias.

2.8 Summary and Discussion

In this work, we have examined the validity of using halo occupation distribution modelling to reproduce galaxy clustering statistics. Halo models provide a simple and computationally inexpensive way to investigate the connection between galaxies and their dark matter haloes, but they rely on the assumption that the role of baryons can be easily parametrized in the modelling procedure. Using two different hydrodynamic simulations, Illustris-2 and EAGLE, we have investigated the accuracy of using a simple five-parameter HOD to reproduce clustering when applied to a high luminosity sample of galaxies as well as a low luminosity sample. The HOD was fit to each simulation and luminosity sample separately, and applied to haloes from the dark matter only counterparts of Illustris and Eagle to create mock galaxy catalogues. Our clustering statistics were measured in the same way on our simulation galaxies as they were on our mock catalogues. Our main results are the following:

- Overall, the vanilla HOD model is more successful when applied to a high luminosity sample of galaxies than it is when applied to a low luminosity sample of galaxies.

- The simple five-parameter HOD model is able to accurately (within 3σ tolerance) reproduce correlation functions, the group multiplicity function, the void probability function, and the singular probability function, for the high luminosity sample of galaxies in both Illustris and EAGLE, as well as the number density in EAGLE.

- In our M_r^{-21} Illustris and EAGLE samples, the errors among our 1000 mocks are much larger than the cosmic variance errors from the 400 SDSS-like mocks. In other words, an SDSS size M_r^{-21} survey would perhaps be sensitive to clustering differences that we are unable to detect in our analysis. In our M_r^{-19} Illustris and EAGLE samples, the errors among our 1000 mocks are smaller than the cosmic variance errors from the 400 SDSS-like mocks. This means that a survey with the precision of SDSS might not be able to detect the differences that we find between hydrodynamic galaxies and the HOD model. A future survey like the Dark Energy Spectroscopic Instrument (DESI, DESI Collaboration et al., 2016), however, will have better precision than the SDSS due to its larger volume, allowing it to potentially detect these small differences in clustering measurements.

- In general, the halo mass function is shifted to higher masses when baryons are not included, resulting in an over prediction of galaxy number density when an HOD is applied to the haloes from the dark matter only simulations. After correcting the dark matter only halo mass function, the vanilla HOD model is able to accurately reproduce all clustering statistics in the high luminosity sample of galaxies in both Illustris and EAGLE. It also able to accurately reproduce galaxy number density in both low luminosity samples.

- Even after correcting the halo mass function, the vanilla HOD model is still unable to accurately (within 3σ tolerance) reproduce most of the other five clustering statistics for the low luminosity samples of galaxies in Illustris-2 and EAGLE. However, after removing the potential effects of spatial, velocity, and assembly bias from the galaxies in the original simulations, the HOD model (with mass function correction) is able to accurately reproduce all clustering statistics in both samples and both simulations.

These results demonstrate the prominent differences between the EAGLE and Illustris simulations, in terms of the ways that baryons affect halo masses and galaxy clustering. For example, the EAGLE and Illustris simulations are very different in terms of the amount of spatial, velocity, and assembly bias they exhibit. Additionally, neither EAGLE nor Illustris reproduces the galaxy luminosity function from the SDSS. Therefore, we cannot use the results from our analysis of the clustering in these two hydrodynamic simulations to draw conclusions about galaxy clustering in the real Universe. Because of this, we do not attempt to infer the true amounts of spatial, velocity, and assembly bias in the real Universe based on this work, but rather recommend that any future work involving HOD modelling should include free parameters for these biases. Moreover, our work suggests that future work aiming to use HOD modelling to study cosmology would benefit from focusing on high luminosity galaxy samples, which seem to be less affected by the aforementioned biases.

Additionally, different clustering statistics are sensitive to different biases. For example, the void probability function seems to be particularly sensitive to the presence of assembly bias, while the redshift space correlation function is sensitive to satellite velocity bias, as can be seen in the low luminosity sample of EAGLE galaxies. Therefore, properly constraining HOD parameters (especially when including spatial, velocity, and assembly bias parameters), necessitates measuring several different clustering statistics.

Of particular note is the difference in how baryons alter the halo mass function between the two different simulations. Any future work hoping to use HOD modelling will have to first correct the dark matter only haloes by shifting the mass function to lower masses, so that it more closely resembles what the mass function would look like with baryons included in the simulation. However, the exact nature of this correction to the halo mass function clearly depends upon which hydrodynamic prescriptions are regarded as the truth. The large difference that we see between the two simulations in Figure 2.8 demonstrates the extent to which mass corrections depend on the details of supernova and AGN feedback physics. This result is somewhat alarming because, unlike the other biases we examine in this study, the effect of baryons on the halo mass function cannot be easily parametrized, making it unclear how one must proceed with halo modelling of observed clustering measurements.

At a minimum, we recommend that future halo modelling efforts repeat their analyses a couple times, applying different corrections to the dark matter only halo masses. This will provide a rough estimate of the systematic uncertainty due to baryonic effects on the

Table 2.4: Our fits to the halo mass ratios in Illustris-2 and EAGLE, as well as TNG100-2. In the third column, x is equal to $\log M_{\rm halo}$.

Simulation	Mass Range	$M_{\rm halo,Hydro}/M_{\rm halo,DMO}$
Illustris-2	$1.00 \cdot 10^{11} < M < 9.57 \cdot 10^{12}$	$-0.10771x + 2.21907$
Illustris-2	$M > 9.57 \cdot 10^{12}$	$0.07174x - 0.10774$
EAGLE	$M > 1.00 \cdot 10^{11}$	$0.05956x + 0.16413$
TNG100-2	$1.00 \cdot 10^{10} < M < 2.74 \cdot 10^{12}$	$-0.10171x^2 + 2.37863x - 12.97684$
TNG100-2	$2.74 \cdot 10^{12} < M < 1.06 \cdot 10^{13}$	$0.00189x + 0.84450$
TNG100-2	$M > 1.06 \cdot 10^{13}$	$0.09429x - 0.35479$

halo mass function. For example, if a study finds strong evidence of assembly bias when applying no correction to the halo masses, but then the evidence disappears when the analysis is repeated using a mass correction, one should not claim any detection of assembly bias. To facilitate such a procedure, we fit simple functions to the mass corrections shown in Figure 2.8. In the case of EAGLE we fit a single line, while for Illustris we fit a broken line. These fits are shown as dashed lines in Figure 2.8. In Table 2.4 we list the parameters for these fits to the mass corrections in Illustris and EAGLE.

We have tested these fits and confirmed that they produce the same results as doing the full abundance matching correction that we performed in our analysis. Additionally, we present fits to the same mass correction in IllustrisTNG (Naiman et al., 2018; Pillepich et al., 2018a; Nelson et al., 2018; Marinacci et al., 2018; Springel et al., 2018). TNG is more recent than both Illustris and EAGLE, and makes use of updated feedback mechanisms, which results in a halo mass correction that is different than what we see in either Illustris or EAGLE. We make no assumptions about which of these simulations produces the correct relationship between the masses of their hydrodynamic and DMO haloes, but we recommend that future halo modelling work makes use of one or more of these corrections.

Rather than viewing these results as evidence that dark matter only simulations are insufficient for halo modelling and should thus not be used to study galaxy clustering, we interpret these results as confirmation that there is no consensus among hydrodynamic simulations. Therefore, dark matter only simulations and halo models are still very relevant tools for investigating the galaxy-halo connection, as long as the halo model is given sufficient freedom, and the effect of baryons on the halo mass function is accounted for.

CHAPTER 3

The impact of baryonic physics on the abundance, clustering, and concentration of halos

This chapter was previously published in the November 2021 edition of The Astrophysical Journal (Beltz-Mohrmann and Berlind, 2021) and is reproduced here, with minor formatting changes, with the permission of the publisher and my co-author, Andreas A. Berlind.

We examine the impact of baryonic physics on the halo distribution in hydrodynamic simulations compared to that in dark matter only (DMO) simulations. We find that, in general, DMO simulations produce halo mass functions (HMF) that are shifted to higher halo masses than their hydrodynamic counterparts, due to the lack of baryonic physics. However, the exact nature of this mass shift is a complex function of mass, halo definition, redshift, and larger-scale environment, and it depends on the specifics of the baryonic physics implemented in the simulation. We present fitting formulae for the corrections one would need to apply to each DMO halo catalogue in order to reproduce the HMF found in its hydrodynamic counterpart. Additionally, we explore the dependence on environment of this HMF discrepancy, and find that, in most cases, halos in low density environments are slightly more impacted by baryonic physics than halos in high density environments. We thus also provide environment-dependent mass correction formulae that can reproduce the conditional, as well as global, HMF. We show that our mass corrections also repair the large-scale clustering of halos, though the environment-dependent corrections are required to achieve an accuracy better than 2%. Finally, we examine the impact of baryonic physics on the halo mass - concentration relation, and find that its slope in hydrodynamic simulations is consistent with that in DMO simulations. Ultimately, we recommend that any future work relying on DMO halo catalogues incorporate our mass corrections to test the robustness of their results to baryonic effects.

3.1 Introduction

Studying the connection between galaxies and their dark matter halos is one of the keys to understanding galaxy formation and evolution, as well as constraining cosmological models. In recent years, using hydrodynamic simulations has become a popular method for investigating this connection (e.g. Vogelsberger et al., 2014a). However, these simulations are not only computationally expensive, but are also inconsistent. We currently lack a consensus on the correct baryonic physics prescriptions to use; thus, different hydrodynamic simulations produce widely varying results.

By contrast, dark matter only (DMO) simulations are much less computationally expensive, and although the only physics involved is gravity, they still allow us to predict the large-scale distribution of matter in the universe. However, a DMO simulation produces a halo mass function (HMF) that on average is shifted to higher masses than the HMF produced by a hydrodynamic simulation with the same cosmology and initial conditions (e.g., Cui et al., 2014). This is because in a hydrodynamic simulation the presence of baryonic physics like stellar feedback, star formation, and feedback from active galactic nuclei (AGN) has an impact on the masses of dark matter halos. This effect is non-trivial: the impact of baryonic physics varies with halo mass, as well as halo definition, environment, and redshift. And of course, the effect of baryonic physics depends on the details of the feedback prescriptions implemented, which varies from one hydrodynamic simulation to the next.

Several recent works have focused on comparing hydrodynamic simulations to models of the galaxy-halo connection (e.g. Hadzhiyska et al., 2020, 2021a,c). Beltz-Mohrmann et al. (2020) examined the accuracy of halo occupation distribution (HOD) modeling compared to hydrodynamic simulations. They extracted HOD models from the galaxies in the Illustris and EAGLE simulations, and subsequently applied these HOD models to the corresponding DMO simulations (i.e. Illustris-Dark and EAGLE-Dark). They found that these HOD models, when applied to the DMO halos, were unable to reproduce the galaxy clustering seen in the hydrodynamic simulations. This was in part because the HMFs in the DMO simulations were shifted to higher masses, leading to an overestimate in the overall number density of galaxies, as well as discrepancies in other clustering statistics (e.g. correlation function and void probability function). The authors found that applying a correction to the DMO halo masses led to a significant improvement in the galaxy clustering.

This issue with DMO halo mass functions could present a challenge for any work attempting to use halo modeling to constrain the galaxy-halo connection (e.g. Sinha et al., 2018), as well as any work attempting to constrain cosmology (e.g. Lange et al., 2019b). If one's goal is merely to produce mock galaxy catalogues with the same clustering as seen in survey data, then one does not need to worry about this mass discrepancy; in this case, the parameterization of the galaxy-halo connection is unimportant, so long as the clustering mimics the real universe. However, if one's goal is to *accurately* constrain the galaxy-halo connection, then this mass discrepancy presents a problem. Moreover, if one is attempting to use DMO simulations plus halo modeling to constrain cosmology, then obtaining the correct halo mass function is essential, unless the HOD is able to fully absorb the effect without perturbing the resulting cosmological parameters.

Several previous works have investigated the effects of baryonic physics on the halo mass function. Cui et al. (2012) compared three simulations: a dark matter only simulation, a hydrodynamic simulation with non-radiative physics, and a hydrodynamic simulation with radiative processes. They found that the fractional difference between halo masses in the hydrodynamic and DMO simulations is almost constant for halos above $\log(M_{\Delta_c}/h^{-1}M_{\odot}) > 13.5$, but that for higher overdensity halos, as well as smaller mass halos, differences in halo mass appear which depend on halo mass as well as baryonic physics.

Later, Cui et al. (2014) further examined the effects of baryonic physics on the halo mass function, focusing on the role of AGN feedback. They found that for both friends-of-friends (FoF) and spherical overdensity (SO) halos, AGN feedback suppresses the HMF to a level below that of DMO simulations. They found that the ratio between the HMFs in the hydrodynamic and DMO simulations increases with overdensity, but does not have any redshift or mass dependence. They found that halos in hydrodynamic simulations have shallower inner density profiles, which lends them to halo mass loss caused by "the sudden displacement of gas induced by thermal AGN feedback." The authors provide fitting functions to describe the halo mass variations between the full-physics and DMO simulations at different overdensities, which can recover the HMFs from hydrodynamic simulations for halo masses larger than $10^{13}h^{-1}M_{\odot}$.

Sawala et al. (2013) examined the effect of baryons on the abundance of structures and substructures and found that halo masses are reduced for halos smaller than $10^{12}M_{\odot}$, and the effect grows with decreasing mass. Later, van Daalen et al. (2014) looked at the impact of baryonic processes on the two-point correlation functions of galaxies, subhalos, and matter in large hydrodynamic simulations, and found that the changes due to the inclusion of baryons are not limited to small scales. The authors found that the large-scale effects are due to the change in subhalo mass caused by feedback associated with galaxy formation. They concluded that predictions of galaxy-galaxy and galaxy-mass clustering from models based on collisionless simulations will have errors greater than 10% on scales < 1 Mpc, unless the simulation results are modified to account for the effects of baryons on the distributions of mass and satellites.

Velliscig et al. (2014) used hydrodynamic simulations from the OverWhelmingly Large Simulations (OWLS) project to study how the physical processes related to galaxy formation (e.g. star formation, supernova and AGN feedback, etc.) impact the properties of halos. They found that the "gas expulsion and associated dark matter expansion induced by supernova-driven winds are important for halos with masses $M_{200} < 10^{13}M_{\odot}$, lowering their masses by up to 20% relative a DM-only model." They also found that AGN feedback impacts halo masses up to cluster scales ($M_{200} \sim 10^{15}M_{\odot}$). Moreover, they found that baryonic physics alters the total mass profiles of halos out to several times the virial radius, which cannot be explained by a change in halo concentration. They concluded that the decrease in total halo mass leads to a decrease in the HMF of 20%. The authors provided analytic fitting formulae to correct halo masses and mass functions from DMO simulations.

Subsequently, Khandai et al. (2015) investigated the properties of halos in the MassiveBlack-II hydrodynamical simulation, and found that baryons strongly affect the halo mass function when compared to dark-matter-only simulations, while Schaller et al. (2015) examined the effects of baryons on halos in the EAGLE simulation at low masses, and Chaves-Montero et al. (2016) used subhalo abundance matching to investigate the effect of baryons on the halo occupation distribution and assembly bias in the EAGLE simulation.

Bocquet et al. (2016) used the Magneticum simulations to investigate the impact of baryons on the halo mass function, and found that baryonic effects globally decrease the masses of galaxy clusters, which, at a given mass, results in a decrease of their number density. They found that this effect disappears at high redshift ($z > 2$) and for high halo masses above $10^{14}h^{-1}M_{\odot}$. They concluded that when using a survey like eROSITA, ignoring the impact of baryonic physics on the halo mass function leads to an underestimate in Ω_m by about 0.01. The authors also provided HMF fitting formulae.

Despali and Vegetti (2017) examined the impact of baryonic physics on the subhalo populations in EAGLE and Illustris, and found that the presence of baryons reduces the number of subhalos, especially at the low-mass end. They found that the variations in the subhalo mass function depend on those in the halo mass function, which is shifted by the effect of stellar and AGN feedback.

Finally, Castro et al. (2021) (and earlier Balaguera-Antolínez and Porciani (2013)) addressed the impact of baryonic physics on clusters. They found that ignoring baryonic effects on the halos mass function and halo bias could significantly alter cosmological parameter constraints, particularly in the upcoming generations of galaxy cluster surveys.

Most of these works concur that baryonic physics lead to a net reduction in the masses of halos, and consequently the HMF. However, the magnitude and mass dependence of the effect differ between different works, likely because of the different hydrodynamic simulations used. This calls for a systematic study that compares the impact of baryons on the HMF across a variety of recent hydrodynamic simulations with different baryonic physics and feedback prescriptions, and over a wide range of redshifts and halo definitions. Furthermore, no previous work has examined the environmental dependence of baryonic effects on the halo mass function, nor simultaneously examined the impact of baryonic physics on halo clustering and halo concentrations as a function of mass, all of which are important ingredients for halo models of galaxy clustering.

In this work, we investigate the impact of baryonic physics on the halo mass function in three different simulations: Illustris, IllustrisTNG, and EAGLE. We consider three different redshifts and five different halo definitions. Moreover, we investigate the environmental dependence of the effect of baryons on the halo mass function. For all these cases, we provide formulae that can be used to correct the halo masses from DMO simulations to match the halo mass functions from their hydrodynamic counterparts. Finally, we investigate the effect of baryonic physics on halo clustering and on the halo mass - concentration relation.

We discuss the simulation data in Section 3.2. In Section 3.3 we discuss the effects of baryons on halo populations, and in Section 3.4 we discuss our corrections. In Section 3.5 we discuss the environmental dependence of the halo mass function, and in Section 3.6 we discuss the halo mass - concentration relation. Finally, in Section 3.7 we summarize our results and conclusions.

3.2 Simulation Data

For this analysis we use three cosmological N-body simulations: Illustris (Nelson et al., 2015; Vogelsberger et al., 2014b,a; Genel et al., 2014), IllustrisTNG (Marinacci et al., 2018; Naiman et al., 2018; Nelson et al., 2018; Pillepich et al., 2018b,a; Springel et al., 2018), and EAGLE (Schaye et al., 2015; McAlpine et al., 2016; The EAGLE team, 2017; Springel, 2005; Crain et al., 2015). Specifically, we use Illustris-1, TNG100-1, TNG300-1, and

Table 3.1: The columns show (from left to right): simulation name, box size in (h^{-1}Mpc), number of dark matter particles, dark matter particle mass (for the hydrodynamical run) in ($h^{-1}M_{\odot}$), gas particle mass in ($h^{-1}M_{\odot}$), redshift used, and cosmological parameters. The dark matter particle mass for Illustris-1-Dark is $5.3 \cdot 10^{6}(h^{-1}M_{\odot})$; for TNG100-1-Dark it is $6.0 \cdot 10^{6}(h^{-1}M_{\odot})$; for TNG300-1-Dark it is $7.0 \cdot 10^{7}(h^{-1}M_{\odot})$; and for EAGLE Dark it is $7.5 \cdot 10^{6}(h^{-1}M_{\odot})$.

Sim.	L_{box}	N_{DM}	m_{DM}	m_{gas}	h	Ω_m	Ω_Λ	Ω_b	σ_8
Illustris-1	75	1820^3	$4.4 \cdot 10^6$	$9.2 \cdot 10^5$	0.704	0.2726	0.7274	0.0456	0.809
TNG100-1	75	1820^3	$5.1 \cdot 10^6$	$9.4 \cdot 10^5$	0.6774	0.3089	0.6911	0.0486	0.8159
TNG300-1	205	2500^3	$4.0 \cdot 10^7$	$7.6 \cdot 10^6$	0.6774	0.3089	0.6911	0.0486	0.8159
EAGLE	67.77	1504^3	$6.6 \cdot 10^6$	$1.2 \cdot 10^6$	0.6777	0.307	0.693	0.04825	0.8288

EAGLE RefL100N1504. Each of these hydrodynamic simulations has a corresponding dark matter only counterpart. A summary of the simulation parameters can be found in Table 3.1.

The Illustris and IllustrisTNG simulations were performed with the AREPO code, while the EAGLE simulation was performed with GADGET-3. All three simulations model star formation, stellar evolution, gas cooling and heating, supernovae feedback, black hole formation, and AGN feedback. GADGET-3 and AREPO contain different numerical hydrodynamical techniques, leading to large discrepancies in the galaxy populations they produce (Scannapieco et al., 2012). Additionally, though Illustris and TNG were performed with the same code, they vary in the strength of their feedback prescriptions. Illustris contains much stronger AGN feedback than either the TNG or EAGLE simulations (Weinberger et al., 2017), which leads to substantial differences in their results.

We choose to compare these three simulations because they are publicly available and all have dark matter only (DMO) counterparts to their hydrodynamic simulations. Moreover, each simulation has adequate resolution and volume to allow us to examine its halo mass function between 10^{10} and $10^{14}h^{-1}M_{\odot}$. Finally, the three simulations have different baryonic physics prescriptions, which allows us to compare how variations in physics impacts our results.

In this work, we utilize halo catalogues from each simulation at $z = 0, 1$, and 2. In all three simulations, halos were identified using a standard friends-of-friends (FoF) algorithm (Davis et al., 1985) with a linking length of $b = 0.2$ times the mean interparticle separation. The FoF algorithm was run on the dark matter particles, and in the hydrodynamic simulations baryonic particles were attached to the same FoF group as their nearest dark matter particle. Each FoF group has a mass, which we refer to in this work as M_{FoF}. Additionally, for each FoF group, masses were calculated for several different spherical overdensity (SO) definitions, which we will make use of in this paper. Specifically, these halo definitions are M_{200b} (the total mass of the group enclosed in a sphere whose mean density is 200 times the mean density of the Universe at the time the halo is considered), M_{200c} and M_{500c} (the total mass of the group enclosed in a sphere whose mean density is 200 or 500 times the critical density of the Universe at the time the halo is considered), and M_{vir} (the total mass of this group enclosed in a sphere whose mean density is Δ_c times the critical density of the Universe at the time the halo is considered, where Δ_c derives from the solution of the collapse of a spherical top-hat perturbation from Bryan and Norman 1998.) For each halo definition and redshift, we only consider halos above $> 10^{10} h^{-1} M_{\odot}$. In Illustris, the number of halos above this threshold ranges from about 60 to 101 thousand, while in IllustrisTNG, the number ranges from about 70 to 120 thousand, and in EAGLE, the number ranges from about 53 to 85 thousand. The exact number of halos depends on the halo definition and redshift, but it generally is lowest for M_{500c} halos at $z = 2$ and is highest for M_{FoF} halos at $z = 1$.

3.3 The Effect of Baryons on the Halo Mass Function

The most straightforward way to investigate the population of halos in a simulation is by looking at the halo mass function (HMF), which displays the abundance of halos as a function of mass. In Figure 3.1, we show the halo mass functions for M_{200b} halos at $z = 0$ for both the hydrodynamic and dark matter only versions of Illustris-1, TNG100-1, and EAGLE. Illustris is plotted in blue, TNG in green, and EAGLE in orange. It should be noted that the halo masses from the hydrodynamic simulations include both dark matter particles and baryonic particles. The hydrodynamic versions are plotted with dashed lines, while the DMO versions are plotted with solid lines. A corrected version of the DMO HMF is plotted with a dotted line and will be discussed further in the next section. In the residual panel, we plot the ratio of the hydrodynamic HMF to the DMO HMF (solid), as well as the ratio of the hydrodynamic HMF to the corrected DMO HMF (dotted) for each simulation.

The residuals reveal sizeable discrepancies between the DMO and hydrodynamic halo mass functions. In general, the hydrodynamic HMFs are shifted to lower masses in all three simulations, but this shift is mass dependent, and varies by simulation. In particular, the Illustris (blue) hydrodynamic HMF is consistently lower than the Illustris-Dark HMF above $10^{12}h^{-1}M_{\odot}$, as well as below $10^{11}h^{-1}M_{\odot}$. In TNG100 (green), the hydrodynamic HMF is lower than the DMO HMF between 10^{12} and $10^{14}h^{-1}M_{\odot}$, as well as below $10^{12}h^{-1}M_{\odot}$. In EAGLE (orange), the hydrodynamic HMF is lower than the DMO HMF at all halo masses below about $10^{13}h^{-1}M_{\odot}$.

In all three simulations, the discrepancies between the hydrodynamic and DMO halo mass functions reach the twenty percent level for much of the halo mass range. We emphasize that not only is this a significant difference, but it is also not trivial to correct. It is not equivalent to a difference in halo definition, which would be a constant offset and would not vary by simulation; this effect exhibits a trend with halo mass, and depends on the feedback implemented in the hydrodynamic simulation.

Our HMF results for Illustris are consistent with the findings of Vogelsberger et al. (2014b), who found that the halo mass function in Illustris is most affected at halo masses below $10^{10}h^{-1}M_{\odot}$ and above $10^{12}h^{-1}M_{\odot}$, where baryonic feedback processes (e.g. reionization and SN/AGN feedback) are strongest, leading to a reduction in halo mass compared to their DMO counterparts. They also found that halos around $10^{11}h^{-1}M_{\odot}$, where star formation is most efficient, tend to be more massive than their DMO counterparts.

Our HMF results for TNG100 are consistent with the findings of Springel et al. (2018), who found that baryons in the TNG simulation have a larger impact on low mass halos and

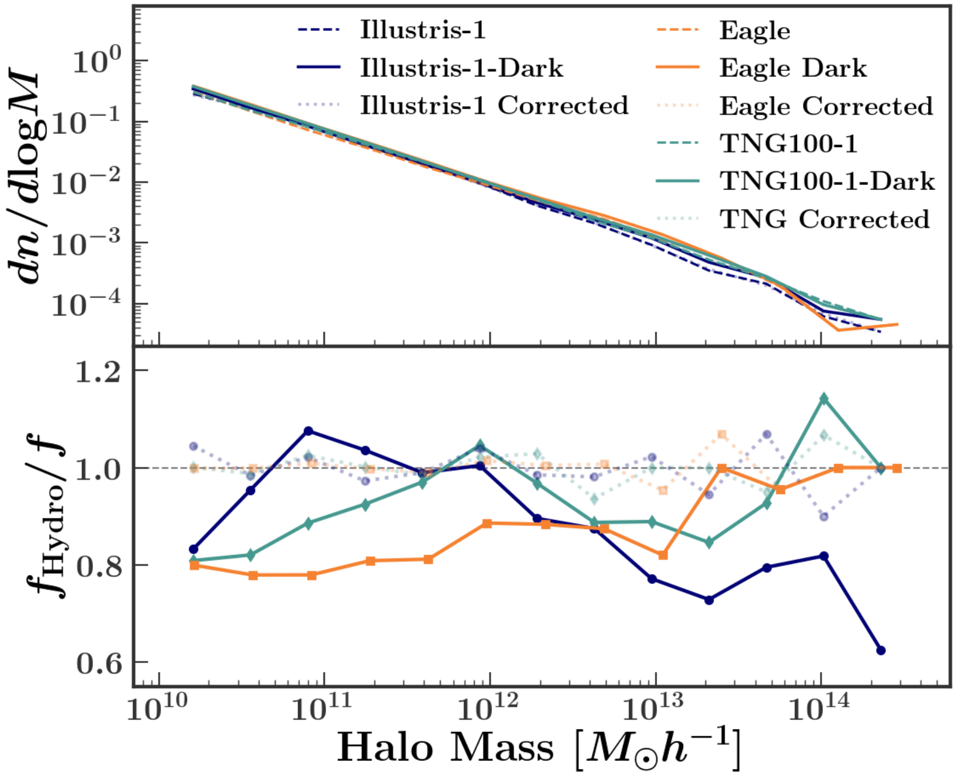

Figure 3.1: Halo mass functions of hydrodynamic compared to dark matter only simulations in the case of Illustris-1 (blue), TNG100-1 (green), and EAGLE RefL100N1504 (orange). The hydrodynamic versions are plotted with dashed lines, while the DMO versions are plotted with solid lines, and the corrected DMO versions are plotted with dotted lines. The bottom panel shows the ratio of the hydrodynamic HMF to either the DMO or the corrected DMO HMF for all three simulations. The halo definition shown is M_{200b}.

a smaller impact on high mass halos compared to Illustris. TNG has weaker AGN feedback than the original Illustris simulation, which leads to there being less discrepancy between the DMO and hydrodynamic HMFs at the high mass end. TNG also has a different wind model than Illustris, which leads to stronger feedback effects on lower mass halos, leading to more discrepancy at the lower mass end of the HMF in TNG.

Our HMF results for the EAGLE simulation are consistent with Desmond et al. (2017) and Schaller et al. (2015), who examined the differences between the halo masses in the EAGLE DMO and hydrodynamic runs, and found the halos to be less massive on average in the hydrodynamic run. Desmond et al. (2017) found that, at low halo masses, stellar feedback in EAGLE strips baryons from the halo, which reduces the growth rate of the halo. At higher halo masses, stellar feedback becomes less effective, but AGN feedback is still capable of expelling baryons in all but the most massive halos.

Another way to visualize the discrepancy between the halo mass distributions from hydrodynamic and DMO simulations is to plot the fractional difference between their masses. In Figures 3.2 and 3.3 we plot this fractional difference for M_{200b} (top) and M_{500c} (bottom) halos in Illustris, TNG100, and EAGLE at $z = 0$ and $z = 2$, respectively. Illustris halos are plotted in blue, TNG100 halos in green, and EALGE halos in orange. The x-axis is the DMO halo mass in units of $h^{-1}M_{\odot}$, and the y-axis is the fractional difference of hydrodynamic and DMO halo mass. In the left-hand column, we use the total hydrodynamic halo mass (i.e. dark matter and baryonic particles) to calculate this fractional difference. In the middle column, we use only the dark matter component of the hydrodynamic halo to calculate this fractional difference, and we normalize by one minus the universal baryonic mass fraction, $1 - \Omega_b/\Omega_m$, or $(\Omega_m - \Omega_b)/\Omega_m$. In the right-hand column, we use only the baryonic component (i.e. gas and star particles) of the hydrodynamic halo to calculate this fractional difference, and we normalize by the universal baryonic mass fraction Ω_b/Ω_m.

In each panel of Figures 3.2 and 3.3 hydrodynamic and DMO halos are paired based on their ranked masses, rather than spatial positions or particle IDs. Thus, the most massive DMO halo is paired with the most massive hydrodynamic halo, and so on. In other words, we essentially “abundance match” the halos in the hydrodynamic and DMO simulations (we do this separately for each redshift and halo definition, as well as each column in the figures; therefore, one cannot compare an individual point from one column or row to the next). We adopt this procedure because the fractional mass differences calculated in this way represent the correction one would need to apply to the halo masses from one of the DMO simulations in order to exactly recover the global halo mass function from the corresponding hydrodynamic simulation. While this abundance matching technique does not produce the exact same results as position or particle matching would, the overall trends seen are the same, although the scatter is drastically reduced by abundance matching

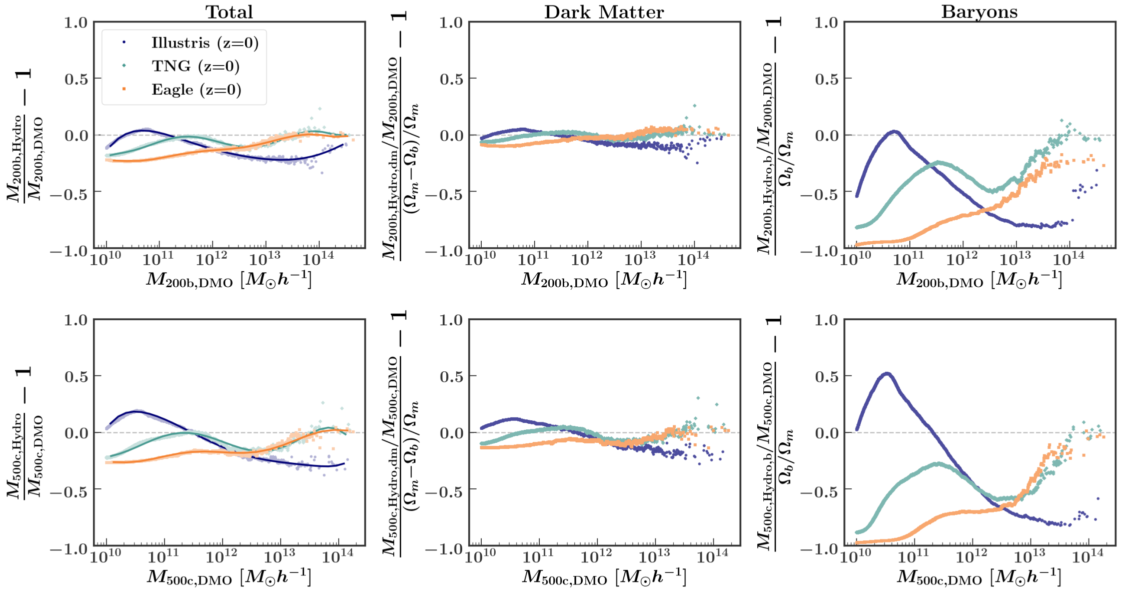

Figure 3.2: The fractional difference between halo masses from the hydrodynamic simulations to halo masses from the DMO simulations, as a function of DMO halo mass, all at $z = 0$. The hydrodynamic mass in the y-axis of each column is (from left to right): total mass, mass of dark matter particles, and mass of baryonic (gas and star) particles. In the latter two cases, masses are normalized to account for the global difference between the total, dark matter, and baryon matter densities. The top row displays M_{200b} halos, and the bottom row displays M_{500c} halos. Hydrodynamic and dark matter only halos are matched by their mass rank, rather than by position. The displayed ratio thus represents the correction factor needed to apply to the dark matter only halos in order to recover the hydrodynamic halo mass function. Polynomial fit correction functions, as described in Section 3.4, are plotted in the first column of panels as solid lines.

(Vogelsberger et al., 2014b; Schaller et al., 2015; Springel et al., 2018).

Examining the top left panel of Figure 3.2, we can see that the same trends are present here as were present in the halo mass functions shown in Figure 3.1. The results at $z = 0$ for M_{200b} halos Illustris (consistent with Vogelsberger et al. 2014b and Springel et al. 2018) indicate that stellar feedback slightly reduces the masses of the lowest mass halos (by up to $\sim 10\%$), while star formation efficiency slightly increases the masses of halos around $10^{11} h^{-1} M_\odot$, and AGN feedback severely reduces the masses of high mass halos (by up to $\sim 20\%$). The results for TNG100 (consistent with Springel et al. 2018) indicate that stellar feedback reduces the masses of low mass halos (by up to $\sim 15\%$), while star formation efficiency peaks at slightly higher masses than in Illustris (but is not quite so efficient as

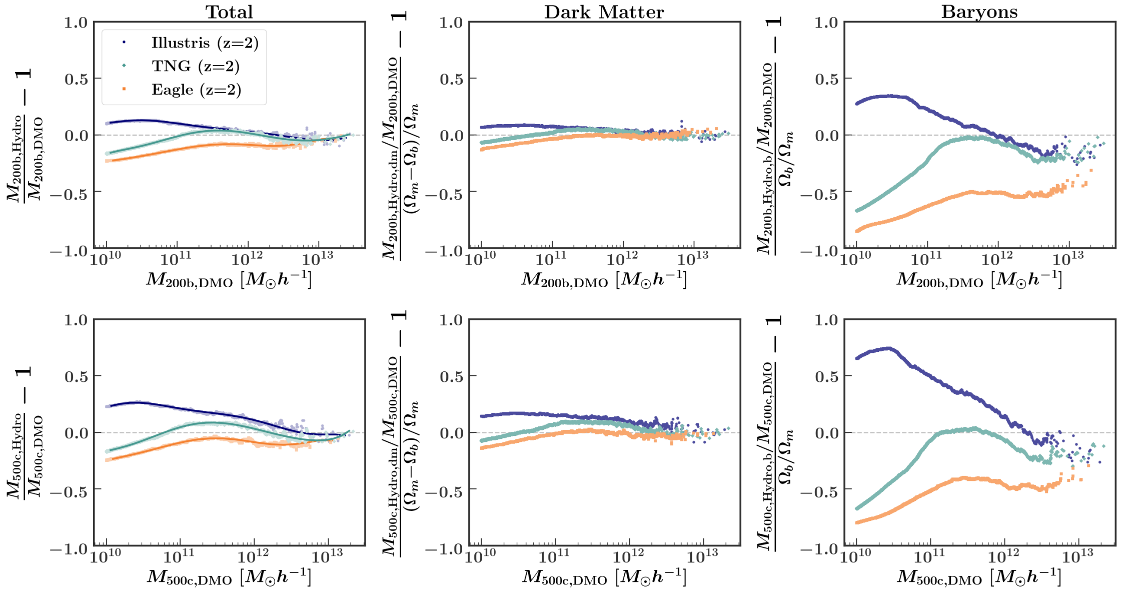

Figure 3.3: The fractional difference between halo masses from the hydrodynamic simulations to halo masses from the DMO simulations, as a function of DMO halo mass, all at $z = 2$. All features are the same as in Fig. 3.2.

to increase halo masses), while AGN feedback is less strong than in Illustris, and does not effect the highest mass halos, but reduces the masses of intermediate mass halos by up to $\sim 10\%$. The results for EAGLE (consistent with Schaller et al. 2015) indicate that stellar feedback severely reduces the masses of low mass halos (by up to $\sim 20\%$), while AGN feedback is similar to that in TNG, and reduces the masses of intermediate mass halos (by $\sim 10\%$) but does not impact the highest mass halos.

Looking at the second and third columns of Figure 3.2, we can see that feedback has a much more extreme effect on the baryons in a halo than it does on the dark matter particles. Thus, most of the mass difference is due to a loss or gain of baryons, and not dark matter. For example, in the EAGLE simulation, there is an extreme lack of baryons at lower halo masses, while in the Illustris simulation, there is a significant lack of baryons at higher halo masses, but for both simulations, the amount of dark matter in each halo remains relatively constant.

The differences between the three simulations in the third column of Figure 3.2 indicate that these three hydrodynamic simulations disagree in terms of the baryon content of their halos. Precise observational constraints on the baryon fraction as a function of halo mass could in principle allow us to differentiate between these three simulations. Gonzalez et al. (2013) measured the baryons contained in both the stellar and hot-gas components for galaxy clusters and groups with $M_{500} > 10^{14} M_{\odot}$ at $z = 0.1$, and found that the weighted mean baryon fraction for halos with $M_{500} > 2 \times 10^{14} M_{\odot}$ is 7% below the universal value when using a Planck cosmology. This is consistent with the results from the TNG and EAGLE simulations at the high mass end, but is not in agreement with the results from the Illustris simulation. The more significant differences between the three simulations, however, occur at lower halo masses. Eckert et al. (2017) examined the baryonic content of halos in the ECO and RESOLVE galaxy surveys (Moffett et al., 2015). While these results do extend to lower halo masses, they only include cold baryonic content, so we cannot make a direct comparison to the baryonic content in the three hydrodynmaic simulations. In the future, more precise observational constraints on the baryon content of low-mass halos could allow us to rule out certain hydrodynamic simulations.

Comparing the top and bottom rows of Figure 3.2, we can see that the discrepancy between hydrodynamic and DMO halo masses is more extreme for M_{500c} halos than it is for M_{200b} halos, indicating that feedback has stronger impact on the inner regions of a halo. This is most evident when looking at the baryonic component of Illustris M_{500c} halos (in the third column of the second row of Figure 3.2), where feedback and star formation have an extreme effect on the percentage of baryons in a halo.

We can also see that the trends present in each simulation at $z = 0$ are quite different from those at $z = 2$, which are shown in Figure 3.3. For example, we can see clearly from the M_{500c} Illustris halos that at $z = 2$, stellar and AGN feedback have not kicked in yet, and star formation efficiency is quite strong, resulting in a strong overabundance of gas and star particles. Because of this, essentially all M_{500c} Illustris halos at $z = 2$ have a higher mass in the hydrodynamic simulation than they do in the DMO simulation (by $\sim 10\%$). However, by $z = 0$, stellar feedback has kicked in for halos at or below $10^{10}h^{-1}M_{\odot}$, while AGN feedback has kicked in for halos above $5 \times 10^{11}h^{-1}M_{\odot}$, resulting in these halos having lower masses in the hydrodynamic simulation than their DMO counterparts. Meanwhile, in TNG100 and EAGLE, it appears that stellar feedback has already kicked in by $z = 2$, but AGN feedback has not, leading to less of a reduction in halo mass at the high mass end, but a similar result at low masses.

3.4 Correcting the DMO Halo Mass Function

Overall, Figures 3.2 and 3.3 emphasize the fact that the effect of baryonic physics on the halo mass function is to shift the HMF to lower masses. This shift can be as large as 25 percent. However, it is clear that this effect varies dramatically from one simulation to the next. This presents a problem for anyone using DMO halo catalogues to constrain the galaxy-halo connection or cosmology (Beltz-Mohrmann et al., 2020). The solution to this problem is not as simple as adding a parameter to the model, because it is a problem with the dark halo population itself. Additionally, the solution is not as simple as applying a constant offset to all DMO halo masses, because the discrepancy depends on the mass regime. Finally, once again, the solution depends on which hydrodynamic simulation is regarded as having the correct baryonic physics.

One possible solution is to apply a correction to each of the halos in a DMO simulation, so that the HMF better mimics the mass function from a hydrodynamic simulation. This correction should serve to adjust each DMO halo so that it has the mass of its corresponding hydrodynamic halo. We can use the fractional difference in halo mass shown in Figures 3.2 and 3.3 to identify a functional form for this correction.

To do this, we fit a polynomial to the fractional difference in halo masses between our hydrodynamic and DMO simulations (i.e., column one of Figures 3.2 and 3.3). The polynomial fits to each halo mass relationship were found using NumPy's polyfit function. After examining the effectiveness of several polynomials to correct the halo mass function, we found that a 7th order polynomial was the lowest order that could accurately capture the halo mass trend and allow us to correct the DMO halo mass function. We have additionally looked at the Bayesian Information Criteria (BIC) for polynomials ranging from 3rd order to 12th order and found that the mean BIC (averaged over each halo definition and redshift combination) continues to decrease until we reach 7th order, but does not continue to decrease significantly for higher orders. Thus, we decided that a 7th order polynomial was the lowest order we could use for our fits and still accurately correct the halo mass function.

The corrected DMO halo masses (in units of $10^{10}h^{-1}M_{\odot}$) are given by

$$M_{\mathrm{h,corrected}} = (y+1) \times M_{\mathrm{h,DMO}} \tag{3.1}$$

where $M_{\mathrm{h,DMO}}$ is the unlogged original halo mass in units of $10^{10}h^{-1}M_{\odot}$, and

$$y = ax^7 + bx^6 + cx^5 + dx^4 + ex^3 + fx^2 + gx + h, \tag{3.2}$$

where $x = \log_{10}(M_{\mathrm{h,DMO}})$ and a through h are the polynomial coefficients for a given simulation and halo definition. These fits (for M_{200b} and M_{500c} halos at redshifts 0 and 2) are plotted with solid lines in the first column panels of Figures 3.2 and 3.3. The fits at redshifts 0, 1, and 2 for all halo definitions are listed in Tables 3.2, 3.3, and 3.4.

Once again, this correction in based on "abundance matching" the halos between hydrodynamic and DMO simulations. Thus, applying this correction will assign to the most massive DMO halo the mass of the most massive hydrodynamic halo, and so on. After applying our corrections to each DMO simulation, we can examine our new corrected DMO halo mass functions. In Figure 3.1, we have plotted these corrected HMFs for M_{200b} halos at $z = 0$ with dotted lines. Looking at the residual panel, we can see that the corrections lead to significant improvement, and almost perfectly reproduce the HMFs from the hydrodynamic simulations (deviations are less than 5%), which was precisely the goal of our abundance matching method.

Unfortunately, the Illustris, TNG100, and EAGLE simulations do not contain any halos with masses above about $4 \times 10^{14} h^{-1} M_{\odot}$ at $z = 0$. This upper limit decreases for $z = 1$ and $z = 2$ halos, and is also somewhat dependent on halo definition. We do not make any assumptions about the masses of our halos beyond our data. Thus, our mass corrections should not be applied to any DMO halos above the limits given in Tables 3.2, 3.3, and 3.4. Rather, any DMO halo above our limits should be left unaltered. (This is very important; because our mass corrections are seventh order polynomials, extending them beyond our mass limits will lead to very large changes in halo mass.)

It is noteworthy that in TNG100 and EAGLE, the mass correction is already almost zero at the high mass end. However, in Illustris, for $z = 0$ halos, this is not the case. This means that applying the Illustris halo mass correction will lead to a slight discontinuity in the halo mass function. This discontinuity could be alleviated by, for example, extrapolating the fit at the high mass end; however, because doing this would not be based on any data, we do not provide any extrapolations of our fits here. Additionally, our mass corrections should only be applied to halos with masses above $10^{10} h^{-1} M_{\odot}$. We do not present corrections for halos below $10^{10} h^{-1} M_{\odot}$ in this work due to the mass resolutions of the simulations that we use. (Once again, it is very important not to apply the corrections to any DMO halos below $10^{10} h^{-1} M_{\odot}$ due to the nature of the seventh order polynomial fits.)

We recommend that any future work using halos from a DMO simulation do the following: for a given halo definition and redshift, apply at least one of our halo mass corrections to correct the halo mass function. Once the mass corrections are applied, the remaining analysis (e.g. HOD, CLF, etc.), can be performed on the corrected halo catalogues. In this way, one can investigate the robustness of their results to changes in the halo mass function. Ideally, we recommend applying all three corrections (i.e. each correction based on Illustris, TNG100, and EAGLE) since the variation among them represents the theoretical uncertainty in baryonic physics.

We have created a PYTHON module for implementing our halo mass corrections, which

Table 3.2: Shown here are the DMO halo mass corrections for $z = 0$ halos. The columns list (from left to right): halo definition, simulation, upper mass limit, and the seventh order polynomial coefficients of the halo mass correction fits, beginning with the highest order.

Halo Def.	Simulation	Upper Mass Lim.	a	b	c	d	e	f	g	h
M_{200b}	Illustris	$3.2 \cdot 10^{14}$	-0.000215	0.024375	-0.004406	-0.05112	0.32264	-0.75977	0.65421	-0.14617
	TNG	$3.6 \cdot 10^{14}$	0.0035204	-0.05293	0.30055	-0.79135	0.94957	-0.45741	0.17801	-0.19176
	EAGLE	$4.3 \cdot 10^{14}$	0.0019695	-0.02991	0.17485	-0.49145	0.66861	-0.36278	0.06311	-0.2312
M_{FoF}	Illustris	$3.5 \cdot 10^{14}$	-0.0007237	0.010774	-0.05997	0.13949	-0.03321	-0.3973	0.47075	-0.10679
	TNG	$4.1 \cdot 10^{14}$	0.002596	-0.038396	0.21334	-0.54396	0.61565	-0.26124	0.11294	-0.18982
	EAGLE	$4.3 \cdot 10^{14}$	0.001614	-0.024094	0.13731	-0.37018	0.46557	-0.2016	0.007495	-0.20007
M_{vir}	Illustris	$2.8 \cdot 10^{14}$	-0.0003252	0.003505	-0.006643	-0.059713	0.3689	-0.083365	0.67904	-0.12001
	TNG	$3.2 \cdot 10^{14}$	0.003404	-0.05052	0.28087	-0.71152	0.78559	-0.30528	0.13288	-0.19697
	EAGLE	$3.4 \cdot 10^{14}$	0.0021115	-0.031718	0.18264	-0.50231	0.66224	-0.34231	0.057532	-0.23903
M_{200c}	Illustris	$2.0 \cdot 10^{14}$	-0.0015556	0.019282	-0.084937	0.12921	0.14682	-0.73173	0.65319	-0.069491
	TNG	$2.4 \cdot 10^{14}$	0.003539	-0.050576	0.26672	-0.61706	0.54882	-0.073071	0.075193	-0.20515
	EAGLE	$2.5 \cdot 10^{14}$	0.0024825	-0.036885	0.20893	-0.55934	0.70466	-0.33643	0.051158	-0.24832
M_{500c}	Illustris	$1.4 \cdot 10^{14}$	0.0016927	-0.024244	0.14307	-0.46591	0.96494	-1.3004	0.76757	0.029836
	TNG	$1.5 \cdot 10^{14}$	0.0017185	-0.026947	0.14529	-0.30347	0.13573	0.14566	0.08892	-0.21709
	EAGLE	$1.7 \cdot 10^{14}$	0.002766	-0.040727	0.22567	-0.57591	0.65174	-0.22731	0.01389	-0.26156

Table 3.3: Shown here are the DMO halo mass corrections for $z = 1$ halos. The columns list (from left to right): halo definition, simulation, upper mass limit, and the seventh order polynomial coefficients of the halo mass correction fits, beginning with the highest order.

Halo Def.	Simulation	Upper Mass Lim.	a	b	c	d	e	f	g	h
M_{200b}	Illustris	$7.8 \cdot 10^{13}$	-0.0091394	0.12078	-0.62164	1.55561	-1.88022	0.82614	0.051559	0.029147
	TNG	$8.7 \cdot 10^{13}$	0.0004585	-0.012752	0.095528	-0.27788	0.28703	-0.036331	0.095493	-0.18408
	EAGLE	$9.1 \cdot 10^{13}$	0.0038916	-0.055802	0.3104	-0.83425	1.09195	-0.61358	0.18135	-0.25625
M_{FoF}	Illustris	$9.9 \cdot 10^{13}$	-0.004083	0.055165	-0.28932	0.72959	-0.84872	0.25271	0.15202	-0.001737
	TNG	$1.1 \cdot 10^{14}$	0.0006069	-0.010626	0.065737	-0.17125	0.15705	-0.0026786	0.057664	-0.18654
	EAGLE	$9.7 \cdot 10^{13}$	0.001926	-0.028518	0.16368	-0.45252	0.60351	-0.33582	0.10452	-0.20726
M_{vir}	Illustris	$7.7 \cdot 10^{13}$	-0.009481	0.12481	-0.63996	1.59564	-1.92256	0.84396	0.05024	0.031429
	TNG	$8.7 \cdot 10^{13}$	0.0005164	-0.01353	0.099543	-0.28774	0.29864	-0.042489	0.09722	-0.18443
	EAGLE	$9.1 \cdot 10^{13}$	0.004431	-0.063071	0.34911	-0.93817	1.24065	-0.72282	0.2166	-0.25995
M_{200c}	Illustris	$7.5 \cdot 10^{13}$	-0.010986	0.1405	-0.69674	1.6675	-1.89517	0.71634	0.12019	0.046023
	TNG	$8.4 \cdot 10^{13}$	0.0007219	-0.015852	0.10836	-0.29737	0.28648	-0.019382	0.09872	-0.18967
	EAGLE	$8.8 \cdot 10^{13}$	0.004001	-0.057765	0.32258	-0.866998	1.12902	-0.62778	0.18489	-0.26094
M_{500c}	Illustris	$6.0 \cdot 10^{13}$	-0.007073	0.090233	-0.42661	0.89096	-0.65131	-0.31909	0.43537	0.14683
	TNG	$6.8 \cdot 10^{13}$	-0.0008956	0.001151	0.038427	-0.14363	0.079659	0.11429	0.11684	-0.20366
	EAGLE	$7.1 \cdot 10^{13}$	0.003834	-0.057735	0.32624	-0.85209	1.00627	-0.43779	0.12879	-0.27083

Table 3.4: Shown here are the DMO halo mass corrections for $z = 2$ halos. The columns list (from left to right): halo definition, simulation, upper mass limit, and the seventh order polynomial coefficients of the halo mass correction fits, beginning with the highest order.

Halo Def.	Simulation	Upper Mass Lim.	a	b	c	d	e	f	g	h
M_{200b}	Illustris	$2.7 \cdot 10^{13}$	-0.003523	0.0452	-0.22138	0.51417	-0.54733	0.14959	0.060834	0.10298
	TNG	$3.0 \cdot 10^{13}$	0.002992	-0.043248	0.24032	-0.62563	0.74363	-0.36722	0.20431	-0.16932
	EAGLE	$2.0 \cdot 10^{13}$	-0.004707	0.040814	-0.12442	0.16518	-0.12594	0.088025	0.064783	-0.23354
M_{FoF}	Illustris	$2.8 \cdot 10^{13}$	-0.00888	0.092242	-0.37524	0.75331	-0.74998	0.27535	0.016797	0.048725
	TNG	$3.3 \cdot 10^{13}$	0.001705	-0.02408	0.13514	-0.36934	0.49252	-0.31233	0.16692	-0.18532
	EAGLE	$2.8 \cdot 10^{13}$	-0.003794	0.038169	-0.14905	0.29626	-0.34349	0.237997	-0.02068	-0.17421
M_{vir}	Illustris	$2.7 \cdot 10^{13}$	-0.003382	0.043078	-0.21112	0.49535	-0.54109	0.16667	0.047084	0.098199
	TNG	$3.0 \cdot 10^{13}$	0.002231	-0.034183	0.19859	-0.5329	0.64231	-0.31672	0.19247	-0.16811
	EAGLE	$2.1 \cdot 10^{13}$	-0.00395	0.03334	-0.096012	0.11237	-0.075638	0.066711	0.064067	-0.23171
M_{200c}	Illustris	$2.6 \cdot 10^{13}$	-0.0034895	0.04379	-0.21029	0.47787	-0.48856	0.10121	0.076317	0.11155
	TNG	$2.9 \cdot 10^{13}$	0.0034024	-0.047361	0.25559	-0.65001	0.75527	-0.36117	0.20235	-0.16962
	EAGLE	$1.9 \cdot 10^{13}$	-0.0066515	0.059723	-0.19541	0.29533	-0.2477	0.14119	0.059818	-0.23506
M_{500c}	Illustris	$1.9 \cdot 10^{13}$	-0.12841	0.13531	-0.53611	0.96895	-0.71259	-0.033619	0.17877	0.222999
	TNG	$2.1 \cdot 10^{13}$	0.006793	-0.077694	0.34484	-0.70242	0.56511	-0.083469	0.1591	-0.16571
	EAGLE	$1.3 \cdot 10^{13}$	-0.003905	0.019662	0.012513	-0.1937	0.28351	-0.12291	0.16284	-0.24759

is publicly available at https://github.com/gbeltzmo/halo_mass_correction. This module takes in an array of halo masses, a halo definition (M_{200b}, $M_{\rm FoF}$, $M_{\rm vir}$, M_{200c}, or M_{500c}), a redshift (0, 1, or 2), and a simulation (Illustris, TNG, or EAGLE), and returns the corresponding corrected halo masses. If a given halo mass is outside the accepted mass range, the code will issue a warning, and will return the original (uncorrected) halo mass.

One question worth investigating is whether applying our mass corrections to a box with a very different resolution produces accurate results. To investigate this, we applied the TNG correction (based on the TNG100-1 simulation) to the TNG300-1-Dark box (which is about a factor of 8 lower in resolution than the TNG100-1-Dark simulation) to see if we could reproduce the halo mass function from the TNG300-1 simulation. Looking only at the $10^{10} - 10^{14} h^{-1} M_{\odot}$ mass range, this correction almost perfectly reproduces the halo mass function from TNG300-1, with all deviations less than 5% (compared to 17% deviations without the correction). This indicates that as long as the mass corrections are only applied within the appropriate mass regime, the corrections are accurate even when used with simulations of different volumes and resolutions.

Szewciw et al. (in prep) have applied our mass corrections to $M_{\rm vir}$ halos (z=0) from a large DMO simulation (Las Damas; McBride et al., 2009) and examined how their HOD parameter constraints on SDSS galaxies varied with the different halo mass corrections. They found that for both their luminosity samples (M_r^{-19} and M_r^{-21}), the different halo mass corrections lead to changes in all their HOD parameters. The biggest changes are seen in log$M_{\rm min}$ and logM_1, and in most cases the Illustris-based mass correction leads to the biggest change, although for the M_r^{-19} sample it is the EAGLE-based correction that leads to the biggest difference in log$M_{\rm min}$. While it is to be expected that the halo mass corrections lead to changes in the mass parameters of the HOD, these changes are simulation dependent, and are not trivial to predict. The halo mass corrections ultimately do not lead to better χ^2 values for the best fitting models, nor do they lead to the models being ruled out.

3.5 Environmental Dependence

Our “abundance matching” technique does not explicitly take halo environment into account when correcting the halo masses. This means that while our mass corrections successfully reproduce the global halo mass function from our hydrodynamic simulations, they will not correct the *conditional* HMF if baryonic effects on halo mass are environment dependent.

Fitting halo mass corrections in which halos are matched between DMO and hydro based on position or particle IDs would inherently take halo environment into account. However, there are several issues with this technique. Firstly, matching halos based on position or particle IDs does not guarantee that every halo in the hydrodynamic simulation has a match in the DMO simulation. Secondly, this method of matching introduces a significant amount of scatter into the hydrodynamic-to-DMO halo mass relationship. Therefore, if one were to use this relationship to correct halos from a large DMO simulation, one would have to either ignore the scatter (in which case the result is essentially the same as the abundance matching correction), or take the scatter into account by binning halos by mass and then drawing their corrected mass from a distribution within that bin. While this accounts for the scatter, it still does not account for halo environment. Additionally, this is not the cleanest method for correcting DMO halo masses; when applied to any of the three simulations used in this work, this method does not successfully reproduce the correct global *or* conditional HMF from the hydrodynamic simulations.

One alternate possibility is to use our original “abundance matching” technique, but to separate halos by environment. We can do this by measuring the large-scale environment around each of our DMO and hydrodynamic halos. We can then split our halos into those with high-density environments and those with low-density environments, and subsequently “abundance match” between DMO and hydro, matching *only* halos within *similar* environments. For example, the most massive DMO halo in a *high*-density environment would be matched with the most massive hydrodynamic halo in a *high*-density environment, while the most massive DMO halo in a *low*-density environment would be matched with the most massive hydrodynamic halo in a *low*-density environment, and so on. This procedure will yield mass corrections that are guaranteed to recover the correct conditional HMF.

To do this for each of our simulations, we first have to measure the large-scale environment around each of our halos. To measure halo environment, we calculate the total mass of *halos* within 5 Mpc spheres centered on each halo of interest (excluding the mass of the halo of interest itself). In other words, we do not sum up all particles, but rather sum up the masses of all halos in the halo catalog whose centers fall within the sphere (excluding the main halo). We do not impose any lower mass limit on the halos included in this sum. We measure environments for all DMO halos above $10^{10}h^{-1}M_{\odot}$, and all hydrodynamic halos that are matched to them using the abundance matching method described above.

We can thus define an environment factor δ for each halo, such that $\delta = (\rho_{\rm sphere}/\rho_{\rm box}) - 1$, where $\rho_{\rm sphere}$ is the mass of halos in a 5 Mpc sphere around the halo divided by the volume of a 5 Mpc sphere, and ρ_{box} is the sum of all halo masses in the box divided by the volume of the box. We measure halo environment in this way for all three of our simulations - hydrodynamic and DMO alike. We repeat this measurement for each of our different halo definitions and redshifts. We then split our halos into "high-" and "low-density" environments based on the median environment ($\delta_{\rm med}$) for that simulation. This is done separately for each halo definition and redshift, and is also done separately for the hydrodynamic and DMO simulations. Thus, each simulation/halo definition/redshift combination has a slightly different $\delta_{\rm med}$. Subsequently, DMO halos in high-density environments are "abundance matched" with hydrodynamic halos also in high-density environments, and likewise for halos in low-density environments.

Shown in Figure 3.4 are the results of this environment-dependent abundance matching technique for M_{200b} halos at $z = 0$ in Illustris-1 (blue), TNG100-1 (green), and EAGLE (orange). Halos in high-density environments are plotted in darker colors, and halos in

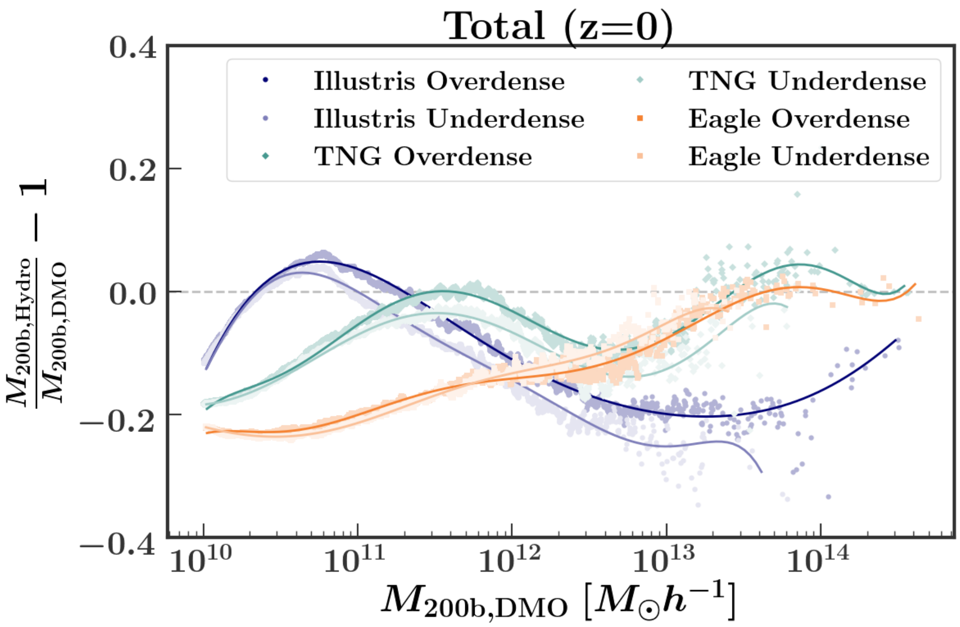

Figure 3.4: Fractional difference between hydrodynamic and DMO M_{200b} halos from Illustris-1 (blue), TNG100-1 (green), and EAGLE (orange) at $z = 0$. In this plot, hydrodynamic and DMO halos are first split into high and low density environments, and then are abundance matched with corresponding halos in similar environments. High density environments are plotted in darker colors, and low density environments are plotted in lighter colors. Each trend is fit with a seventh order polynomial, which is given in Table 3.5.

low-density environments are plotted in lighter colors. Each of these relationships is once again fit with a seventh order polynomial, which is plotted here with a solid line. The fits for M_{200b} and $M_{\rm vir}$ halos at $z = 0$, along with the $\delta_{\rm med}$ and mass limit, are given in Table 3.5.

We can see from Figure 3.4 that for each simulation, the relationship between hydrodynamic and DMO halos closely resembles that seen in Figure 3.2 for M_{200b} halos, but we do detect a difference between the high- and low-density environments. In Illustris and TNG100, the differences between environments appear for all DMO halos above $5\times10^{10}h^{-1}M_\odot$, where hydrodynamic halos in high-density environments are slightly more massive relative to DMO halos than those in low-density environments. In EAGLE, below $2\times10^{10}h^{-1}M_\odot$ the high- and low-density trends are the same; between 2×10^{10} and $5\times10^{11}h^{-1}M_\odot$ the hydrodynamic halos in high-density environments are slightly more massive relative to DMO halos than those in low-density environments; at higher masses this relationship is reversed. Additionally, for all three simulations, the highest mass hydrodynamic and DMO halos are exclusively found in high-density environments, as expected.

While our original "abundance matching" halo mass correction reproduces the global halo mass function from a given hydrodynamic simulation, our environment-dependent halo mass correction reproduces both the global *and* the conditional halo mass function. To examine whether it is important to implement this more complicated halo mass correction, we can see how well our different mass corrections reproduce the halo clustering found in hydrodynamic simulations by measuring the halo correlation function, $\xi(r)$.

In Figure 3.5 we plot the halo correlation functions for M_{200b} halos at $z = 0$ in TNG300-1. We use the TNG300-1 box for this analysis because the smaller hydrodynamic boxes contain too few halos to accurately examine the clustering of halos. We calculate the halo correlation function in two different mass bins: halos greater than $10^{11}h^{-1}M_{\odot}$ and halos greater than $10^{12}h^{-1}M_{\odot}$. For the lower mass sample, we measure ξ in 13 bins of separation r between 0.49 and $15h^{-1}$Mpc, and for the higher mass sample we measure ξ in 10 bins of separation r between 1.07 and $15h^{-1}$Mpc. We compute ξ with the blazing fast code CORRFUNC (Sinha and Garrison, 2017, 2019). We measure ξ on halos from the following versions of TNG300: (i) hydrodynamic simulation, (ii) the DMO simulation, (iii) the DMO simulation with the global "abundance matching" halo mass correction, (iv) and the DMO simulation with the environment-dependent halo mass correction. In this figure, the corrections are done on a halo-by-halo basis; in other words, we do not use our polynomial fits to make these corrections, but rather we directly assign DMO halos the exact masses of the corresponding "abundance matched" hydrodynamic halos. In this way, we can assess the ability of a given mass correction to reproduce the bias seen in the full-physics version of TNG300 without introducing any error due to our fits. In the bottom panel of Figure 3.5 we plot the ratio of each subsequent ξ (DMO, DMO with "abundance matching" correction, or DMO with environment dependent correction) to the hydrodynamic ξ.

For both mass samples, the DMO halos do not exhibit the same correlation function as do the hydrodynamic halos; the residual panel shows that the difference in ξ for the low (high) mass sample is about 8% (7%) at the smallest scales, and for both mass samples it is about 4% for all scales above $2h^{-1}$Mpc. After applying the global "abundance matching" halo mass correction (dot-dashed lines), the difference in ξ reduces to about 2% for both mass samples for all scales above $2h^{-1}$Mpc (the difference is slightly larger for the high mass sample). This difference reaches about 6% on the smallest scales for both mass samples. We emphasize that this lingering discrepancy in ξ after applying the abundance matching mass correction is not due to the accuracy of our seventh-order polynomial fit, because in this figure we are not using our fits, but rather matching individual halos. In other words, at this point the global halo mass function has been perfectly corrected, so the remaining differences in ξ must be due to other factors.

After applying the environment-dependent halo mass correction (solid lines), the discrepancies in ξ shrink once again. For both mass samples, the discrepancy between the environment-dependent corrected DMO halo correlation function and the hydrodynamic halo correlation function is less than 1% for all scales above $2h^{-1}$Mpc. At smaller scales, the environment-dependent correction is still an improvement over the original correction or no correction, although the discrepancy still reaches about 5 or 6% at the smallest scales we consider. This is due to the fact that halo exclusion occurs at smaller scales in the hydrodynamic simulation than it does in the DMO simulation. Because our halo mass corrections have no impact on the sizes or positions of halos, they cannot improve the halo correlation function until it reaches a scale where both the DMO and hydrodynamic simulations are not lacking halo pairs due to halo exclusion.

These results indicate that the impact of baryonic physics on the halo mass function is dependent on the environment of the halo, which in turn affects the ability of a DMO simulation to reproduce the halo clustering observed in hydrodynamic simulations. Our original "abundance matching" halo mass corrections are able to reproduce the global halo mass function seen in hydrodynamic simulations, and are able to reproduce the halo clus-

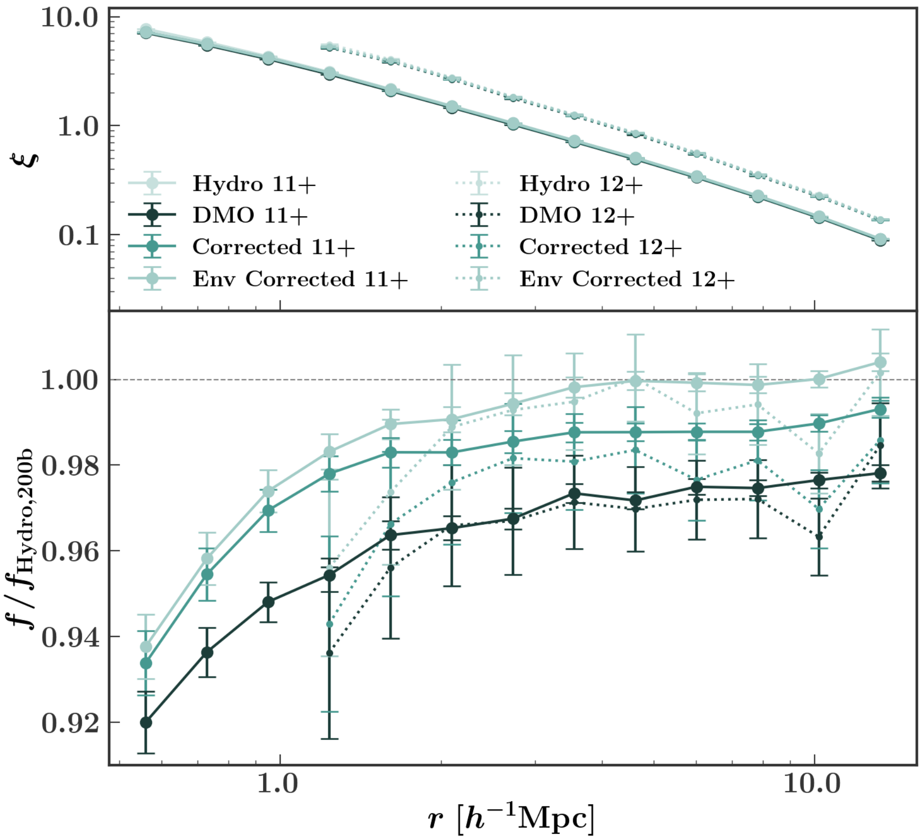

Figure 3.5: The halo correlation functions for M_{200b} halos at $z = 0$ in TNG300-1 (light green), TNG300-1-Dark (dark green), TNG300-1-Dark with the abundance matching correction (green), and TNG300-1-Dark with the environment-dependent abundance matching correction (light green). Halo correlation functions are plotted for halos greater than $10^{11}h^{-1}M_\odot$ (solid), and greater than $10^{12}h^{-1}M_\odot$ (dotted). The bottom panel shows the ratios of each DMO halo correlation function for a given mass bin.

tering to within a few percent. However, if higher accuracy is needed, it is important to account for environment-dependent baryonic effects. Our environment-dependent halo mass corrections reproduce both the global and the conditional halo mass function from hydrodynamic simulations, and they reproduce the halo clustering to within less than 1%.

In Tables 3.5, 3.6, and 3.7 we provide our environment-dependent halo mass corrections for all halo definitions in Illustris, TNG, and EAGLE at $z = 0$, $z = 1$, and $z = 2$, respectively. Additionally, corrections are also available in our PYTHON module (https://github.com/gbeltzmo/halo_mass_correction).

Figure 3.6 shows the results of applying these environment-dependent halo mass cor-

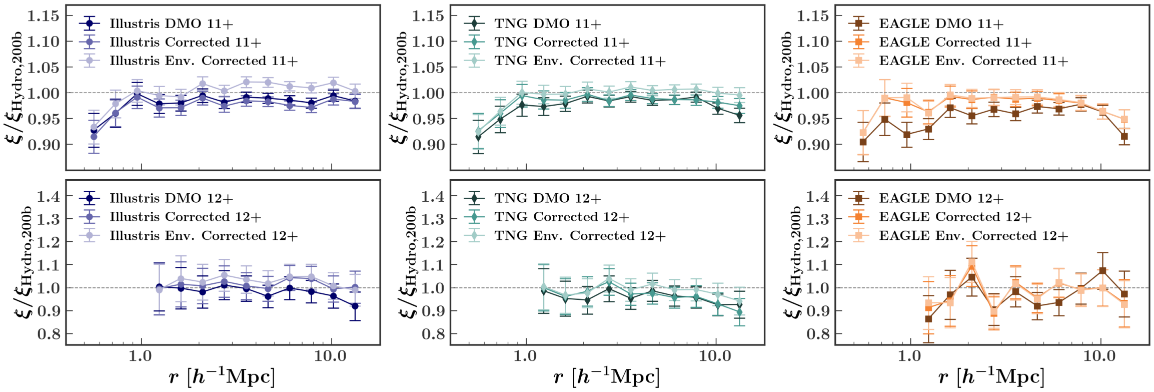

Figure 3.6: The discrepancy in halo correlation functions for M_{200b} halos at $z = 0$ in Illustris (left), TNG (middle), and EAGLE (right). In each panel, we show the discrepancy between the hydrodynamic halo correlation function and that in DMO (dark), DMO with the original halo mass correction (medium), and DMO with the environment dependent halo mass correction (light). The top panels are the correlation function discrepancies for halos greater than $10^{11}h^{-1}M_{\odot}$, and the bottom panels are for halos greater than $10^{12}h^{-1}M_{\odot}$.

rections to Illustris (left), TNG (middle), and EAGLE (right). Each panel shows the discrepancy between the hydrodynamic correlation function and that in DMO (dark), DMO with the original halo mass correction (medium), and DMO with the environment dependent halo mass correction (light). In this figure, the mass corrections used are the fits given in Tables 3.2 and 3.5 for M_{200b} halos at $z = 0$. The top panels show the correlation function discrepancies for halos above than $10^{11}M_{\odot}$, and the bottom panels show the same for halos above $10^{12}M_{\odot}$. (We do not show halos below $10^{11}M_{\odot}$ because in all three simulations the environment dependent corrections do not deviate from the original corrections below $10^{11}M_{\odot}$. Additionally, the clustering of low mass halos is more complicated, and its investigation is beyond the scope of this paper.)

In each simulation, the DMO halos exhibit some clustering discrepancy compared to the hydrodynamic halos. For the lower mass sample (top panels), DMO underestimates the clustering by about 3% on average in Illustris, and TNG, and by about 5% on average in EAGLE. In Illustris, the original mass correction actually makes the clustering slightly worse, while the environment-dependent correction is an improvement on small-scales, and a slight overcorrection on large scales. In TNG, the original mass correction provides little to no improvement over DMO, but the environment-dependent correction improves the clustering almost completely. In EAGLE, the original and the environment-dependent corrections are in very close agreement, and both improve the clustering almost entirely compared to DMO.

For the higher mass sample (bottom panels), the error bars are large due to a lack of high mass halos. However, in each case, the DMO simulations underestimate the clustering of halos compared to the hydrodynamic simulations. This discrepancy is smallest in Illustris, and as a result, both the original and the environment-dependent mass corrections yield a slight overestimation of clustering. In TNG, the original mass correction provides little to no improvement, while the environment-dependent correction once again improves the clustering almost completely. In EAGLE, both the original and the environment-dependent corrections provide a slight improvement compared to DMO.

Applying these corrections to a large DMO simulation is slightly more complicated, because it requires calculating the large-scale environment for each halo of interest (only those above $10^{10}h^{-1}M_{\odot}$), and then separating these halos into "high" and "low" density environments based on the median environment $\delta_{\rm med}$, and applying the corresponding mass correction. (In our module, an environment argument must be passed, which can be "all," "high," or "low," wherein the code assumes it is applying the correction to all halos, only halos in high density environments, or only halos in low density environments, respectively.)

We have provided the $\delta_{\rm med}$ that we found for each simulation/halo definition/redshift in Tables 3.5, 3.6, and 3.7. If applying our environment dependent corrections, we suggest using the $\delta_{\rm med}$ that we provide for a given correction, rather than finding a new $\delta_{\rm med}$ that is specific to your sample. This is because our corrections are based on splitting halos into high- and low-density environments using a $\delta_{\rm med}$ for all halos above $10^{10}h^{-1}M_{\odot}$. However, $\delta_{\rm med}$ does increase with halo mass. Thus, for a halo catalogue that only includes halos with masses above $10^{12}h^{-1}M_{\odot}$, for example, the pertinent $\delta_{\rm med}$ would be higher. This

Table 3.5: Environment dependent fits for $z = 0$ halos in Illustris (I), TNG (T), and EAGLE (E). The first four columns are the halo definition, simulation, the median environment factor δ, and the type of environment. The fifth column is the upper mass limit for each fit. (The lower limit is $10^{10}h^{-1}M_\odot$ for all fits.) The remaining columns are the polynomial coefficients for the fits.

Halo	Sim.	δ_{med}	Env.	Mass Limit	a	b	c	d	e	f	g	h
M_{200b}	I	0.056	High	$3.2 \cdot 10^{14}$	-0.000282	0.00451	-0.025829	0.050903	0.082441	-0.49837	0.55978	-0.13646
			Low	$6.0 \cdot 10^{13}$	-0.009371	0.11086	-0.50638	1.0971	-1.0281	0.016647	0.45108	-0.13212
	T	0.03	High	$3.6 \cdot 10^{14}$	0.004054	-0.06133	0.3523	-0.94844	1.19187	-0.63492	0.23833	-0.19604
			Low	$7.1 \cdot 10^{13}$	-0.000712	0.002265	0.018051	-0.074849	0.015843	0.1256	0.027276	-0.1839
	E	0.064	High	$4.3 \cdot 10^{14}$	0.001995	-0.030087	0.17401	-0.48138	0.64179	-0.34464	0.069041	-0.23138
			Low	$2.9 \cdot 10^{13}$	0.000006	-0.005095	0.050413	-0.17942	0.26066	-0.089445	-0.027582	-0.2228
$M_{\rm FoF}$	I	0.037	High	$3.5 \cdot 10^{14}$	-0.001322	0.020666	-0.1249	0.35426	-0.40672	-0.082729	0.38557	-0.10176
			Low	$5.4 \cdot 10^{13}$	-0.003437	0.043915	-0.21498	0.48554	-0.40377	-0.23471	0.44261	-0.110298
	T	0.006	High	$4.1 \cdot 10^{14}$	0.003159	-0.046315	0.25607	-0.65495	0.75757	-0.3424	0.13639	-0.19178
			Low	$6.7 \cdot 10^{13}$	-0.002906	0.033108	-0.14702	0.345	-0.49753	0.39455	-0.038197	-0.1823
	E	0.059	High	$4.3 \cdot 10^{14}$	0.001609	-0.023019	0.12344	-0.3034	0.32554	-0.087478	-0.014541	-0.19728
			Low	$2.7 \cdot 10^{13}$	0.000044	-0.005998	0.058549	-0.2139	0.3297	-0.15123	-0.011213	-0.19971
$M_{\rm vir}$	I	0.068	High	$2.8 \cdot 10^{14}$	0.000327	-0.006513	0.053503	-0.23813	0.64263	-1.05189	0.7843	-0.13659
			Low	$4.9 \cdot 10^{13}$	-0.008	0.090915	-0.39228	0.76673	-0.50891	-0.42139	0.60965	-0.12534
	T	0.039	High	$3.2 \cdot 10^{14}$	0.003833	-0.057633	0.32688	-0.8577	1.021834	-0.49292	0.20724	-0.20621
			Low	$6.4 \cdot 10^{13}$	-0.009157	0.10013	-0.42418	0.91773	-1.1416	0.77501	-0.098584	-0.18654
	E	0.073	High	$3.4 \cdot 10^{14}$	0.002174	-0.032846	0.19068	-0.53199	0.72558	-0.4227	0.10743	-0.24488
			Low	$2.5 \cdot 10^{13}$	0.001936	-0.029656	0.17358	-0.4863	0.65787	-0.35221	0.059215	-0.24129
M_{200c}	I	0.098	High	$2.0 \cdot 10^{14}$	0.001173	-0.019618	0.13432	-0.49339	1.08891	-1.4831	0.95941	-0.10775
			Low	$3.5 \cdot 10^{13}$	-0.007623	0.081204	-0.31079	0.45385	0.097309	-1.00228	0.82077	-0.10244
	T	0.063	High	$2.4 \cdot 10^{14}$	0.004882	-0.071596	0.39701	-1.02283	1.21394	-0.62421	0.28046	-0.22167
			Low	$5.6 \cdot 10^{13}$	-0.016976	0.18901	-0.81739	1.77293	-2.0835	1.2395	-0.14973	-0.19893
	E	0.088	High	$2.5 \cdot 10^{14}$	0.002764	-0.04149	0.23974	-0.66758	0.91638	-0.55755	0.15461	-0.26261
			Low	$2.1 \cdot 10^{13}$	-0.000626	-0.005441	0.091492	-0.36922	0.60402	-0.37903	0.094691	-0.25489
M_{500c}	I	0.146	High	$1.4 \cdot 10^{14}$	0.004316	-0.064464	0.38943	-1.23356	2.25144	-2.44401	1.27652	-0.037299
			Low	$2.6 \cdot 10^{13}$	0.008843	-0.11192	0.58452	-1.63652	2.69776	-2.68389	1.27195	-0.048703
	T	0.113	High	$1.5 \cdot 10^{14}$	0.004952	-0.073248	0.4092	-1.064929	1.30501	-0.77502	0.4149	-0.24536
			Low	$4.3 \cdot 10^{13}$	-0.013201	0.12858	-0.45147	0.69274	-0.42261	-0.085257	0.34902	-0.24587
	E	0.135	High	$1.7 \cdot 10^{14}$	0.006497	-0.092612	0.510003	-1.35855	1.79593	-1.090549	0.30516	-0.28725
			Low	$1.6 \cdot 10^{13}$	-0.00035	-0.008961	0.1102	-0.41233	0.62834	-0.34813	0.080577	-0.27024

Table 3.6: Environment dependent fits for $z = 1$ halos in Illustris (I), TNG (T), and EAGLE (E). The first four columns are the halo definition, simulation, the median environment factor δ, and the type of environment. The fifth column is the upper mass limit for each fit. (The lower limit is $10^{10}h^{-1}M_{\odot}$ for all fits.) The remaining columns are the polynomial coefficients for the fits.

Halo	Sim.	δ_{med}	Env.	Mass Limit	a	b	c	d	e	f	g	h
M_{200b}	I	0.10599	High	$7.8 \cdot 10^{13}$	-0.0074725	0.100308	-0.523482	1.324619	-1.604935	0.664136	0.099003	0.03675
			Low	$1.2 \cdot 10^{13}$	-0.010227	0.141074	-0.734014	1.834347	-2.225889	1.050426	-0.0303705	0.027386
	T	0.08711	High	$8.7 \cdot 10^{13}$	-0.000475	-0.000194	0.029589	-0.10719	0.061838	0.101182	0.06785	-0.181415
			Low	$1.4 \cdot 10^{13}$	0.007072	-0.069453	0.2886	-0.6129	0.600835	-0.18297	0.11624	-0.18539
	E	0.09353	High	$9.1 \cdot 10^{13}$	0.003517	-0.0496995	0.27162	-0.7136	0.90439	-0.48197	0.14894	-0.2541
			Low	$1.1 \cdot 10^{13}$	-0.006006	0.055583	-0.1833	0.25838	-0.17733	0.13851	-0.021422	-0.23899
$M_{\rm FoF}$	I	0.10508	High	$9.9 \cdot 10^{13}$	-0.0038987	0.053565	-0.28691	0.74388	-0.90188	0.30641	0.14011	0.0097592
			Low	$1.2 \cdot 10^{13}$	-0.012759	0.15123	-0.70827	1.64118	-1.88347	0.84521	-0.007369	0.001575
	T	0.07584	High	$1.1 \cdot 10^{14}$	-0.000118	-0.000252	0.007728	-0.011479	-0.067499	0.14646	0.023399	-0.18334
			Low	$1.4 \cdot 10^{13}$	0.000479	-0.009243	0.06816	-0.21263	0.26048	-0.094668	0.079228	-0.18782
	E	0.083702	High	$9.7 \cdot 10^{13}$	0.001467	-0.022737	0.13515	-0.38237	0.513565	-0.2783	0.0851875	-0.20423
			Low	$1.1 \cdot 10^{13}$	-0.004819	0.0312925	-0.041542	-0.10184	0.27748	-0.162625	0.056434	-0.20258
$M_{\rm vir}$	I	0.10736	High	$7.7 \cdot 10^{13}$	-0.007327	0.0981775	-0.51147	1.29182	-1.56032	0.63605	0.10432	0.039672
			Low	$1.2 \cdot 10^{13}$	-0.010028	0.14207	-0.75163	1.89881	-2.32218	1.10993	-0.043551	0.030321
	T	0.08732	High	$8.7 \cdot 10^{13}$	-0.000489	0.000038	0.027968	-0.10122	0.0504345	0.11079	0.065778	-0.18126
			Low	$1.4 \cdot 10^{13}$	0.007257	-0.07183	0.30055	-0.64285	0.63984	-0.20756	0.12265	-0.18632
	E	0.09377	High	$9.1 \cdot 10^{13}$	0.003673	-0.051746	0.28257	-0.745065	0.95693	-0.52998	0.16749	-0.25643
			Low	$1.1 \cdot 10^{13}$	-0.003532	0.031089	-0.088055	0.073326	0.012294	0.039115	0.001966	-0.24085
M_{200c}	I	0.11297	High	$7.5 \cdot 10^{13}$	-0.008147	0.10478	-0.52113	1.24371	-1.37842	0.41825	0.188215	0.056463
			Low	$1.1 \cdot 10^{13}$	-0.016437	0.20509	-0.98707	2.30626	-2.62546	1.15052	-0.015322	0.047557
	T	0.09068	High	$8.4 \cdot 10^{13}$	0.0004415	-0.0121	0.089187	-0.2507	0.23287	-0.0005205	0.10798	-0.18964
			Low	$1.3 \cdot 10^{13}$	0.018484	-0.18578	0.74942	-1.50578	1.47238	-0.584875	0.19915	-0.19541
	E	0.095404	High	$8.8 \cdot 10^{13}$	0.002957	-0.043446	0.24636	-0.67084	0.88455	-0.50383	0.1736	-0.26358
			Low	$1.1 \cdot 10^{13}$	-0.010396	0.09199	-0.292916	0.40304	-0.25632	0.14676	-0.010236	-0.24538
M_{500c}	I	0.14951	High	$6.0 \cdot 10^{13}$	-0.003715	0.044544	-0.18354	0.25271	0.210546	-0.89752	0.61044	0.15031
			Low	$7.8 \cdot 10^{12}$	-0.034305	0.35908	-1.45797	2.84532	-2.56545	0.60461	0.2429	0.13191
	T	0.12154	High	$6.8 \cdot 10^{13}$	0.003303	-0.045958	0.24589	-0.60355	0.63199	-0.25061	0.24103	-0.2156
			Low	$9.3 \cdot 10^{12}$	-0.009472	0.1201	-0.55419	1.25785	-1.60611	1.1218	-0.14866	-0.18471
	E	0.11713	High	$7.1 \cdot 10^{13}$	0.004588	-0.071084	0.41551	-1.14448	1.50791	-0.87629	0.29411	-0.29091
			Low	$7.5 \cdot 10^{12}$	-0.009276	0.073287	-0.19254	0.17829	-0.060174	0.1048	0.014796	-0.26143

Table 3.7: Environment dependent fits for $z = 2$ halos in Illustris (I), TNG (T), and EAGLE (E). The first four columns are the halo definition, simulation, the median environment factor δ, and the type of environment. The fifth column is the upper mass limit for each fit. (The lower limit is $10^{10}h^{-1}M_{\odot}$ for all fits.) The remaining columns are the polynomial coefficients for the fits.

Halo	Sim.	δ_{med}	Env.	Mass Limit	a	b	c	d	e	f	g	h
M_{200b}	I	0.17955	High	$2.7 \cdot 10^{13}$	-0.003731	0.048442	-0.24134	0.57591	-0.64481	0.21476	0.046485	0.13021
			Low	$7.8 \cdot 10^{12}$	0.023188	-0.18283	0.51902	-0.63447	0.32652	-0.14135	0.078106	0.081198
	T	0.14689	High	$3.0 \cdot 10^{13}$	0.005096	-0.066039	0.33594	-0.821535	0.94879	-0.48134	0.24659	-0.168485
			Low	$8.5 \cdot 10^{12}$	0.009864	-0.112245	0.498236	-1.08559	1.15849	-0.55096	0.22659	-0.17404
	E	0.14676	High	$2.0 \cdot 10^{13}$	-0.000368	-0.007503	0.083543	-0.27045	0.33387	-0.13932	0.10477	-0.2383
			Low	$6.7 \cdot 10^{12}$	0.022763	-0.17532	0.51585	-0.72836	0.47585	-0.087887	0.081362	-0.22936
M_{FoF}	I	0.17581	High	$2.8 \cdot 10^{13}$	-0.010735	0.1129	-0.46652	0.95622	-0.98294	0.39474	-0.002468	0.067686
			Low	$8.5 \cdot 10^{12}$	0.03765	-0.306197	0.972865	-1.54527	1.33716	-0.69812	0.21531	0.019684
	T	0.14349	High	$3.3 \cdot 10^{13}$	0.002372	-0.030459	0.157025	-0.40171	0.51151	-0.31237	0.16074	-0.18052
			Low	$1.0 \cdot 10^{13}$	-0.004453	0.036516	-0.10105	0.09468	-0.004631	-0.018069	0.08073	-0.180235
	E	0.14364	High	$2.8 \cdot 10^{13}$	-0.0006545	0.003492	-0.001252	-0.009285	-0.026647	0.085026	0.003363	-0.172798
			Low	$6.5 \cdot 10^{12}$	-0.001431	0.028575	-0.16361	0.41671	-0.54417	0.36996	-0.050129	-0.17439
M_{vir}	Is	0.1777	High	$2.7 \cdot 10^{13}$	-0.003737	0.04692	-0.22826	0.53662	-0.59528	0.19207	0.049992	0.12173
			Low	$8.2 \cdot 10^{12}$	0.030668	-0.25239	0.77104	-1.083923	0.73852	-0.32675	0.11433	0.073593
	T	0.1453	High	$3.0 \cdot 10^{13}$	0.005433	-0.070022	0.35457	-0.866545	1.008868	-0.52069	0.25328	-0.16851
			Low	$9.0 \cdot 10^{12}$	0.000086	-0.0182245	0.145583	-0.434094	0.54356	-0.27228	0.17328	-0.17024
	E	0.14621	High	$2.1 \cdot 10^{13}$	-0.001168	0.002568	0.033903	-0.14905	0.17904	-0.038044	0.0708725	-0.23216
			Low	$6.8 \cdot 10^{12}$	0.019864	-0.15534	0.46681	-0.677998	0.45768	-0.085613	0.0767575	-0.2289
M_{200c}	I	0.18239	High	$2.6 \cdot 10^{13}$	-0.006626	0.081877	-0.392755	0.91643	-1.043495	0.44873	-0.015381	0.148625
			Low	$7.3 \cdot 10^{12}$	0.020516	-0.15355	0.40556	-0.43574	0.16161	-0.083549	0.069883	0.087871
	T	0.14878	High	$2.9 \cdot 10^{13}$	0.005585	-0.070937	0.3538	-0.84875	0.96079	-0.47725	0.24784	-0.16883
			Low	$7.9 \cdot 10^{12}$	0.017738	-0.18518	0.763505	-1.56374	1.59917	-0.74486	0.2643	-0.17738
	E	0.14769	High	$1.9 \cdot 10^{13}$	-0.001459	0.003902	0.03811	-0.18529	0.26048	-0.1197	0.11165	-0.24072
			Low	$6.5 \cdot 10^{12}$	0.006712	-0.032013	0.0168	0.13133	-0.29405	0.25469	0.020221	-0.2277
M_{500c}	I	0.20976	High	$1.9 \cdot 10^{13}$	-0.012781	0.13226	-0.511965	0.88291	-0.55058	-0.18947	0.22475	0.26724
			Low	$3.6 \cdot 10^{12}$	-0.033887	0.34532	-1.35019	2.529825	-2.26757	0.735285	0.016232	0.189495
	T	0.17413	High	$2.1 \cdot 10^{13}$	0.017809	-0.19951	0.873875	-1.84515	1.84797	-0.79746	0.329197	-0.170197
			Low	$8.3 \cdot 10^{12}$	0.055304	-0.50264	1.818375	-3.29399	3.015438	-1.273185	0.39279	-0.18187
	E	0.17109	High	$1.3 \cdot 10^{13}$	-0.004448	0.03157	-0.068125	0.0513	-0.081447	0.13874	0.079324	-0.24213
			Low	$4.2 \cdot 10^{12}$	-0.02677	0.24861	-0.875225	1.49326	-1.36135	0.6674	0.003002	-0.23464

means that halos which we identified as being in high-density environments might now be identified as being in low-density environments. In this case, it would not be appropriate to apply our environment-dependent halo mass corrections to this sample. This is admittedly a limitation of our environment-dependent halo mass corrections. The effect of using the “wrong” value of $\delta_{\rm med}$ is very slight, but we still recommend using the value of $\delta_{\rm med}$ that we provide to ensure that the mass corrections are applied to the appropriate halos.

We once again provide upper mass limits for our halo mass corrections, and emphasize the importance of not applying the corrections to any halos above these mass limits or below $10^{10}h^{-1}M_{\odot}$.

3.6 Halo Mass - Concentration Relation

Halo modeling often uses the internal structure of halos to determine the placement of satellite galaxies within the halo (e.g. Zehavi et al., 2011; Zentner et al., 2014). This often involves assuming that the spatial distribution of satellite galaxies within halos follows an NFW profile (Navarro et al., 1996), which includes a parameter for the concentration of the halo. Up to this point, we have investigated and quantified the impact of baryonic physics on the halo mass function, as well as the halo correlation function. We now examine the impact of baryonic physics on halo concentration.

The concentration c of a halo is defined as the ratio of the virial radius $R_{\rm vir}$ of the halo to the scale radius R_s (Navarro et al., 1997). The relationship between halo mass and concentration has been previously studied in simulations, and it has been found that halo concentration has a weak power-law dependence on halo mass, with a slope of approximately -0.13 (Bullock et al., 2001). This relationship has a great deal of scatter, and is predicated on the assumption that the density within a dark matter halo follows an NFW profile.

Ragagnin et al. (2019) investigated the dependence of halo concentration on mass and redshift in the Magneticum hydrodynamic simulations, and later Ragagnin et al. (2021)

examined the cosmology dependence of halo masses and concentrations in hydrodynamic simulations. Bose and Loeb (2020) examined the mass and concentration of host halos in IllustrisTNG as they relate to velocity dispersion. Additionally, Wang et al. (2020) studied the density profiles of early-type galaxies (ETGs) in IllustrisTNG at $z = 0$, and found that the profiles are steeper in the hydrodynamic simulation than their counterparts in the DMO simulation. They also found that the density profiles of the ETG dark matter halos are well described by steeper than NFW profiles.

In order to investigate the effect of baryonic physics on the halo mass - concentration relation in Illustris, TNG100, and EAGLE, we must first calculate the concentration for each halo (above $10^{10}h^{-1}M_{\odot}$) in these simulations. While concentration is not directly provided in any of these simulations' halo catalogues, we can estimate the concentration from the various halo mass definitions that are provided: M_{200b}, M_{200c}, and M_{500c}. The method works as follows. Assuming that each halo obeys an NFW profile, we can integrate the profile and use the virial mass and virial radius of the halo to determine the scale density ρ_0 of the halo, substituting $R_{\rm vir}/c$ for R_s:

$$M_{\rm vir} = \int_0^{R_{\rm vir}} 4\pi r^2 \rho(r) dr \tag{3.3}$$

$$\rho_0 = \frac{M_{\rm vir} c^3}{4\pi R_{\rm vir}^3 [ln(1+c) - \frac{c}{1+c}]}. \tag{3.4}$$

We can then substitute this to solve for any other halo mass M_h and radius R_h (e.g. M_{200b} and R_{200b}):

$$M_h = \frac{M_{\rm vir}}{ln(1+C) - \frac{c}{1+c}} [ln(1 + cR_h/R_{\rm vir}) - \frac{cR_h}{cR_h + R_{\rm vir}}]. \tag{3.5}$$

We can rearrange this equation to find the ratio of $M_{\rm vir}$ to M_h as a function of $R_{\rm vir}$, R_h, and c:

$$\frac{M_{\rm vir}}{M_h} = \frac{ln(1+c) - \frac{c}{1+c}}{ln(1 + cR_h/R_{\rm vir}) - \frac{cR_h}{cR_h + R_{\rm vir}}}. \tag{3.6}$$

For each of our simulations, we have three known values of M_h: M_{200b}, M_{200c}, and M_{500c}, and three corresponding values of R_h. To find the value of concentration for each halo, we loop over possible values of c between 0 and 2000 and find the one that minimizes the sum of the squared fractional difference between the left- and right-hand side of the previous equation over the three halo definitions. In other words,

$$A = \frac{M_{\rm vir}}{M_h} \tag{3.7}$$

and

$$B = \frac{ln(1+c) - \frac{c}{1+c}}{ln(1+cR_h/R_{\rm vir}) - \frac{cR_h}{cR_h+R_{\rm vir}}}, \tag{3.8}$$

and we find the value of c for each halo that minimizes

$$(\frac{B_{200b}-A_{200b}}{A_{200b}})^2 + (\frac{B_{200c}-A_{200c}}{A_{200c}})^2 + (\frac{B_{500c}-A_{500c}}{A_{500c}})^2. \tag{3.9}$$

We have tested this method of determining concentration on a DMO halo catalogue for which we know the concentration values for each halo (calculated by the ROCKSTAR; Behroozi et al. 2013). For this particular DMO halo catalogue, using the given values of concentration led to a halo mass - concentration relation with a slope of -0.104 ± 0.004, while using our method of determining concentration led to a slope of -0.098 ± 0.006. Therefore, we can conclude that our method leads to the correct overall halo mass - concentration relation.

After confirming its accuracy, we applied this method to each of our hydrodynamic and DMO simulations to determine the concentration of each halo. In Figure 3.7 we plot log(c) as a function of log($M_{\rm vir}$) (in units of $h^{-1}M_\odot$) for Illustris (left), TNG100 (middle), and EAGLE (right) halos. The results for the DMO simulations are plotted as gray points, while the results for the hydrodynamic simulations are plotted in blue, green, and orange, respectively. We then bin the points by mass, and plot the mean and standard deviation of

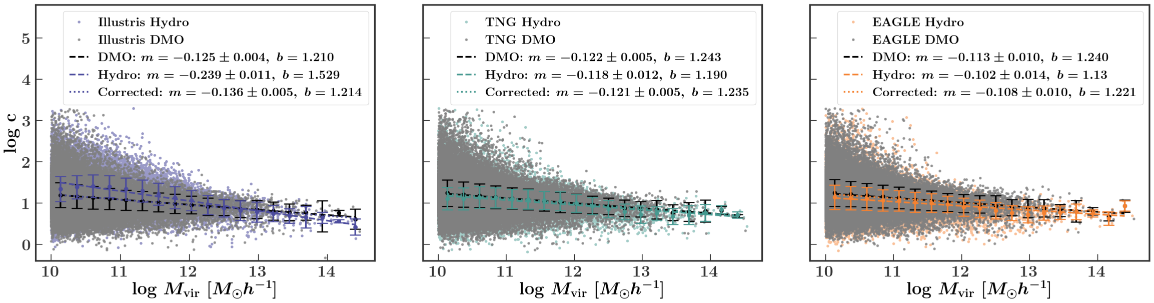

Figure 3.7: Concentration as a function of $M_{\rm vir}$ halo mass (in units of $h^{-1}M_\odot$) for Illustris (left), TNG100 (center), and EAGLE (right) halos at $z = 0$. DMO halos are plotted in gray for each simulation, while hydrodynamic halos are plotted in blue (Illustris), green (TNG100), and orange (EAGLE). The larger points in each panel are the mean concentrations in bins of halo mass for each DMO (black) and hydrodynamic (blue/green/orange) simulation, along with their standard deviations. Additionally, we fit a line to these binned points for each simulation, with the corresponding slope and y-intercept shown in the legend of each panel.

these bins (in black for DMO or blue/green/orange for the hydrodynamic simulations). We subsequently fit a line to these means, which we plot with a dashed line. (Note that the dashed line is not connecting the points, but rather is a fit to the points.) The slope and y-intercept for each simulation are given in Figure 3.7, along with the standard error in the slope.

Based on these results, we can see that all three DMO simulations have halo mass - concentration relations that are consistent with each other and with what we expect from previous studies. The Illustris DMO simulation has a halo mass - concentration relation with a slope of -0.125 ± 0.004, while the TNG DMO simulation has a halo mass - concentration relation with a slope of -0.122 ± 0.005, and the EAGLE DMO simulation has a halo mass - concentration relation with a slope of -0.113 ± 0.010 (only slightly shallower than expected).

The Illustris hydrodynamic simulation has a halo mass - concentration relation with a slope of -0.239 ± 0.011, which is steeper than we would expect, and not consistent with that of the DMO simulation. The TNG100 hydrodynamic simulation has a halo mass - concentration relation with a slope of -0.118 ± 0.012, and the EAGLE hydrodynamic simulation has a slope of -0.102 ± 0.014, both of which are consistent with their corresponding

DMO simulations. In each of these cases, the scatter among both the DMO and hydrodynamic halos is quite large, but there does not appear to be a systematic offset between the DMO and hydrodynamic distributions for any simulation. Based on these results, we can conclude that in TNG and EAGLE, baryonic physics does not significantly impact the halo mass - concentration relation, while in Illustris, baryonic physics results in a slightly steeper halo mass - concentration relation compared to that in Illustris-Dark.

We would also like to know what the halo mass - concentration relation would look like in each of these simulations if we corrected the DMO halo masses, but did not alter their concentrations. Another way of saying this is if someone were to apply our mass correction to a large DMO box, but not change the halo concentration parameter, how would that impact the halo mass - concentration relation? To investigate this, we apply our "abundance matching" halo mass correction to our DMO halos for each simulation, and then plot the original DMO halo concentrations as a function of these corrected masses. In this case, we once again do a direct halo-by-halo correction rather than applying our seventh-order polynomial fit, in order to perfectly reproduce the mass function from the hydrodynamic simulation. Thus, the concentration (which we found via the method outlined above using the original halo masses and radii) of the most massive DMO halo is plotted against the mass of the most massive hydrodynamic halo (i.e. the "corrected" DMO halo mass), and so on. We then find the mean concentration in bins of mass, and fit a new line to this relationship. This new fit is plotted as a dotted line in each panel of Figure 3.7, and the slope and y-intercept are given in the legend of each panel.

In each case, the "corrected" halo mass - concentration relation has a slope that is still consistent with what we would expect for halos obeying an NFW profile. In all three cases, applying the correction achieves a slope that is slightly closer to the hydrodynamic results. This is most significant in Illustris, where the slope becomes steeper, although still does not agree fully with the hydrodynamic results. In TNG and EAGLE, the DMO, hydrodynamic, and corrected results are all consistent with one another.

Based on this, for any work relying on halo masses and concentrations from a DMO simulation, we believe that depending on how one wants to employ the halo mass - concentration relation, one can either correct only the halo masses (and leave the concentrations untouched, with little impact on the results), or one can correct the halo concentrations as well using the slopes from the full-physics versions of Illustris, TNG, or EAGLE (shown in Figure 3.7). Either way, the impact of baryonic physics on the concentration-mass relation is likely to be small.

3.7 Conclusions

The implementation of baryonic physics in hydrodynamic simulations results in halo mass functions that are generally shifted to lower masses than those produced by dark matter only simulations. This is because stellar and AGN feedback removes baryonic particles (as well as some dark matter particles) from halos over time. This effect varies with halo mass: stellar feedback has more of an impact on lower mass halos, while AGN feedback has more of an impact on higher mass halos. Additionally, particularly efficient star formation can serve to *increase* the masses of some hydrodynamic halos (compared to their DMO counterparts).

In this work, we have quantified the relationship between the masses of halos in hydrodynamic simulations and those in corresponding DMO simulations. The impact of baryonic physics on the halo mass function depends on redshift, as well as halo definition. Additionally, because different hydrodynamic simulations contain different baryonic physics prescriptions, the halo mass discrepancy between hydrodynamic and DMO halos varies widely from one simulation to the next. Furthermore, the impact of baryonic physics on halo mass depends somewhat on the large-scale environment of the halo: in Illustris and TNG, halos in low-density environments are more impacted by baryonic physics, leading to greater discrepancies in their masses compared to their DMO counterparts. In EAGLE, halos in high-density environments exhibit a greater mass discrepancy between hydrodynamic and DMO simulations. This indicates that in hydrodynamic simulations in general, the strength of feedback (and its ability to impact halo mass) has a dependence on the density of the halo's environment.

We have found that this relationship as a function of DMO halo mass is well-fit with a seventh-order polynomial. We provide these fits for Illustris, IllustrisTNG100, and EAGLE halos for M_{200b}, $M_{\rm FoF}$, $M_{\rm vir}$, M_{200c}, and M_{500c} halos at $z = 0, 1$, and 2. These fits are based on matching halos by mass (i.e. "abundance matching") across hydrodynamic and DMO simulations. In other words, these are the corrections one would need to apply to the halos from Illustris-Dark, for example, to reproduce the halo mass function in the full-physics version of Illustris. We also provide these same fits after taking halo environment into account, which can be used to reproduce the conditional as well as the global mass function. Furthermore, we have shown that these corrections for halo mass also reproduce the large-scale clustering of halos, though the environment-dependent corrections are required to achieve an accuracy better than 2%. Finally, we have shown that baryonic effects do not impact the halo concentration-mass relation substantially. Our halo mass corrections are publicly available as a PYTHON module at https://github.com/gbeltzmo/halo_mass_correction.

Any work relying on halo catalogues from DMO simulations (e.g., halo occupation distribution modeling, conditional luminosity function modeling, stellar-to-halo mass relation modeling, etc.) could potentially be impacted by inaccuracies in the halo mass function. In particular, as these types of analyses start to be used more to constrain cosmological parameters, it is imperative that any conclusions are robust to changes in the halo mass function on the order of what we find in this work. For example, a DMO simulation created with a given set of cosmological parameters can produce a halo mass function that, after being adjusted with one of our mass corrections, resembles a HMF from a different cosmology. Thus, without understanding the uncertainty in the halo mass function due to baryonic physics, it may be challenging to distinguish between these cosmological models.

We recommend that any future work utilizing a halo catalogue from a DMO simulation repeat their analysis after applying at least one of our halo mass corrections (and ideally more than one). This will provide a rough estimate of the systematic uncertainty in one's results due to baryonic effects on the halo mass function. We make no assumptions about which of the three hydrodynamic simulations used in this work produces the "correct" halo mass function, but rather provide our halo mass corrections as a method for determining the robustness of one's modelling results to changes in the halo mass function.

CHAPTER 4

Toward Accurate Modeling of Galaxy Clustering on Small Scales: Extending the Halo Model

Halo models provide a simple and flexible framework for accurately modeling the small-scale clustering of galaxies. The standard Halo Occupation Distribution (HOD) model relies on the assumption that the number of galaxies in a given halo is solely dependent on the halo's mass. In this work, we employ a "decorated" HOD model, which allows for the possibility that halo occupation exhibits some dependence on a secondary halo property (a phenomenon known as assembly bias, or secondary bias). We choose to use halo concentration as the secondary halo property on which to model this assembly bias. Building on the framework established in Szewciw et al. (2022) (hereafter S22), we identify an optimal set of clustering statistics measured on a variety of scales to constrain this decorated HOD model. We use this modeling framework to constrain the galaxy-halo connection in SDSS DR7. For low-luminosity galaxies, our constraints indicate the presence of strong central galaxy assembly bias and moderate satellite galaxy assembly bias. Additionally, our best-fit model exhibits significantly less tension with the clustering of SDSS galaxies than was found in S22. For high-luminosity galaxies, our results are consistent with zero assembly bias for both central and satellite galaxies, and our best-fit model does not relieve the tension found in S22. These results emphasize the importance of including secondary bias parameters in the HOD modeling framework, as well as the value of using a variety of clustering statistics to probe different aspects of the galaxy-halo connection.

4.1 Introduction

Halo models are motivated by our understanding that galaxies form and reside in gravitationally bound, virialized regions of dark matter known as halos (e.g. Neyman and Scott, 1952; Peebles, 1974; McClelland and Silk, 1977; Scherrer and Bertschinger, 1991; Kauffmann et al., 1997; Jing et al., 1998; Baugh et al., 1999; Kauffmann et al., 1999; Benson et al., 2000; Ma and Fry, 2000; Peacock and Smith, 2000; Seljak, 2000; Scoccimarro et al., 2001; Sheth et al., 2001; White et al., 2001; Cooray and Sheth, 2002). These models assume that the clustering of galaxies can be fully described by (i) the clustering of their host halos and (ii) the way in which galaxies occupy these halos.

A key ingredient of the halo model is the Halo Occupation Distribution (HOD), which specifies via a few parameters the probability that a halo of mass M contains N galaxies (above some luminosity threshold) (Berlind and Weinberg, 2002; Berlind et al., 2003). The standard form of the HOD (Zheng et al., 2005) contains at most five free parameters that specify the mean occupation number of galaxies and assumes that galaxies trace the dark matter inside halos. Constraining these parameters when fitting to observational data provides a useful empirical measurement against which we can test competing theories of galaxy formation and evolution.

Many works have used the standard HOD to model the clustering of galaxies in recent redshift surveys (e.g., Zehavi et al., 2011; Guo et al., 2016). However, several of these studies yield fits which would rule out the ΛCDM + HOD model if taken at face value. The errors used in these studies are typically derived via the jackknife method, which has been shown to produce biased results (Norberg et al., 2009). Sinha et al. (2018) (S18 hereafter) developed a numerical mock-based modeling procedure that significantly improved the accuracy of HOD modeling. They compared the clustering of galaxies in the Sloan Digital Sky Survey (SDSS, York et al., 2000) to a ΛCDM + standard HOD model, measuring the projected correlation function, group multiplicity function, and galaxy number density. Carefully controlling for systematic errors allowed them to interpret the goodness of fit of their model. They found that their best-fit HOD model was unable to jointly fit the clustering statistics, revealing significant tension between SDSS and their ΛCDM + HOD model. Because this tension did not exist when they considered only measurements of the projected correlation function (as is done in many studies), S18 demonstrated the value of adding additional statistics in small-scale clustering analyses.

Szewciw et al. (2022) extended the procedure used in S18 in order to maximize the return from spectroscopic surveys. They included the same clustering statistics used in S18 (galaxy number density, projected correlation function, and group multiplicity function) as well as four additional clustering statistics: redshift-space correlation function, group velocity dispersion, mark correlation function, and counts-in-cells statistics. They were able to significantly tighten the constraints on their five-parameter HOD model, as well as increase the tension found in S18.

This increase in tension suggests a need to expand the standard HOD model to include additional features. For example, the standard HOD model assigns galaxies to halos based solely on the halo's mass, but it is possible that halo occupation depends on additional (secondary) features of the halo (e.g. concentration), a phenomenon known as assembly bias (Gao et al., 2005; Berlind et al., 2006b; Wechsler et al., 2006; Croton et al., 2007).

Several works have examined the potential for the presence of assembly bias to affect constraints on the galaxy-halo connection and cosmology. For example, Zentner et al. (2014) examined the potential for assembly bias to induce systematic errors in inferred halo occupation statistics. They built mock galaxy catalogs that exhibited assembly bias as well as companion mock catalogs with identical HODs but no assembly bias. They fit HOD models to the galaxy clustering in each catalog, and found that the inferred HODs described the true HODs well in the mocks without assembly bias, but in the mocks with assembly bias the inferred HODs exhibited significant systematic errors. Later, McCarthy et al. (2019) used a mock galaxy catalog with assembly bias to study how assembly bias might affect cosmological constrains. Specifically, they used the large-scale redshift-space distortion to probe $f\sigma_8$. They found that on small scales (e.g. a few to tens of h^{-1}Mpc) galaxy assembly bias can introduce systematic uncertainties in cosmological constraints. They concluded that galaxy assembly bias can only be ignored when modeling scales above 8 h^{-1}Mpc, where clustering is determined purely by the large scale bias.

A number of works have investigated the presence of assembly bias in dark matter only simulations. For example, Villarreal et al. (2017) used dark matter only simulations to examine the dependence of assembly bias on halo definition, and found that the effect of assembly bias for low mass halos can be mitigated through the use of a mass-dependent halo definition with a smaller spherical overdensity threshold for lower mass halos. This mass-dependent halo definition subsumes backsplash halos into larger host halos. They also found shape- and spin-dependent clustering to be significant for all halo definitions, with weaker mass dependence. They conclude that no halo definition mitigates all manifestations of assembly bias.

Salcedo et al. (2018) explored halo assembly bias in the Large Suite of Dark Matter Simulations and found that a clustering bias exists if halos are binned by mass or by any other halo property, indicating that no single halo property encompasses all the spatial clustering information of the halo population. They also found that the mean values of some halo properties depend on their halo's distance to a more massive neighbor, and concluded that this "neighbor bias" largely accounts for the secondary bias seen in halos binned by mass and split by concentration or age. However, they also found that halos binned by other mass-like properties still show a secondary bias even when the neighbor bias is removed. Meanwhile, Mao et al. (2018) presented a summary of secondary halo biases of high-mass halos due to various halo properties (e.g. concentration, spin, several proxies of assembly history, and subhalo properties) in the MultiDark Planck 2 simulation. They found that, while concentration, spin, and the abundance and radial distribution of subhalos exhibit significant secondary biases, properties that directly quantify halo assembly history do not.

Mansfield and Kravtsov (2020) used the Bolshoi simulations to examine the physical processes that lead to halo assembly bias, focusing on the origin of assembly bias in the mass range corresponding to the hosts of typical galaxies. Using halo concentration as a proxy of halo formation time, they found that splashback subhalos are responsible for two-thirds of the assembly bias signal, but do not account for the entire effect. After splashback subhalos are removed, they found that the remaining assembly bias signal is due to a small fraction of halos in dense regions. They tested several additional physical processes thought to contribute to assembly bias, and concluded that three main processes modify the assembly bias of small-mass halos: large-scale tidal fields, gravitational heating due to the collapse of large-scale structures, and splashback subhalos located outside the virial radius.

Behroozi et al. (2021) examined the correlation of different properties of dark matter halos (e.g. growth rate, spin, concentration) with environment in the Small MultiDark Planck simulation. They demonstrated that these halo properties imprint distinct signatures in the galaxy two-point correlation function and in the distribution of distances to galaxies' kth nearest neighbors. Finally, they computed two-point correlation functions for SDSS galaxies binned by half-mass radius at $z = 0$, showing that classic galaxy size models (i.e. galaxy size being proportional to halo spin) as well as other recent proposals show significant tensions with observational data. They demonstrated that the agreement with observed clustering can be improved with a simple empirical model in which galaxy size correlates with halo growth.

Several other works have explored assembly bias in semi-analytic models of galaxy formation. For example, Pujol and Gaztañaga (2014) looked at two-point clustering in semi-analytic models from the Millennium Simulation, and found evidence that galaxy clustering is affected by assembly bias in low-mass halos ($M < 3 \times 10^{11} h^{-1} M_{\odot}$). Later, Pujol et al. (2017) examined the bias of central galaxies in semi-analytic models and found that using local density as a secondary property correctly predicts galaxy bias, while using solely halo mass does not. Zehavi et al. (2018) examined the dependence of the galaxy occupation of dark matter halos on large-scale environment and halo formation time using semi-analytic models applied to the Millennium simulation. They found that early-forming halos (and to a lesser extent halos in denser environments) are more likely to host central galaxies at lower halo mass, but also tend to host fewer satellite galaxies.

Contreras et al. (2019) examined the evolution of assembly bias using a semi-analytic model of galaxy formation applied to the Millennium-WMAP7 N-body simulation. They found that at $z = 0$, the dependence of both halo clustering and halo occupation on halo concentration and halo formation time are similar. At higher redshift, the assembly bias signature weakens for halos selected by age, and reverses and increases for halos selected by concentration. Meanwhile, the halo occupation variation with halo age stays mostly constant with increasing redshift, but decreases for concentration.

Finally, a handful of works have examined assembly bias in hydrodynamic simulations. For example, Artale et al. (2018) used the EAGLE and Illustris hydrodynamic simulations to examine the variations in galaxy occupancy of dark matter halos with the large-scale environment and halo formation time. They found that for low-mass halos at fixed halo mass, halos in denser environments are more likely to host a central galaxy, and early-formed halos are even more likely to host a central galaxy. Additionally, these halos are likely to host more massive central galaxies. Meanwhile, early-formed halos host fewer satellite galaxies. Later, Bose et al. (2019) examined the galaxy-halo connection in the IllustrisTNG simulations, and found that halos in dense environments, with low concentrations, late formation times, and high angular momenta are richest in their satellite population. Additionally, at low mass, halos with high-concentrations in overdense environments are more likely to host a central galaxy. They conclude that at fixed halo mass, concentration is a strong predictor of the stellar mass of the central galaxy.

Beltz-Mohrmann et al. (2020) investigated the ability of the standard HOD model to reproduce the small-scale galaxy clustering seen in hydryodynamic simulations. Comparing to the Illustris and EAGLE simulations, they found strong evidence for the need to extend the standard HOD model to include assembly bias parameters, particularly when investigating the clustering of lower-luminosity galaxies. Meanwhile, Xu and Zheng (2020) examined central galaxy assembly bias in the Illustris simulation and found that galaxy stellar mass has a tighter correlation with peak maximum halo circular velocity than with halo mass. Once the correlation with peak velocity is accounted for, stellar mass has nearly no dependence on any other halo assembly variables.

Hadzhiyska et al. (2020) and Hadzhiyska et al. (2021a) investigated the assumptions of the standard HOD model by comparing predictions of galaxy clustering to the IllustrisTNG simulations, and found that the standard HOD model fails to predict the correct galaxy clustering. They also explored using different secondary parameters in an HOD model that included assembly bias, and found that the local environment of the halo and the velocity dispersion anisotropy are the most effective measures of assembly bias for predicting clustering consistent with IllustrisTNG. They also found that at fixed halo mass, galaxies in one type of environment cluster differently from galaxies in another, and concluded that combining mass and local environment information about the halo leads to a more complete model of the galaxy-halo connection. Hadzhiyska et al. (2021c) further explored halo occupation in IllustrisTNG, and found that at fixed halo mass, galaxies in high-density environments cluster ten times more strongly than those in low-density regions. Finally, Hadzhiyska et al. (2021b) compared galaxy clustering in IllustrisTNG and the Santa-Cruz semi-analytic model and found good agreement between the two models for two-point clustering and galaxy assembly bias signatures. They also found that both models exhibited a similar response in halo occupancy and clustering to secondary halo properties such as formation history and concentration.

Contreras et al. (2021) used the IllustrisTNG simulation, the SAGE semi-analytic model, and subhalo abundance matching (SHAM) to investigate the differences in predictions of galaxy assembly bias, and found that all three models produced an assembly bias signal of different magnitude, redshift evolution, and dependence with selection criteria and number density. They found that by including an extension to SHAM that allows for arbitrary amounts of assembly bias, they were able to reproduce the galaxy assembly bias signature in both SAGE and IllustrisTNG, for all redshifts and galaxy number densities.

A few recent works have attempted to constrain assembly bias in observational surveys. Zentner et al. (2019) used the projected correlation function to constrain an HOD model with assembly bias in SDSS. They detected central galaxy assembly bias in the $M_r < -20$ and $M_r < -20.5$ samples, and detected satellite galaxy assembly bias in the $M_r < -19$ sample, but found no evidence for galaxy assembly bias in the $M_r < -21$ sample. Meanwhile, Vakili and Hahn (2019) used the Small MultiDark-Planck high resolution N-body simulation to model galaxy clustering in SDSS. They examined the concentration-dependence of halo occupation, and found that the satellite population is not correlated with halo concentration at fixed halo mass. They also found no correlation between the occupation of centrals and halo concentration in the most luminous samples ($M_r < -21.5, -21$), and modest correlation in the $M_r < -20.5, -20, -19.5$ samples. Additionally, Salcedo et al. (2020) investigated the level of galaxy assembly bias in the Sloan Digital Sky Survey using the ELUCID simulation, and found no evidence for significant galaxy assembly bias in the local Universe for galaxies above a stellar mass threshold of $10^{10.2} h^{-1} M_{\odot}$.

More recently, Lange et al. (2022) used an HOD model with both assembly bias and velocity bias parameters to marginalize over uncertainties in the galaxy-halo connection and obtain cosmological constraints from the BOSS LOWZ sample. Using $V_{\max}$ as their assembly bias property, they do not find significant evidence of either central or satellite galaxy assembly bias. However, Lange et al. (2019a) explored how galaxy assembly bias affects cosmological inference and found a degeneracy between assembly bias and $f\sigma_8$. Ultimately, they found that not including galaxy assembly bias in the model leads to a small shift in the posterior of $f\sigma_8$, indicating that it is important to account for galaxy assembly bias to obtain unbiased cosmological constraints.

Finally, McCarthy et al. (2022) used HOD modeling to investigate assembly bias in redshift (velocity) space in SDSS, with an extended construction of early- and late-forming galaxies. They found that while early- and late-forming central galaxies have consistent host halo masses, early-forming central galaxies exhibit large velocity bias, with central galaxies moving at more than 50 percent of the dark matter velocity dispersion inside host halos, signaling an assembly bias effect.

In this work, we model the clustering of SDSS galaxies using an HOD model with assembly bias parameters to constrain the galaxy-halo connection for both low- and high-luminosity galaxies. We build on the procedures developed in Sinha et al. (2018) and Szewciw et al. (2022), using a fully numerical, mock-based modeling procedure, with a wide variety of galaxy clustering statistics and a careful treatment of systematic errors. In Section 4.2 we describe our data, and in Section 4.3 we describe our simulations and halo catalogs. In Section 4.4 we describe our halo model, and in Section 4.5 we describe our full modeling procedure (including our mock galaxy catalogs, covariance matrices, clustering measurements, and MCMC framework). In Section 4.6 we describe our selection of optimal observables for constraining our HOD model, and in Section 4.7 we describe our results. We summarize our findings in Section 4.8.

4.2 Observational Data

In this work, we use the same observational dataset as that used in S22. We utilize the large scale structure samples from the NYU Value Added Galaxy Catalog (NYU-VAGC; Blanton et al., 2005) from the seventh data release (DR7; Abazajian et al., 2009) of the Sloan Digital Sky Survey (SDSS; York et al., 2000). The absolute magnitudes of the galaxies in this sample have been k-corrected to rest-frame magnitudes at redshift $z = 0.1$, but have not been corrected for passive luminosity evolution. From this sample, we construct two volume-limited subsamples, each complete down to a specified r-band absolute magnitude threshold ($M_r < -19$ and $M_r < -21$). We refer to these samples as the -19 and -21 samples throughout this paper. The luminosity thresholds, redshift limits, median redshifts, effective volumes, and number densities of our samples are listed in Table 4.1. The comoving distances of the SDSS galaxies in our samples are determined using a flat ΛCDM cosmological model with $\Omega_\mathrm{m} = 0.302$ and $h = 1$. Our distances are in units of h^{-1}Mpc, and our absolute magnitudes are actually $M_r + 5\log h$[1]. Fiber collisions are handled in the same way as in S22. For more details, see Sinha et al. (2018) and S22.

[1] Throughout this paper, log refers to $\log_{10}$.

Table 4.1: The columns list (from left to right): the absolute magnitude threshold of each sample at $z = 0.1$; the minimum, maximum, and median redshifts; the effective volume; and the galaxy number density of each sample. The volumes and number densities of the samples are corrected for survey incompleteness.

$M_r^{\lim}$	$z_{\min}$	$z_{\max}$	z_{median}	$V_{\text{eff}}(h^{-3}\text{Mpc}^3)$	$n_g(h^3\text{Mpc}^{-3})$
-19	0.02	0.07	0.0562	6,087,119	0.01453
-21	0.02	0.158	0.1285	67,174,396	0.00123

4.3 Simulations and Halo Catalogs

In our modeling procedure, we make use of the same dark matter only (DMO) cosmological N-body simulations as those used in S22. These simulations are from the Large Suite of Dark Matter Simulations project (LasDamas; McBride et al., 2009), and were run on the Texas Advanced Computing Center's Stampede supercomputer using the public code GADGET-2 (Springel, 2005). Power spectra were generated with CMBFAST (Seljak and Zaldarriaga, 1996; Zaldarriaga et al., 1998; Zaldarriaga and Seljak, 2000), and initial conditions were generated with 2LPTIC (Scoccimarro, 1998; Crocce et al., 2006, 2012). All simulations were run with the following cosmological parameters, based on results from the Planck experiment (Planck Collaboration et al., 2014): $\Omega_{\text{m}} = 0.302$, $\Omega_{\Lambda} = 0.698$, $\Omega_{\text{b}} = 0.048$, $h = 0.681$, $\sigma_8 = 0.828$, and $n_s = 0.96$. The details of these simulations are given in Table 4.2. We identify halos with a spherical over-density (SO; Lacey and Cole, 1994) threshold of M_{vir} (Bryan and Norman, 1998) using the ROCKSTAR phase-space temporal halo finder (Behroozi et al., 2013). Finally, for computational purposes, we randomly downsample to keep only 5% of the dark matter particles in each halo, with no loss of accuracy (see S22).

4.4 Halo Model

The Halo Occupation Distribution framework governs the number, positions, and velocities of galaxies within dark matter halos. The standard HOD model assigns galaxies to halos based on five free parameters, which depend only on the halo's mass (Zheng et al., 2007). Galaxies are split into centrals and satellites within their halos (Kravtsov et al., 2004; Zheng

Table 4.2: Simulation parameters. The columns list (from left to right): what each simulation is used for, the absolute magnitude threshold of the corresponding SDSS sample, the name of the simulation, the seeds used, the (comoving) boxsize (in h^{-1}Mpc), number of particles, mass resolution (in $h^{-1}\mathrm{M}_{\odot}$), (comoving) force softening (in h^{-1}kpc), and the number of simulations.

Use	Sample	Simulation	Seeds	L_{box}	N_{part}	m_{part}	ε	N_{sim}
Matrix	-19	Consuelo	4001 - 4100	420	1400^3	$2.26 \cdot 10^9$	8	100
Matrix	-21	Carmen	2001 - 2100	1000	1120^3	$5.97 \cdot 10^{10}$	25	100
MCMC	-19	ConsueloHD	4002, 4022	420	2240^3	$5.53 \cdot 10^8$	5	2
MCMC	-21	CarmenHD	2007, 2023	1000	2240^3	$7.46 \cdot 10^9$	12	2

et al., 2005). In this model, the mean number of central galaxies in a halo of mass M is described by

$$\langle N_{\mathrm{cen}} \rangle = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\log M - \log M_{\mathrm{min}}}{\sigma_{\log M}}\right)\right], \tag{4.1}$$

where M_{min} is the mass at which half of halos host a central galaxy, $\sigma_{\log M}$ is the scatter around this halo mass, and erf(x) is the error function, $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-y^2)dy$.

The central galaxy is always placed at the center of the halo, and assigned the mean velocity of the halo. The number of satellite galaxies in a given halo is drawn from a Poisson distribution with mean

$$\langle N_{\mathrm{sat}} \rangle = \langle N_{\mathrm{cen}} \rangle \times \left(\frac{M - M_0}{M_1}\right)^{\alpha}, \tag{4.2}$$

where M_0 is the halo mass below which there are no satellite galaxies, M_1 is the mass where halos contain one satellite galaxy on average, and α is the slope of the power-law occupation function at high masses. Each satellite galaxy is given the position and velocity of a randomly selected dark matter particle within the halo.

In this work we use the decorated HOD (dHOD) model of Hearin et al. (2016), used previously in Zentner et al. (2019). In this model, galaxies are assigned to halos based on both the halo's mass *and* a secondary halo property. In order to apply the decorated HOD, we first split halos by mass into bins of width 0.05 dex. Then, within each mass bin, we split halos into two groups based on the median value of the secondary property s in each bin. We then assign galaxies to halos based on

$$\langle N_{\rm cen}|M, s_{\rm high}\rangle = \langle N_{\rm cen}|M\rangle + \delta N_{\rm cen}, \tag{4.3}$$

$$\langle N_{\rm cen}|M, s_{\rm low}\rangle = \langle N_{\rm cen}|M\rangle - \delta N_{\rm cen}, \tag{4.4}$$

$$\langle N_{\rm sat}|M, s_{\rm high}\rangle = \langle N_{\rm sat}|M\rangle + \delta N_{\rm sat}, \tag{4.5}$$

and

$$\langle N_{\rm sat}|M, s_{\rm low}\rangle = \langle N_{\rm sat}|M\rangle - \delta N_{\rm sat}, \tag{4.6}$$

where

$$\delta N_{\rm cen} = A_{\rm cen}{\rm MIN}[\langle N_{\rm cen}|M\rangle, 1 - \langle N_{\rm cen}|M\rangle] \tag{4.7}$$

for central galaxies and

$$\delta N_{\rm sat} = A_{\rm sat}\langle N_{\rm sat}|M\rangle \tag{4.8}$$

for satellite galaxies.

$A_{\rm cen}$ and $A_{\rm sat}$ are between -1 and 1; values of 0 indicate no assembly bias. A key point is that, regardless of the strength of the assembly bias, $\langle N_{\rm cen}\rangle$ and $\langle N_{\rm sat}\rangle$ are preserved for a given halo mass. In other words, at fixed mass, for the same 5-parameter standard HOD model, the decorated HOD has the same halo occupation distribution when averaged over all halos.

One frequently used secondary property is halo concentration, c, which is defined as the ratio of the virial radius $R_{\rm vir}$ of the halo to the scale radius R_s (Navarro et al., 1997). For a given halo, concentration can be found using the relationship between virial mass, maximum circular velocity, and concentration at $z = 0$:

$$v_{\rm circ}(M_{\rm vir}) = \frac{6.72\times10^{-3} M_{\rm vir}^{1/3}\sqrt{c}}{\sqrt{ln(1+c) - c/(1+c)}} \tag{4.9}$$

where $M_{\rm vir}$ is the virial mass of the halo in units of $h^{-1}M_{\odot}$, and $v_{\rm circ}$ is the maximum circular velocity of the halo in units of km/s (Klypin et al., 2011). In our case, we implement halo concentration as our secondary bias property, and use $v_{\rm circ}/M_{\rm vir}^{1/3}$ as a proxy for halo concentration.

4.5 Modeling Procedure

4.5.1 Building mock galaxy catalogs

We build mock galaxy catalogs to use as our model by populating the two high-resolution simulations for each sample (ConsueloHD and CarmenHD) with galaxies. Once we populate our dark matter halos with galaxies, we build realistic mock galaxy catalogs that resemble our SDSS samples of interest. To do this, we transpose the mock galaxies from Cartesian to spherical coordinates by positioning an observer at the center of the box and converting the positions of the galaxies into RA, DEC, and comoving distances. We can then carve out four independent mock galaxy catalogs from each simulation box, and incorporate the same systematic effects that plague our observational dataset, such as redshift-space distortions, sample geometry, and incompleteness. For more details, see S22.

4.5.2 Covariance Matrices

If we wish to take advantage of the information present at small scales to constrain the galaxy-halo connection, it is essential that we carefully understand and minimize the uncertainty in our modeling procedure. To do this, we run 100 low-resolution simulations for each sample (Consuelo and Carmen) which differ in the phases of the density modes of the power spectrum, which is controlled by a seed supplied to 2LPTIC. We populate these low-resolution simulations with galaxies using the HOD parameters listed in Table 4.3. We then build 400 mock galaxy catalogs for each sample, from which we can construct a covariance matrix to represent cosmic variance. The elements of the covariance matrix are

Table 4.3: Fiducial HOD parameters for covariance matrices. The HOD parameters used to construct the covariance matrices in our analysis.

$M_r^{\rm lim}$	$\log M_{\rm min}$	$\sigma_{\log M}$	$\log M_0$	$\log M_1$	α	$A_{\rm cen}$	$A_{\rm sat}$
-19	11.54	0.22	12.01	12.74	0.92	0	0
-21	12.72	0.46	7.87	13.95	1.17	0	0

given by

$$C_{ij} = \frac{1}{N-1}\sum_{1}^{N}(y_i - \overline{y_i})(y_j - \overline{y_j}) \tag{4.10}$$

where the sum is taken over the $N = 400$ mocks. The values y_i and y_j are the ith and jth observables measured on each mock, while $\overline{y_i}$ and $\overline{y_j}$ are the mean values of the ith and jth observables, respectively. Each diagonal element, C_{ii}, of the matrix is the variance across 400 mocks for observable i, and $\sqrt{C_{ii}}$ is the cosmic variance uncertainty of observable i. For an arbitrary observable, we refer to this uncertainty as $\sigma_{\rm obs}$.

4.5.3 Clustering Statistics

Several works have demonstrated the power of using a variety of different clustering statistics to constrain the galaxy-halo connection (Berlind and Weinberg, 2002; Sinha et al., 2018; Hadzhiyska et al., 2021a; Szewciw et al., 2022). In our analysis, we employ the following clustering statistics: the projected correlation function $w_{\rm p}(r_{\rm p})$ (e.g. Zehavi et al., 2002, 2004; Zheng, 2004; Zehavi et al., 2005; Zheng et al., 2007; Zehavi et al., 2011; Leauthaud et al., 2012; Zentner et al., 2014; Coupon et al., 2015), the redshift-space correlation function $\xi(s)$ (e.g. Tinker et al., 2006b; Parejko et al., 2013; Guo et al., 2015a; Padilla et al., 2019; Beltz-Mohrmann et al., 2020; Tonegawa et al., 2020), the group multiplicity function $n(N)$ (e.g. Berlind et al., 2006a; Zheng and Weinberg, 2007; Sinha et al., 2018; Beltz-Mohrmann et al., 2020), the average group velocity dispersion function $\sigma_v(N)$, the mark correlation function mcf(s) (e.g. Zu and Mandelbaum, 2018), and two special cases of counts-in-cells $P_N(R)$: the void probability function P_0 (VPF(R)) and the singular probability function P_0 (SPF(R)) (e.g. Tinker et al., 2006a, 2008; McCullagh et al., 2017; Walsh and Tinker, 2019; Wang et al., 2019; Beltz-Mohrmann et al., 2020). A detailed description of each of these clustering statistics is given in S22. To calculate $w_{\rm p}(r_{\rm p})$, $\xi(s)$, mcf(s), VPF(R), and SPF(R) we make use of the publicly available code CORRFUNC (Sinha and Garrison, 2017, 2019). In our modeling procedure, we measure each clustering statistic in the same way (i.e., either on the full box/es or on the mock galaxy catalogs) as was done in S22.

4.5.4 MCMC

We explore the HOD parameter space with a Markov Chain Monte Carlo (MCMC) algorithm, using a privately developed C-implementation of the popular affine-invariant sampler EMCEE (Foreman-Mackey et al., 2013), which we call EMCEE_IN_C[2]. We impose flat priors on the same parameter ranges given in Sinha et al. (2018), as well as flat priors of [-1.0,1.0] on $A_{\rm cen}$ and $A_{\rm sat}$ for both samples.

At each point in the chain, we evaluate the likelihood that a particular HOD model could have generated a dataset with the same clustering as SDSS. This likelihood is given by

$$\mathscr{L}(\mathbf{D}|\mathbf{M}) = \frac{\exp(-\frac{1}{2}(\mathbf{D}-\mathbf{M})\mathbf{C}^{-1}(\mathbf{D}-\mathbf{M})^{T})}{\sqrt{(2\pi)^{K}\det(\mathbf{C})}}, \tag{4.11}$$

where $\mathbf{D}$ is the K-dimensional vector of observables measured on the SDSS dataset, $\mathbf{M}$ is the corresponding vector of observables measured on the HOD model, and $\mathbf{C}$ is the K-dimensional covariance matrix of these observables representing cosmic variance (see Equation 4.10). (The the term within the exponential is essentially χ^2, multiplied by a factor of $-1/2$.)

In the HOD framework, the process of populating halos with galaxies in the is stochastic, and is controlled with a "population seed." For a fixed HOD model, changes in this population seed can lead to significant differences in clustering statistics. To minimize the noise in our results due to this random variation, at each point in the chain we populate

[2] https://github.com/aszewciw/emcee_in_c

halos four times, using four fixed population seeds. Thus the clustering measurements for a given point in HOD parameter space are the average measurements over these four population seeds.

4.6 Choosing Optimal Observables

In order to constrain the dHOD when fit to SDSS, we must first choose a set of observables to use in our MCMC. We seek a subset of observables that produce the tightest constraints on our HOD parameters, at the cost of little noise. Noise is introduced into the covariance matrix due to the fact that we are constructing it from only 400 mocks. This noise propagates into the likelihood function and thus into our posterior results. Increasing the number of observables we use increases this noise, highlighting the need to choose our observables wisely.

To choose an "optimal" set of high-information, low-noise observables, we employ the importance sampling algorithm described in S22. In this algorithm, we first run MCMCs on four mock galaxy catalogs to create fiducial non-uniform grids of HOD points. When constructing these four fiducial grids, the likelihood of each point is calculated using only n_{gal} and $w_{\mathrm{p}}(r_{\mathrm{p}} \sim 0.3\ h^{-1}\mathrm{Mpc})$. We then use importance sampling on these grids to explore the constraining power of different combinations of clustering statistics. The algorithm chooses observables one by one, each time selecting the observable that, when combined with all previously chosen observables, produces the tightest projected constraints on all HOD parameters of interest. When choosing an observable, we consider how it performs on across all four grids, minimizing any bias due to cosmic variance. Thus, at the end of running this algorithm, we have a list of observables (ordered in terms of cumulative constraining power) and a corresponding list of cumulative projected constraints for each sample. (We refer the reader to S22 for a more complete description of this procedure.)

To choose the fiducial HOD from which the mocks used in this algorithm are constructed, we first run chains on our SDSS samples using matrices made from the HOD

Table 4.4: Initial SDSS best-fit results.

$M_r^{\rm lim}$	$\log M_{\rm min}$	$\sigma_{\log M}$	$\log M_0$	$\log M_1$	α	$A_{\rm cen}$	$A_{\rm sat}$
-19	11.456	0.141	11.64	12.701	0.94	0.75	-0.33
-21	12.763	0.538	11.14	13.939	1.05	-0.35	-0.25

parameters given in Table 4.3 and the 36 optimal observables chosen for each sample in S22 (listed in Table 4.5 under "vHOD"). These chains are run using a dHOD with concentration as the assembly bias parameter. The best-fit HOD parameters from these chains are listed in Table 4.4. We use these best-fit HOD parameters to build four mock galaxy catalogs for each sample, which are then used to construct the four grids used in our algorithm, as described above.

There are two key differences in our implementation of this algorithm compared to S22. First, when choosing the third observable for each sample, we only attempt to constrain $A_{\rm cen}$ and $A_{\rm sat}$. This is because these parameters are entirely unconstrained when using only $n_{\rm gal}$ and $w_{\rm p}(r_{\rm p} \sim 0.3\ h^{-1}{\rm Mpc})$, which causes the MCMC to explore unrealistic HOD models; thus, it is essential to choose an observable early on that provides constraining power for these parameters. After the third observable is chosen, we make all successive choices by attempting to jointly constrain all HOD parameters (excluding $\log M_0$ for the -21 sample). Second, in the S22 algorithm, new grids are created (by running new MCMCs using the already chosen observables) whenever the old grids become insufficiently dense for importance sampling. S22 creates these new grids after choosing five observables for each sample, and again for the -19 sample after choosing eight observables. In our case, we build denser grids after choosing three, five, ten, and twenty observables for each sample.

In Figure 4.1, we show our estimated constraint for each HOD parameter (excluding $\log M_0$) as we choose successive observables. The results for the -19 sample are shown in blue, and the results for the -21 sample are shown in red. The solid lines show the average constraint across the four mocks used in the algorithm described above. In Table 4.5, we list the observables chosen (in order) that we use for each sample (labeled "dHOD"). We

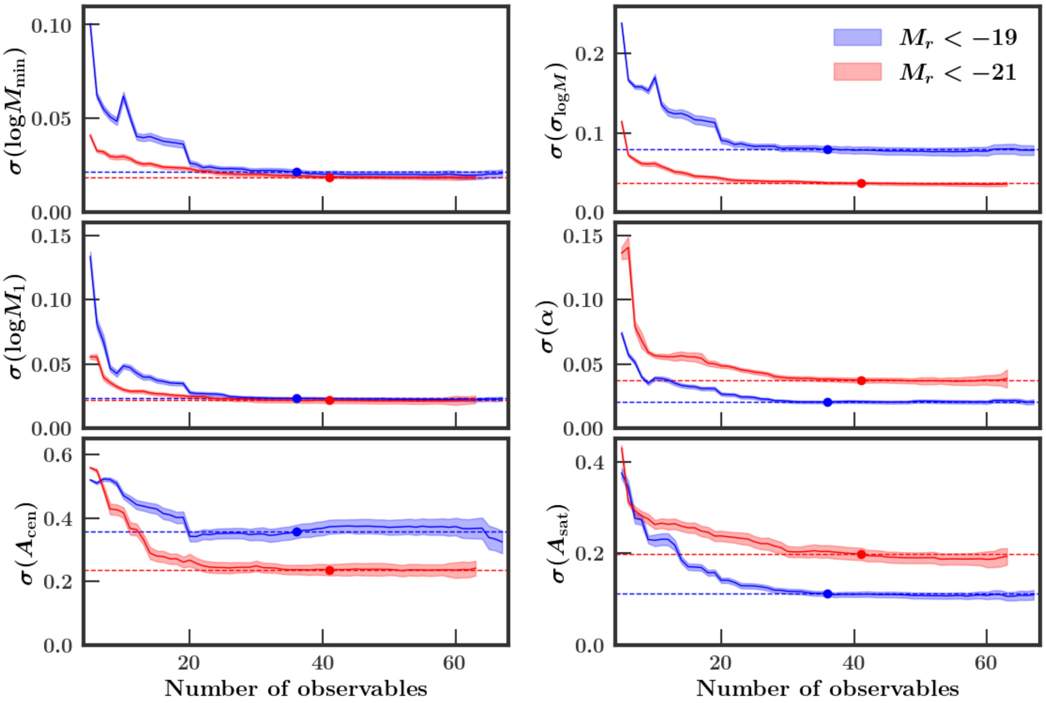

Figure 4.1: Constraints on each HOD parameter as we increase the number of observables, for the -19 sample (blue) and the -21 sample (red). The solid line in each panel shows the average mock constraint (1-σ) across four mocks, and the shaded region is an estimate of the uncertainty (inner 68%) in our constraints. The dot indicates the optimal number of observables for each sample, and the dashed line indicates the corresponding constraining power for each parameter.

also list the observables chosen in the previous analysis using a "vanilla" HOD model (i.e., no assembly bias, labeled "vHOD"). The observables chosen in this work that were *not* chosen in the previous analysis are shown in bold.

After ordering the observables from greatest to least constraining power, we need to choose the total number of observables to use in our analysis. To do this, we employ the same procedure as S22. Briefly, we estimate an error associated with each projected constraint (for a given number of observables K) by resampling the covariance matrix 100 times, and then importance sampling the chain with each of these resampled matrices. Doing so lets us approximate the error in our constraint due to the number (and specific combination) of observables we are using in our analysis. The shaded regions in each panel of Figure 4.1 show this error for each HOD parameter as we increase K. We choose the lowest value of K such that the constraint at this value is within one standard error of the constraint at all higher values of K. We require that this condition is met for all HOD parameters (except $\log M_0$ for the -21 sample). The optimal number of observables for each sample is indicated with a dot in each panel, and the corresponding constraining power is shown with a dashed line. For the -19 sample, the optimal number of observables is 36. For the -21 sample, the optimal number of observables is 41. Using these observables, we confirm that we can recover the truth when running chains on mocks created with different HOD parameters (i.e. different amounts of assembly bias) for each sample.

It is noteworthy that for both the -19 and -21 samples, the third observable (chosen to constrain only A_{cen} and A_{sat}) is a small bin of the average group velocity dispersion function ($\sigma_v(N)$ 3 for -19 and $\sigma_v(N)$ 1 for -21). It is also noteworthy that for both samples, the majority of the first twenty observables chosen in this analysis (16/20 or 17/20) were also chosen in the previous analysis to constrain an HOD model without assembly bias. Meanwhile, about half of the observables chosen beyond the initial twenty (8/16 or 9/21) are unique to this analysis. In particular, all of the observables chosen for the -21 beyond the first 36 are unique to this analysis. This possibly indicates that the initial observables are chosen for their ability to constrain the standard HOD parameters, while the later observables are selected for their ability to constrain the assembly bias parameters. This may also indicated that it is difficult to constrain assembly bias until the standard HOD parameters are constrained.

For the -19 sample, the unique observables chosen for this analysis include a large and small scale of $\xi(s)$, five scales of $P_N(R)$, and four large scales of $\mathrm{mcf}(s)$. For the -21 sample, the unique observables chosen for this analysis include two bins of $\sigma_v(N)$, two intermediate scales of $w_{\mathrm{p}}(r_{\mathrm{p}})$, one small scale and two large scales of $\mathrm{mcf}(s)$, one intermediate bin of $n(N)$, two large scales of $\xi(s)$, and two bins of $\mathrm{VPF}(R)$. It worth mentioning that for the -19 sample, it is difficult to accurately constrain the decorated HOD model until the parameter $\log M_0$ is constrained. This occurs by about 15 observables, particularly

Table 4.5: Optimal Observable Order. The type of clustering statistic and the bin number (1-indexing) for the observables chosen (in order) for each sample. “vHOD” refers to the observables chosen for each sample in S22. “dHOD” refers to the observables chosen in this work.

Index	-19 vHOD	-19 dHOD	-21 vHOD	-21 dHOD
1	$n_{\rm gal}$	$n_{\rm gal}$	$n_{\rm gal}$	$n_{\rm gal}$
2	$w_{\rm p}(r_{\rm p})$ 2	$w_{\rm p}(r_{\rm p})$ 2	$w_{\rm p}(r_{\rm p})$ 2	$w_{\rm p}(r_{\rm p})$ 2
3	$w_{\rm p}(r_{\rm p})$ 4	$\sigma_v(N)$ 3	$\xi(s)$ 8	$\sigma_v(N)$ **1**
4	VPF(R) 3	$\xi(s)$ **8**	$w_{\rm p}(r_{\rm p})$ 4	$\xi(s)$ 9
5	$w_{\rm p}(r_{\rm p})$ 8	$n(N)$ 3	mcf(s) 9	$\xi(s)$ 3
6	$\xi(s)$ 1	SPF(R) **1**	$w_{\rm p}(r_{\rm p})$ 1	mcf(s) 10
7	$n(N)$ 3	$w_{\rm p}(r_{\rm p})$ 3	$\xi(s)$ 9	$w_{\rm p}(r_{\rm p})$ **5**
8	$\xi(s)$ 5	$n(N)$ 2	mcf(s) 7	$n(N)$ 1
9	$n(N)$ 2	$w_{\rm p}(r_{\rm p})$ 8	$\xi(s)$ 4	$\sigma_v(N)$ 3
10	$n(N)$ 4	$\xi(s)$ 1	$\xi(s)$ 7	mcf(s) 3
11	$n(N)$ 1	$w_{\rm p}(r_{\rm p})$ 4	mcf(s) 10	$\xi(s)$ 1
12	SPF(R) 4	VPF(R) **2**	$\xi(s)$ 1	$\xi(s)$ 8
13	$\xi(s)$ 13	mcf(s) 1	$w_{\rm p}(r_{\rm p})$ 14	$\xi(s)$ 5
14	mcf(s) 14	$\xi(s)$ 10	$n(N)$ 1	$w_{\rm p}(r_{\rm p})$ 1
15	$\xi(s)$ 6	SPF(R) 2	SPF(R) 4	$n(N)$ 2
16	$n(N)$ 5	$\xi(s)$ **4**	mcf(s) 3	SPF(R) 4
17	$\xi(s)$ 2	$n(N)$ 1	$\xi(s)$ 6	mcf(s) 5
18	SPF(R) 2	$n(N)$ 5	$\sigma_v(N)$ 4	$\sigma_v(N)$ 4
19	$\xi(s)$ 10	$w_{\rm p}(r_{\rm p})$ 1	$\xi(s)$ 5	mcf(s) **14**
20	mcf(s) 2	SPF(R) 4	$\xi(s)$ 3	SPF(R) 3
21	mcf(s) 3	mcf(s) **7**	$n(N)$ 4	$w_{\rm p}(r_{\rm p})$ 3
22	$\sigma_v(N)$ 1	mcf(s) **11**	$w_{\rm p}(r_{\rm p})$ 7	$\sigma_v(N)$ **5**
23	$\sigma_v(N)$ 3	$\sigma_v(N)$ 5	$w_{\rm p}(r_{\rm p})$ 3	$\sigma_v(N)$ 2
24	$\xi(s)$ 9	SPF(R) **3**	mcf(s) 8	$\xi(s)$ 7
25	$\sigma_v(N)$ 4	$\xi(s)$ **3**	VPF(R) 3	$n(N)$ **3**
26	mcf(s) 1	$n(N)$ 4	$\xi(s)$ 2	$n(N)$ 4
27	$\sigma_v(N)$ 2	mcf(s) 2	$n(N)$ 5	$\xi(s)$ 4
28	$n(N)$ 6	$\sigma_v(N)$ 2	$n(N)$ 2	$\xi(s)$ 2
29	VPF(R) 1	VPF(R) **4**	$\xi(s)$ 11	$n(N)$ 5
30	$w_{\rm p}(r_{\rm p})$ 1	mcf(s) **8**	$\sigma_v(N)$ 3	$w_{\rm p}(r_{\rm p})$ **8**
31	$w_{\rm p}(r_{\rm p})$ 6	$w_{\rm p}(r_{\rm p})$ 6	$\sigma_v(N)$ 2	$w_{\rm p}(r_{\rm p})$ 4
32	$w_{\rm p}(r_{\rm p})$ 5	$\xi(s)$ 9	SPF(R) 2	$\xi(s)$ 6
33	$\sigma_v(N)$ 5	$n(N)$ 6	mcf(s) 5	mcf(s) 8
34	$w_{\rm p}(r_{\rm p})$ 3	mcf(s) 14	SPF(R) 3	$\xi(s)$ **10**
35	$\sigma_v(N)$ 7	VPF(R) **5**	mcf(s) 4	$\xi(s)$ 11
36	$n(N)$ 7	mcf(s) **12**	SPF(R) 1	mcf(s) 7
37	–	–	–	mcf(s) **12**
38	–	–	–	$\xi(s)$ **14**
39	–	–	–	VPF(R) **5**
40	–	–	–	VPF(R) **2**
41	–	–	–	mcf(s) **1**

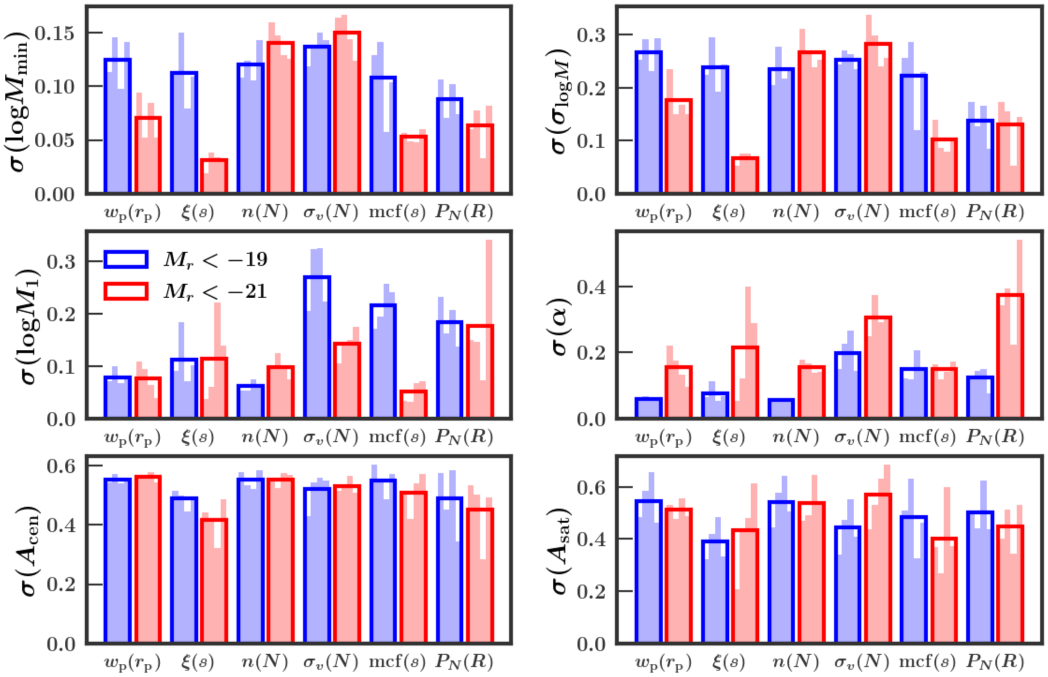

Figure 4.2: Projected constraints (1-σ) of each clustering statistic (combined with $n_{\rm gal}$) for each HOD parameter. The constraints for the -19 and -21 mocks are shown in blue and red, respectively. The height of each smaller vertical bar shows the projected constraints on one mock, while the larger open bar shows the average constraint across four mocks.

after $\xi(s)$ 1 and $w_{\rm p}(r_{\rm p})$ 4 are included. Our analyses using mock galaxy catalogs indicate that a failure to include these particular observables leads to biased constraints on $A_{\rm cen}$ and $A_{\rm sat}$. In the -21 sample, the parameter $\log M_0$ remains unconstrained. This is consistent with the results of S22, which found that constraining $\log M_0$ is important for obtaining accurate results in the -19 sample, but not in the -21 sample.

Given the results of the chains run on mocks using only $n_{\rm gal}$, $w_{\rm p}(r_{\rm p} \sim 0.3\, h^{-1}{\rm Mpc})$, and the third chosen observable for each sample, we can use importance sampling to estimate the constraining power we would achieve for each HOD parameter had we run a chain using only one clustering statistic (e.g. $w_{\rm p}(r_{\rm p})$) plus $n_{\rm gal}$. We display the results of this exercise in Figure 4.2. In each panel, the y-axis shows the projected constraint (1-σ) for a particular HOD parameter as we use different clustering statistics. The constraints for the -19 and -21 mocks are shown in blue and red, respectively. The height of each smaller vertical bar shows the projected constraints on one mock, while the larger open bar shows the average constraint across four mocks.

For the central and satellite parameters, our results are similar (though not identical) to the results from S22. For the assembly bias parameters, it is interesting to note that for both samples, no single clustering statistic provides significant constraining power for either $A_{\rm cen}$ or $A_{\rm sat}$. $\xi(s)$ seems to have the most constraining power for both $A_{\rm cen}$ and $A_{\rm sat}$, for both samples, but it performs only slightly better than the other clustering statistics. Due to the nature of importance sampling, these results should be interpreted as *estimates*, purely for visual purposes. However, this figure illustrates that while no single clustering statistic provides significant constraining power for assembly bias, the *combination* of different scales of different clustering statistics is able to produce tighter constraints on the assembly bias parameters than any one statistic.

4.7 Results

In this section we present the results from using the optimal observables we identified in the previous section to constrain the galaxy-halo connection of SDSS galaxies using a decorated HOD model with concentration-based assembly bias. The results for the -19 sample are shown in Figure 4.3, while the results for the -21 sample are shown in Figure 4.4. Dark and light blue regions depict the 1- and 2-σ regions, respectively. The best-fit parameters are listed in Table 4.6, along with their corresponding p-values (labeled “dHOD”), as well as the results from previous analyses using a standard (“vanilla”) HOD model (Sinha et al., 2018; Szewciw et al., 2022). The constraints for each parameter are listed in Table 4.7.

For the -19 sample, our best-fit results indicate strong positive central galaxy assembly bias ($A_{\rm cen}$ = 0.793), and moderate negative satellite galaxy assembly bias ($A_{\rm sat}$ = -0.368). In other words, central galaxies strongly preferentially reside in halos with higher concentrations, while satellite galaxies preferentially reside in halos with lower concentrations. This is consistent with previous results (e.g. Lange et al., 2022) which also found positive

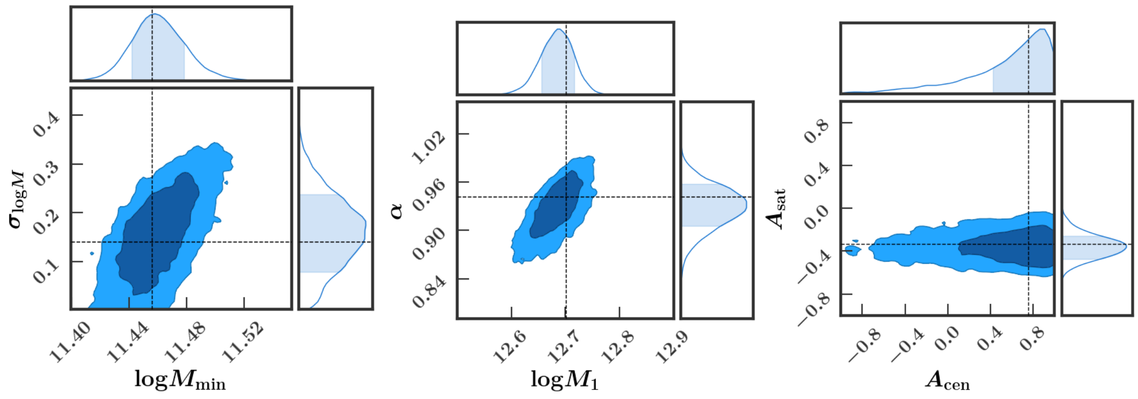

Figure 4.3: HOD parameter constraints for the SDSS -19 sample, using concentration as the secondary halo property and the “dHOD” optimal observables (listed in Table 4.5).

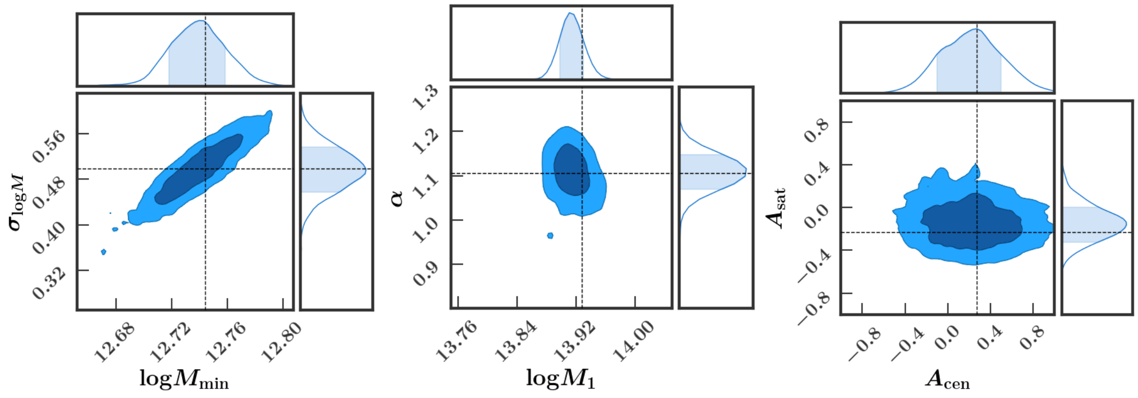

Figure 4.4: HOD parameter constraints for the SDSS -21 sample, using concentration as the secondary halo property and the 41 “dHOD” optimal observables (listed in Table 4.5). The best-fit parameters are indicated with crosshairs.

central galaxy assembly bias and negative satellite galaxy assembly bias. Additionally, this best-fit model yields a significant decrease in tension compared to the results of S22 (2.0σ compared to 4.5σ). Unfortunately, even for our optimal combination of observables, it is difficult to tightly constrain central galaxy assembly bias for this sample. This challenge is akin to the difficulty in tightly constraining $\sigma_{\log M}$ in the -19 sample; both $A_{\rm cen}$ and $\sigma_{\log M}$ result in a change in the scatter in central galaxy occupation for a given halo mass. Despite the lack of tight constraints on $A_{\rm cen}$, we are able to strongly rule out a model with zero assembly bias.

For the -21 sample, we are able to obtain slightly tighter constraints on $A_{\rm cen}$ than we

Table 4.6: Final SDSS best-fit results. Best-fit HOD parameters from all chains. We also indicate the goodness of fit of each parameter combination with a p-value.

M_r^{lim}	Model/Obs.	$\log M_{\text{min}}$	$\sigma_{\log M}$	$\log M_0$	$\log M_1$	α	A_{cen}	A_{sat}	p-val.
-19	vHOD/NWG	11.552	0.229	12.107	12.707	0.905	–	–	0.384
	vHOD/OPT36	11.445	0.099	11.651	12.703	0.958	–	–	10^{-6}
	dHOD	11.455	0.141	11.757	12.685	0.925	0.793	-0.368	0.047
-21	vHOD/NWG	12.691	0.377	12.075	13.938	1.191	–	–	0.151
	vHOD/OPT36	12.728	0.467	9.015	13.929	1.112	–	–	10^{-5}
	dHOD	12.774	0.554	9.447	13.926	1.067	-0.090	-0.240	10^{-6}

are able to obtain in the -19 sample. However, our best-fit results are consistent with zero assembly bias. Additionally, this model does not result in any decrease in tension compared to the results from S22. (In fact, the tension actually increased slightly compared to the previous analysis.) This finding is consistent with the results of Beltz-Mohrmann et al. (2020), which found assembly bias to be present in hydrodynamic simulations for lower luminosity galaxies, but not a significant source of clustering discrepancy for higher luminosity galaxies. It is thus to be expected that for the -21 sample, the addition of assembly bias parameters to the model did not result in any relief of tension.

The remaining tension found for both the -19 and -21 samples could indicate that the HOD model needs to be made even more flexible with the inclusion of spatial and velocity bias parameters (Beltz-Mohrmann et al., 2020). Additionally, these results are for a fixed cosmology sample; it is possible that a slight change in cosmological parameter values could also result in a further relief of this tension. Finally, it is possible that a different secondary halo property could be more strongly correlated with galaxy clustering, and that using a property other than concentration (like environment) could result in a better model.

4.8 Conclusions

In this work we have explored extending the standard HOD model to include parameters for assembly bias, using halo concentration as the secondary halo characteristic for modeling this assembly bias. We have identified an optimal set of observables for constraining

Table 4.7: SDSS Constraints. Marginalized constraints on SDSS for each chain. We present the median parameter values along with upper and lower limits corresponding to the 84 and 16 percentiles respectively.

M_r^{lim}	HOD Param.	vHOD/NWG	vHOD/OPT36	dHOD
-19	$\log M_{\mathrm{min}}$	$11.597^{+0.124}_{-0.055}$	$11.442^{+0.016}_{-0.015}$	$11.469^{+0.019}_{-0.017}$
	$\sigma_{\log M}$	$0.289^{+0.293}_{-0.192}$	$0.106^{+0.074}_{-0.065}$	$0.159^{+0.074}_{-0.077}$
	$\log M_0$	$10.385^{+1.519}_{-2.935}$	$11.674^{+0.089}_{-0.094}$	$11.750^{+0.093}_{-0.095}$
	$\log M_1$	$12.803^{+0.046}_{-0.058}$	$12.691^{+0.028}_{-0.029}$	$12.685^{+0.029}_{-0.031}$
	α	$0.969^{+0.028}_{-0.047}$	$0.954^{+0.019}_{-0.019}$	$0.930^{+0.025}_{-0.028}$
	A_{cen}	–	–	$0.673^{+0.245}_{-0.529}$
	A_{sat}	–	–	$-0.361^{+0.107}_{-0.103}$
-21	$\log M_{\mathrm{min}}$	$12.694^{+0.071}_{-0.058}$	$12.748^{+0.015}_{-0.015}$	$12.737^{+0.019}_{-0.020}$
	$\sigma_{\log M}$	$0.391^{+0.150}_{-0.201}$	$0.517^{+0.029}_{-0.029}$	$0.494^{+0.038}_{-0.040}$
	$\log M_0$	$9.220^{+2.136}_{-2.183}$	$9.015^{+2.017}_{-2.036}$	$8.980^{+2.019}_{-2.013}$
	$\log M_1$	$13.941^{+0.021}_{-0.024}$	$13.919^{+0.014}_{-0.014}$	$13.914^{+0.015}_{-0.015}$
	α	$1.195^{+0.051}_{-0.057}$	$1.088^{+0.031}_{-0.033}$	$1.110^{+0.035}_{-0.039}$
	A_{cen}	–	–	$0.236^{+0.290}_{-0.297}$
	A_{sat}	–	–	$-0.148^{+0.181}_{-0.156}$

this model using the algorithm presented in S22. Our best-fit results indicate the presence of strong positive central galaxy assembly bias and moderate negative satellite galaxy assembly bias for low-luminosity galaxies, with a model that does not include assembly bias significantly ruled out. This result also yields a decrease in tension compared to a previous analysis without assembly bias parameters in the model. For high-luminosity galaxies, we do not find significant evidence for assembly bias, nor do we find any significant reduction in tension by including assembly bias parameters in the model. It is possible that using a different secondary halo property to model assembly bias could yield improved results, and so further exploration is needed.

It is also possible that galaxies do not trace the spatial distribution of dark matter within halos (i.e. there is spatial bias Watson et al., 2012; Piscionere et al., 2015; Beltz-Mohrmann et al., 2020), or that they do not trace the velocity distribution of dark matter within halos (i.e. there is velocity bias Van den Bosch et al., 2005; Guo et al., 2015b,a; Beltz-Mohrmann et al., 2020). Additionally, the standard HOD model assumes that the number of satellite galaxies in each halo is governed by a Poisson distribution, but recent results indicate that this is probably not the case (Boylan-Kolchin et al., 2010; Mao et al., 2015; Jiménez et al., 2019). Finally, it is possible that a change in cosmological parameters could lead to better clustering agreement. In future work, we hope to explore all of these possibilities.

Additionally, it is possible that a change in halo definition or the removal of backsplash halos from our sample could lead to a reduction in the assembly bias signature that we find for low-luminosity galaxies (Villarreal et al., 2017; Mansfield and Kravtsov, 2020). In future work, it is worth investigating whether accounting for this possibility leads to improved agreement between our model and the observed clustering of galaxies.

CHAPTER 5

Conclusions

In order to take advantage of the power of small-scale galaxy clustering to probe and constrain cosmological models, it is crucial that we can confidently and accurately marginalize over the uncertainty of galaxy formation physics. In particular, upcoming spectroscopic surveys like the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al., 2016; Levi et al., 2019) will make unprecedentedly precise measurements of the distribution of galaxies, which will allow us to detect minute differences in clustering. Obtaining unbiased cosmological constraints from these measurements requires that we have a full understanding of the connection between galaxies and the dark matter halos in which they reside. In this dissertation I have worked to develop an accurate model of the galaxy-halo connection, through a combination of studies performed on hydrodynamic simulations and analyses of small-scale galaxy clustering in the Sloan Digital Sky Survey.

5.1 Summary

In Chapter 2, I examined the ability of halo occupation distribution (HOD) modelling to reproduce the galaxy clustering found in two different hydrodynamic simulations. I fit a simple five parameter HOD model to each simulation, and applied it to the corresponding dark matter only simulations. I then measured several galaxy clustering statistics on the galaxies from the hydrodynamic simulations and the galaxies from the HOD model. I first found that the halo mass function is shifted to lower masses in the hydrodynamic simulations, which resulted in a galaxy number density that was too high when the HOD model was applied to the dark matter only simulations. After applying a correction to the halo mass function in each simulation, I found that the HOD is able to accurately reproduce all clustering statistics for a high luminosity sample of galaxies. For a low luminosity sample, I found evidence that in addition to correcting the halo mass function, including spatial, velocity, and assembly bias parameters in the HOD is necessary to accurately reproduce clustering statistics.

In Chapter 3, I examined the impact of baryonic physics on the halo distribution in hydrodynamic simulations compared to that in dark matter only (DMO) simulations. I found that, in general, DMO simulations produce halo mass functions (HMFs) that are shifted to higher halo masses than their hydrodynamic counterparts, due to the lack of baryonic physics. However, the exact nature of this mass shift is a complex function of mass, halo definition, redshift, and larger-scale environment, and it depends on the specifics of the baryonic physics implemented in the simulation. I provided fitting formulae for the corrections one would need to apply to each DMO halo catalog in order to reproduce the HMF found in its hydrodynamic counterpart. Additionally, I explored the dependence on environment of this HMF discrepancy, and find that, in most cases, halos in low density environments are slightly more impacted by baryonic physics than halos in high density environments. Therefore, I also provided environment-dependent mass correction formulae which can reproduce the conditional, as well as global, HMF. I showed that these mass corrections also repair the large-scale clustering of halos, though the environment-dependent corrections are required to achieve an accuracy better than 2%. Finally, I examined the impact of baryonic physics on the halo mass - concentration relation, and found that its slope in hydrodynamic simulations is consistent with that in DMO simulations.

In Chapter 4, I employed a decorated HOD model that includes parameters for central and satellite galaxy assembly bias to model the clustering of SDSS galaxies. Using concentration as the secondary halo property to model assembly bias, I identified an optimal set of clustering statistics to constrain this decorated HOD model in both a high-luminosity sample and a low-luminosity sample. Ultimately, I found evidence for strong positive central galaxy assembly bias and moderate negative satellite galaxy assembly bias among low-luminosity galaxies, with zero assembly bias significantly ruled out. This model led to a significant reduction in tension compared to previous analyses that did not model assembly bias. For high-luminosity galaxies, I did not find significant evidence for assembly bias, nor did I find any significant reduction in tension by including assembly bias parameters in the model.

5.2 Future Work

In future work, it is worth investigating additional secondary halo bias properties that may contribute to the observed assembly bias signature. For example, it is possible that using halo environment would lead to better agreement between the model and the clustering of SDSS galaxies. It is also worth investigating whether misidentification of backsplash halos as host halos contributes significantly to the observed phenomenon of assembly bias.

Furthermore, due to the lingering tension found between the best-fit decorated HOD model and the clustering of SDSS galaxies in Chapter 4, it is imperative that we add even more flexibility to the HOD model in future work. We can accomplish through the addition of spatial and velocity bias parameters, which allow for the possibility that galaxies do not trace the distribution of dark matter within their host halos.

Finally, the ultimate goal of all of this work is to develop an accurate model of the galaxy-halo connection that can be implemented to constrain our cosmological model using small scales. Using small-scale galaxy clustering to constrain our cosmological model requires generating accurate, high-resolution, and large-volume predictions of large-scale structure formation while varying cosmological parameters. Running enough N-body simulations to finely sample the cosmological parameter space for this purpose is computationally infeasible. One alternative is to use the cosmological "rescaling" method presented by Angulo and White (2010), in which the output of a simulation with one cosmological model is rescaled to mimic the output of a simulation with a different cosmological model. This method involves running only a few N-body simulations with different parameters, and then rescaling them to explore the parameter space. Thus, we could explore cosmological and HOD parameter space simultaneously, and forward model small-scale galaxy clustering as a function of both cosmology and HOD parameters.

Another option is to run a *large* number of cosmological N-body simulations and build an “emulator.” In this framework, we would measure a variety of galaxy clustering statistics on each simulation as we varied HOD parameters, and then effectively interpolate between different measurements in order to estimate the clustering in any arbitrary cosmology. This method is computationally intensive, but has been used in a variety of recent studies (e.g. DeRose et al., 2019). Ultimately, given a method for accurately predicting structure formation in different cosmologies, combined with a flexible model of the galaxy-halo connection, small-scale galaxy clustering has the potential to become a powerful probe of both our cosmological model, as well as our understanding of galaxy formation and evolution.

References

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., Allende Prieto, C., An, D., Anderson, K. S. J., Anderson, S. F., Annis, J., Bahcall, N. A., and et al. (2009). The Seventh Data Release of the Sloan Digital Sky Survey. *The Astrophysical Journal Supplement*, 182:543–558.

Angulo, R. E. and White, S. D. M. (2010). One simulation to fit them all - changing the background parameters of a cosmological N-body simulation. *Monthly Notices of the Royal Astronomical Society*, 405(1):143–154.

Artale, M. C., Zehavi, I., Contreras, S., and Norberg, P. (2018). The impact of assembly bias on the halo occupation in hydrodynamical simulations. *Monthly Notices of the Royal Astronomical Society*, 480:3978–3992.

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., Günther, H. M., Lim, P. L., Crawford, S. M., Conseil, S., Shupe, D. L., Craig, M. W., Dencheva, N., Ginsburg, A., VanderPlas, J. T., Bradley, L. D., Pérez-Suárez, D., de Val-Borro, M., Aldcroft, T. L., Cruz, K. L., Robitaille, T. P., Tollerud, E. J., Ardelean, C., Babej, T., Bach, Y. P., Bachetti, M., Bakanov, A. V., Bamford, S. P., Barentsen, G., Barmby, P., Baumbach, A., Berry, K. L., Biscani, F., Boquien, M., Bostroem, K. A., Bouma, L. G., Brammer, G. B., Bray, E. M., Breytenbach, H., Buddelmeijer, H., Burke, D. J., Calderone, G., Cano Rodríguez, J. L., Cara, M., Cardoso, J. V. M., Cheedella, S., Copin, Y., Corrales, L., Crichton, D., D'Avella, D., Deil, C., Depagne, É., Dietrich, J. P., Donath, A., Droettboom, M., Earl, N., Erben, T., Fabbro, S., Ferreira, L. A., Finethy, T., Fox, R. T., Garrison, L. H., Gibbons, S. L. J., Goldstein, D. A., Gommers, R., Greco, J. P., Greenfield, P., Groener, A. M., Grollier, F., Hagen, A., Hirst, P., Homeier, D., Horton, A. J., Hosseinzadeh, G., Hu, L., Hunkeler, J. S., Ivezić, Ž., Jain, A., Jenness, T., Kanarek, G., Kendrew, S., Kern, N. S., Kerzendorf, W. E., Khvalko, A., King, J., Kirkby, D., Kulkarni, A. M., Kumar, A., Lee, A., Lenz, D., Littlefair, S. P., Ma, Z., Macleod, D. M., Mastropietro, M., McCully, C., Montagnac, S., Morris, B. M., Mueller, M., Mumford, S. J., Muna, D., Murphy, N. A., Nelson, S., Nguyen, G. H., Ninan, J. P., Nöthe, M., Ogaz, S., Oh, S., Parejko, J. K., Parley, N., Pascual, S., Patil, R., Patil, A. A., Plunkett, A. L., Prochaska, J. X., Rastogi, T., Reddy Janga, V., Sabater, J., Sakurikar, P., Seifert, M., Sherbert, L. E., Sherwood-Taylor, H., Shih, A. Y., Sick, J., Silbiger, M. T., Singanamalla, S., Singer, L. P., Sladen, P. H., Sooley, K. A., Sornarajah, S., Streicher, O., Teuben, P., Thomas, S. W., Tremblay, G. R., Turner, J. E. H., Terrón, V., van Kerkwijk, M. H., de la Vega, A., Watkins, L. L., Weaver, B. A., Whitmore, J. B., Woillez, J., Zabalza, V., and Astropy Contributors (2018). The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package. *The Astronomical Journal*, 156(3):123.

Balaguera-Antolínez, A. and Porciani, C. (2013). Counts of galaxy clusters as cosmological probes: the impact of baryonic physics. *Journal of Cosmology and Astroparticle Physics*, 2013(4):022.

Baugh, C. M., Benson, A. J., Cole, S., Frenk, C. S., and Lacey, C. G. (1999). Modelling the evolution of galaxy clustering. *Monthly Notices of the Royal Astronomical Society*, 305:L21–L25.

Behroozi, P., Hearin, A., and Moster, B. P. (2021). Observational Measures of Halo Properties Beyond Mass. *arXiv e-prints*, page arXiv:2101.05280.

Behroozi, P. S., Wechsler, R. H., and Wu, H.-Y. (2013). The ROCKSTAR Phase-space Temporal Halo Finder and the Velocity Offsets of Cluster Cores. *The Astrophysical Journal*, 762(2):109.

Beltz-Mohrmann, G. D. and Berlind, A. A. (2021). The Impact of Baryonic Physics on the Abundance, Clustering, and Concentration of Halos. *The Astrophysical Journal*, 921(2):112.

Beltz-Mohrmann, G. D., Berlind, A. A., and Szewciw, A. O. (2020). Testing the accuracy of halo occupation distribution modelling using hydrodynamic simulations. *Monthly Notices of the Royal Astronomical Society*, 491(4):5771–5788.

Benson, A. J., Cole, S., Frenk, C. S., Baugh, C. M., and Lacey, C. G. (2000). The nature of galaxy bias and clustering. *Monthly Notices of the Royal Astronomical Society*, 311:793–808.

Berlind, A. A., Frieman, J., Weinberg, D. H., Blanton, M. R., Warren, M. S., Abazajian, K., Scranton, R., Hogg, D. W., Scoccimarro, R., Bahcall, N. A., Brinkmann, J., Gott, III, J. R., Kleinman, S. J., Krzesinski, J., Lee, B. C., Miller, C. J., Nitta, A., Schneider, D. P., Tucker, D. L., Zehavi, I., and SDSS Collaboration (2006a). Percolation Galaxy Groups and Clusters in the SDSS Redshift Survey: Identification, Catalogs, and the Multiplicity Function. *The Astrophysical Journal Supplement*, 167:1–25.

Berlind, A. A., Kazin, E., Blanton, M. R., Pueblas, S., Scoccimarro, R., and Hogg, D. W. (2006b). The Clustering of Galaxy Groups: Dependence on Mass and Other Properties. *arXiv e-prints*.

Berlind, A. A. and Weinberg, D. H. (2002). The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass. *The Astrophysical Journal*, 575:587–616.

Berlind, A. A., Weinberg, D. H., Benson, A. J., Baugh, C. M., Cole, S., Davé, R., Frenk, C. S., Jenkins, A., Katz, N., and Lacey, C. G. (2003). The Halo Occupation Distribution and the Physics of Galaxy Formation. *The Astrophysical Journal*, 593:1–25.

Beutler, F., Blake, C., Colless, M., Jones, D. H., Staveley-Smith, L., Campbell, L., Parker, Q., Saunders, W., and Watson, F. (2013). The 6dF Galaxy Survey: dependence of halo occupation on stellar mass. *Monthly Notices of the Royal Astronomical Society*, 429:3604–3618.

Blanton, M. R., Schlegel, D. J., Strauss, M. A., Brinkmann, J., Finkbeiner, D., Fukugita, M., Gunn, J. E., Hogg, D. W., Ivezić, Ž., Knapp, G. R., Lupton, R. H., Munn, J. A., Schneider, D. P., Tegmark, M., and Zehavi, I. (2005). New York University Value-Added Galaxy Catalog: A Galaxy Catalog Based on New Public Surveys. *The Astronomical Journal*, 129(6):2562–2578.

Bocquet, S., Saro, A., Dolag, K., and Mohr, J. J. (2016). Halo mass function: baryon impact, fitting formulae, and implications for cluster cosmology. *Monthly Notices of the Royal Astronomical Society*, 456:2361–2373.

Bose, S., Eisenstein, D. J., Hernquist, L., Pillepich, A., Nelson, D., Marinacci, F., Springel, V., and Vogelsberger, M. (2019). Revealing the galaxy-halo connection in IllustrisTNG. *Monthly Notices of the Royal Astronomical Society*, 490(4):5693–5711.

Bose, S. and Loeb, A. (2020). Measuring the mass and concentration of dark matter halos from the velocity dispersion profile of their stars. *arXiv e-prints*, page arXiv:2010.15123.

Boylan-Kolchin, M., Springel, V., White, S. D. M., and Jenkins, A. (2010). There's no place like home? Statistics of Milky Way-mass dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 406:896–912.

Bryan, G. L. and Norman, M. L. (1998). Statistical Properties of X-Ray Clusters: Analytic and Numerical Comparisons. *The Astrophysical Journal*, 495(1):80–99.

Bullock, J. S., Kolatt, T. S., Sigad, Y., Somerville, R. S., Kravtsov, A. V., Klypin, A. A., Primack, J. R., and Dekel, A. (2001). Profiles of dark haloes: evolution, scatter and environment. *Monthly Notices of the Royal Astronomical Society*, 321(3):559–575.

Bullock, J. S., Wechsler, R. H., and Somerville, R. S. (2002). Galaxy halo occupation at high redshift. *Monthly Notices of the Royal Astronomical Society*, 329:246–256.

Castro, T., Borgani, S., Dolag, K., Marra, V., Quartin, M., Saro, A., and Sefusatti, E. (2021). On the impact of baryons on the halo mass function, bias, and cluster cosmology. *Monthly Notices of the Royal Astronomical Society*, 500(2):2316–2335.

Chaves-Montero, J., Angulo, R. E., Schaye, J., Schaller, M., Crain, R. A., Furlong, M., and Theuns, T. (2016). Subhalo abundance matching and assembly bias in the EAGLE simulation. *Monthly Notices of the Royal Astronomical Society*, 460:3100–3118.

Clowe, D., Bradač, M., Gonzalez, A. H., Markevitch, M., Randall, S. W., Jones, C., and Zaritsky, D. (2006). A Direct Empirical Proof of the Existence of Dark Matter. *The Astrophysical Journal Letters*, 648(2):L109–L113.

Colless, M., Dalton, G., Maddox, S., Sutherland, W., Norberg, P., Cole, S., Bland-Hawthorn, J., Bridges, T., Cannon, R., Collins, C., Couch, W., Cross, N., Deeley, K., De Propris, R., Driver, S. P., Efstathiou, G., Ellis, R. S., Frenk, C. S., Glazebrook, K., Jackson, C., Lahav, O., Lewis, I., Lumsden, S., Madgwick, D., Peacock, J. A., Peterson, B. A., Price, I., Seaborne, M., and Taylor, K. (2001). The 2dF Galaxy Redshift Survey:

spectra and redshifts. *Monthly Notices of the Royal Astronomical Society*, 328:1039–1063.

Collister, A. A. and Lahav, O. (2005). Distribution of red and blue galaxies in groups: an empirical test of the halo model. *Monthly Notices of the Royal Astronomical Society*, 361:415–427.

Contreras, S., Angulo, R. E., and Zennaro, M. (2021). A flexible modelling of galaxy assembly bias. *Monthly Notices of the Royal Astronomical Society*, 504(4):5205–5220.

Contreras, S., Zehavi, I., Padilla, N., Baugh, C. M., Jiménez, E., and Lacerna, I. (2019). The evolution of assembly bias. *Monthly Notices of the Royal Astronomical Society*, 484:1133–1148.

Cooray, A. and Sheth, R. (2002). Halo models of large scale structure. *Physics Reports*, 372:1–129.

Coupon, J., Arnouts, S., van Waerbeke, L., Moutard, T., Ilbert, O., van Uitert, E., Erben, T., Garilli, B., Guzzo, L., Heymans, C., Hildebrandt, H., Hoekstra, H., Kilbinger, M., Kitching, T., Mellier, Y., Miller, L., Scodeggio, M., Bonnett, C., Branchini, E., Davidzon, I., De Lucia, G., Fritz, A., Fu, L., Hudelot, P., Hudson, M. J., Kuijken, K., Leauthaud, A., Le Fèvre, O., McCracken, H. J., Moscardini, L., Rowe, B. T. P., Schrabback, T., Semboloni, E., and Velander, M. (2015). The galaxy-halo connection from a joint lensing, clustering and abundance analysis in the CFHTLenS/VIPERS field. *Monthly Notices of the Royal Astronomical Society*, 449(2):1352–1379.

Crain, R. A., Schaye, J., Bower, R. G., Furlong, M., Schaller, M., Theuns, T., Dalla Vecchia, C., Frenk, C. S., McCarthy, I. G., Helly, J. C., Jenkins, A., Rosas-Guevara, Y. M., White, S. D. M., and Trayford, J. W. (2015). The EAGLE simulations of galaxy formation: calibration of subgrid physics and model variations. *Monthly Notices of the Royal Astronomical Society*, 450:1937–1961.

Crocce, M., Pueblas, S., and Scoccimarro, R. (2006). Transients from initial conditions in cosmological simulations. *Monthly Notices of the Royal Astronomical Society*, 373(1):369–381.

Crocce, M., Pueblas, S., and Scoccimarro, R. (2012). 2LPTIC: 2nd-order Lagrangian Perturbation Theory Initial Conditions. Astrophysics Source Code Library.

Croton, D. J., Gao, L., and White, S. D. M. (2007). Halo assembly bias and its effects on galaxy clustering. *Monthly Notices of the Royal Astronomical Society*, 374:1303–1309.

Cui, W., Borgani, S., Dolag, K., Murante, G., and Tornatore, L. (2012). The effects of baryons on the halo mass function. *Monthly Notices of the Royal Astronomical Society*, 423:2279–2287.

Cui, W., Borgani, S., and Murante, G. (2014). The effect of active galactic nuclei feedback on the halo mass function. *Monthly Notices of the Royal Astronomical Society*, 441(2):1769–1782.

Davis, M., Efstathiou, G., Frenk, C. S., and White, S. D. M. (1985). The evolution of large-scale structure in a universe dominated by cold dark matter. *The Astrophysical Journal*, 292:371–394.

Dawson, K. S., Schlegel, D. J., Ahn, C. P., Anderson, S. F., Aubourg, É., Bailey, S., Barkhouser, R. H., Bautista, J. E., Beifiori, A., Berlind, A. A., Bhardwaj, V., Bizyaev, D., Blake, C. H., Blanton, M. R., Blomqvist, M., Bolton, A. S., Borde, A., Bovy, J., Brandt, W. N., Brewington, H., Brinkmann, J., Brown, P. J., Brownstein, J. R., Bundy, K., Busca, N. G., Carithers, W., Carnero, A. R., Carr, M. A., Chen, Y., Comparat, J., Connolly, N., Cope, F., Croft, R. A. C., Cuesta, A. J., da Costa, L. N., Davenport, J. R. A., Delubac, T., de Putter, R., Dhital, S., Ealet, A., Ebelke, G. L., Eisenstein, D. J., Escoffier, S., Fan, X., Filiz Ak, N., Finley, H., Font-Ribera, A., Génova-Santos, R., Gunn, J. E., Guo, H., Haggard, D., Hall, P. B., Hamilton, J.-C., Harris, B., Harris, D. W., Ho, S., Hogg, D. W., Holder, D., Honscheid, K., Huehnerhoff, J., Jordan, B., Jordan, W. P., Kauffmann, G., Kazin, E. A., Kirkby, D., Klaene, M. A., Kneib, J.-P., Le Goff, J.-M., Lee, K.-G., Long, D. C., Loomis, C. P., Lundgren, B., Lupton, R. H., Maia, M. A. G., Makler, M., Malanushenko, E., Malanushenko, V., Mandelbaum, R., Manera, M., Maraston, C., Margala, D., Masters, K. L., McBride, C. K., McDonald, P., McGreer, I. D., McMahon, R. G., Mena, O., Miralda-Escudé, J., Montero-Dorta, A. D., Montesano, F., Muna, D., Myers, A. D., Naugle, T., Nichol, R. C., Noterdaeme, P., Nuza, S. E., Olmstead, M. D., Oravetz, A., Oravetz, D. J., Owen, R., Padmanabhan, N., Palanque-Delabrouille, N., Pan, K., Parejko, J. K., Pâris, I., Percival, W. J., Pérez-Fournon, I., Pérez-Ràfols, I., Petitjean, P., Pfaffenberger, R., Pforr, J., Pieri, M. M., Prada, F., Price-Whelan, A. M., Raddick, M. J., Rebolo, R., Rich, J., Richards, G. T., Rockosi, C. M., Roe, N. A., Ross, A. J., Ross, N. P., Rossi, G., Rubiño-Martin, J. A., Samushia, L., Sánchez, A. G., Sayres, C., Schmidt, S. J., Schneider, D. P., Scóccola, C. G., Seo, H.-J., Shelden, A., Sheldon, E., Shen, Y., Shu, Y., Slosar, A., Smee, S. A., Snedden, S. A., Stauffer, F., Steele, O., Strauss, M. A., Streblyanska, A., Suzuki, N., Swanson, M. E. C., Tal, T., Tanaka, M., Thomas, D., Tinker, J. L., Tojeiro, R., Tremonti, C. A., Vargas Magaña, M., Verde, L., Viel, M., Wake, D. A., Watson, M., Weaver, B. A., Weinberg, D. H., Weiner, B. J., West, A. A., White, M., Wood-Vasey, W. M., Yeche, C., Zehavi, I., Zhao, G.-B., and Zheng, Z. (2013). The Baryon Oscillation Spectroscopic Survey of SDSS-III. *The Astronomical Journal*, 145:10.

DeRose, J., Wechsler, R. H., Tinker, J. L., Becker, M. R., Mao, Y.-Y., McClintock, T., McLaughlin, S., Rozo, E., and Zhai, Z. (2019). The AEMULUS Project. I. Numerical Simulations for Precision Cosmology. *The Astrophysical Journal*, 875(1):69.

DESI Collaboration, Aghamousa, A., Aguilar, J., Ahlen, S., Alam, S., Allen, L. E., Allende Prieto, C., Annis, J., Bailey, S., Balland, C., and et al. (2016). The DESI Experiment Part I: Science,Targeting, and Survey Design. *arXiv e-prints*.

Desmond, H., Mao, Y.-Y., Wechsler, R. H., Crain, R. A., and Schaye, J. (2017). On the galaxy-halo connection in the EAGLE simulation. *Monthly Notices of the Royal Astronomical Society*, 471:L11–L15.

Despali, G. and Vegetti, S. (2017). The impact of baryonic physics on the subhalo mass function and implications for gravitational lensing. *Monthly Notices of the Royal Astronomical Society*, 469(2):1997–2010.

Dicke, R. H., Peebles, P. J. E., Roll, P. G., and Wilkinson, D. T. (1965). Cosmic Black-Body Radiation. *The Astrophysical Journal*, 142:414–419.

Eckert, K. D., Kannappan, S. J., Lagos, C. d. P., Baker, A. D., Berlind, A. A., Stark, D. V., Moffett, A. J., Nasipak, Z., and Norris, M. A. (2017). The Baryonic Collapse Efficiency of Galaxy Groups in the RESOLVE and ECO Surveys. *The Astrophysical Journal*, 849(1):20.

Eisenstein, D. J., Zehavi, I., Hogg, D. W., Scoccimarro, R., Blanton, M. R., Nichol, R. C., Scranton, R., Seo, H.-J., Tegmark, M., Zheng, Z., Anderson, S. F., Annis, J., Bahcall, N., Brinkmann, J., Burles, S., Castander, F. J., Connolly, A., Csabai, I., Doi, M., Fukugita, M., Frieman, J. A., Glazebrook, K., Gunn, J. E., Hendry, J. S., Hennessy, G., Ivezić, Z., Kent, S., Knapp, G. R., Lin, H., Loh, Y.-S., Lupton, R. H., Margon, B., McKay, T. A., Meiksin, A., Munn, J. A., Pope, A., Richmond, M. W., Schlegel, D., Schneider, D. P., Shimasaku, K., Stoughton, C., Strauss, M. A., SubbaRao, M., Szalay, A. S., Szapudi, I., Tucker, D. L., Yanny, B., and York, D. G. (2005). Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies. *The Astrophysical Journal*, 633(2):560–574.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. (2013). emcee: The MCMC Hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306.

Gao, F. and Han, L. (2012). Implementing the nelder-mead simplex algorithm with adaptive parameters. *Computational Optimization and Applications*, 51:259–277.

Gao, L., Springel, V., and White, S. D. M. (2005). The age dependence of halo clustering. *Monthly Notices of the Royal Astronomical Society*, 363:L66–L70.

Genel, S., Vogelsberger, M., Springel, V., Sijacki, D., Nelson, D., Snyder, G., Rodriguez-Gomez, V., Torrey, P., and Hernquist, L. (2014). Introducing the Illustris project: the evolution of galaxy populations across cosmic time. *Monthly Notices of the Royal Astronomical Society*, 445:175–200.

Gonzalez, A. H., Sivanandam, S., Zabludoff, A. I., and Zaritsky, D. (2013). Galaxy Cluster Baryon Fractions Revisited. *The Astrophysical Journal*, 778(1):14.

Guo, H., Zheng, Z., Behroozi, P. S., Zehavi, I., Chuang, C.-H., Comparat, J., Favole, G., Gottloeber, S., Klypin, A., Prada, F., Rodríguez-Torres, S. A., Weinberg, D. H., and Yepes, G. (2016). Modelling galaxy clustering: halo occupation distribution versus subhalo matching. *Monthly Notices of the Royal Astronomical Society*, 459:3040–3058.

Guo, H., Zheng, Z., Zehavi, I., Behroozi, P. S., Chuang, C.-H., Comparat, J., Favole, G., Gottloeber, S., Klypin, A., Prada, F., Weinberg, D. H., and Yepes, G. (2015a). Redshift-space clustering of SDSS galaxies - luminosity dependence, halo occupation distribution, and velocity bias. *Monthly Notices of the Royal Astronomical Society*, 453:4368–4383.

Guo, H., Zheng, Z., Zehavi, I., Dawson, K., Skibba, R. A., Tinker, J. L., Weinberg, D. H., White, M., and Schneider, D. P. (2015b). Velocity bias from the small-scale clustering of SDSS-III BOSS galaxies. *Monthly Notices of the Royal Astronomical Society*, 446:578–594.

Hadzhiyska, B., Bose, S., Eisenstein, D., and Hernquist, L. (2021a). Extensions to models of the galaxy-halo connection. *Monthly Notices of the Royal Astronomical Society*, 501(2):1603–1620.

Hadzhiyska, B., Bose, S., Eisenstein, D., Hernquist, L., and Spergel, D. N. (2020). Limitations to the 'basic' HOD model and beyond. *Monthly Notices of the Royal Astronomical Society*, 493(4):5506–5519.

Hadzhiyska, B., Liu, S., Somerville, R. S., Gabrielpillai, A., Bose, S., Eisenstein, D., and Hernquist, L. (2021b). Galaxy assembly bias and large-scale distribution: a comparison between IllustrisTNG and a semi-analytic model. *Monthly Notices of the Royal Astronomical Society*, 508(1):698–718.

Hadzhiyska, B., Tacchella, S., Bose, S., and Eisenstein, D. J. (2021c). The galaxy-halo connection of emission-line galaxies in IllustrisTNG. *Monthly Notices of the Royal Astronomical Society*, 502(3):3599–3617.

Hamana, T., Ouchi, M., Shimasaku, K., Kayo, I., and Suto, Y. (2004). Properties of host haloes of Lyman-break galaxies and Lyman α emitters from their number densities and angular clustering. *Monthly Notices of the Royal Astronomical Society*, 347:813–823.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del R'ıo, J. F., Wiebe, M., Peterson, P., G'erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hearin, A. P., Zentner, A. R., van den Bosch, F. C., Campbell, D., and Tollerud, E. (2016). Introducing decorated HODs: modelling assembly bias in the galaxy-halo connection. *Monthly Notices of the Royal Astronomical Society*, 460:2552–2570.

Hinton, S. R. (2016). ChainConsumer. *The Journal of Open Source Software*, 1(4):00045.

Hubble, E. (1929). A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *Proceedings of the National Academy of Science*, 15(3):168–173.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95.

Jiménez, E., Contreras, S., Padilla, N., Zehavi, I., Baugh, C. M., and Gonzalez-Perez, V. (2019). Extensions to the halo occupation distribution model for more accurate clustering predictions. *Monthly Notices of the Royal Astronomical Society*, 490(3):3532–3544.

Jing, Y. P., Mo, H. J., and Börner, G. (1998). Spatial Correlation Function and Pairwise Velocity Dispersion of Galaxies: Cold Dark Matter Models versus the Las Campanas Survey. *The Astrophysical Journal*, 494:1–12.

Jones, D. H., Saunders, W., Colless, M., Read, M. A., Parker, Q. A., Watson, F. G., Campbell, L. A., Burkey, D., Mauch, T., Moore, L., Hartley, M., Cass, P., James, D., Russell, K., Fiegert, K., Dawe, J., Huchra, J., Jarrett, T., Lahav, O., Lucey, J., Mamon, G. A., Proust, D., Sadler, E. M., and Wakamatsu, K.-i. (2004). The 6dF Galaxy Survey: samples, observational techniques and the first data release. *Monthly Notices of the Royal Astronomical Society*, 355:747–763.

Jones, E., Oliphant, T., Peterson, P., and Others (2001). SciPy: Open source scientific tools for python.

Jose, C., Subramanian, K., Srianand, R., and Samui, S. (2013). Spatial clustering of high-redshift Lyman-break galaxies. *Monthly Notices of the Royal Astronomical Society*, 429:2333–2350.

Kauffmann, G., Colberg, J. M., Diaferio, A., and White, S. D. M. (1999). Clustering of galaxies in a hierarchical universe - I. Methods and results at z=0. *Monthly Notices of the Royal Astronomical Society*, 303:188–206.

Kauffmann, G., Nusser, A., and Steinmetz, M. (1997). Galaxy formation and large-scale bias. *Monthly Notices of the Royal Astronomical Society*, 286:795–811.

Khandai, N., Di Matteo, T., Croft, R., Wilkins, S., Feng, Y., Tucker, E., DeGraf, C., and Liu, M.-S. (2015). The MassiveBlack-II simulation: the evolution of haloes and galaxies to z $\sim$ 0. *Monthly Notices of the Royal Astronomical Society*, 450(2):1349–1374.

Kim, J.-W., Edge, A. C., Wake, D. A., Gonzalez-Perez, V., Baugh, C. M., Lacey, C. G., Yamada, T., Sato, Y., Burgett, W. S., Chambers, K. C., Price, P. A., Foucaud, S., Draper, P., and Kaiser, N. (2014). Clustering of extremely red objects in Elais-N1 from the UKIDSS DXS with optical photometry from Pan-STARRS 1 and Subaru. *Monthly Notices of the Royal Astronomical Society*, 438:825–840.

Klypin, A. A., Trujillo-Gomez, S., and Primack, J. (2011). Dark Matter Halos in the Standard Cosmological Model: Results from the Bolshoi Simulation. *The Astrophysical Journal*, 740(2):102.

Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., Klypin, A. A., Gottlöber, S., Allgood, B., and Primack, J. R. (2004). The Dark Side of the Halo Occupation Distribution. *The Astrophysical Journal*, 609:35–49.

Lacey, C. and Cole, S. (1994). Merger Rates in Hierarchical Models of Galaxy Formation - Part Two - Comparison with N-Body Simulations. *Monthly Notices of the Royal Astronomical Society*, 271:676.

Lange, J. U., Hearin, A. P., Leauthaud, A., van den Bosch, F. C., Guo, H., and DeRose, J. (2022). Five per cent measurements of the growth rate from simulation-based modelling of redshift-space clustering in BOSS LOWZ. *Monthly Notices of the Royal Astronomical Society*, 509(2):1779–1804.

Lange, J. U., van den Bosch, F. C., Zentner, A. R., Wang, K., Hearin, A. P., and Guo, H. (2019a). Cosmological Evidence Modelling: a new simulation-based approach to constrain cosmology on non-linear scales. *Monthly Notices of the Royal Astronomical Society*, 490(2):1870–1878.

Lange, J. U., Yang, X., Guo, H., Luo, W., and van den Bosch, F. C. (2019b). New perspectives on the BOSS small-scale lensing discrepancy for the Planck ΛCDM cosmology. *Monthly Notices of the Royal Astronomical Society*, 488(4):5771–5787.

Leauthaud, A., Tinker, J., Bundy, K., Behroozi, P. S., Massey, R., Rhodes, J., George, M. R., Kneib, J.-P., Benson, A., Wechsler, R. H., Busha, M. T., Capak, P., Cortês, M., Ilbert, O., Koekemoer, A. M., Le Fèvre, O., Lilly, S., McCracken, H. J., Salvato, M., Schrabback, T., Scoville, N., Smith, T., and Taylor, J. E. (2012). New Constraints on the Evolution of the Stellar-to-dark Matter Connection: A Combined Analysis of Galaxy-Galaxy Lensing, Clustering, and Stellar Mass Functions from z = 0.2 to z =1. *The Astrophysical Journal*, 744(2):159.

Lee, K.-S., Giavalisco, M., Gnedin, O. Y., Somerville, R. S., Ferguson, H. C., Dickinson, M., and Ouchi, M. (2006). The Large-Scale and Small-Scale Clustering of Lyman Break Galaxies at 3.5 ¡ z ¡ 5.5 from the GOODS Survey. *The Astrophysical Journal*, 642:63–80.

Levi, M., Allen, L. E., Raichoor, A., Baltay, C., BenZvi, S., Beutler, F., Bolton, A., Castander, F. J., Chuang, C.-H., Cooper, A., Cuby, J.-G., Dey, A., Eisenstein, D., Fan, X., Flaugher, B., Frenk, C., Gonzalez-Morales, A. X., Graur, O., Guy, J., Habib, S., Honscheid, K., Juneau, S., Kneib, J.-P., Lahav, O., Lang, D., Leauthaud, A., Lusso, B., de la Macorra, A., Manera, M., Martini, P., Mao, S., Newman, J. A., Palanque-Delabrouille, N., Percival, W. J., Allende Prieto, C., Rockosi, C. M., Ruhlmann-Kleider, V., Schlegel, D., Seo, H.-J., Song, Y.-S., Tarle, G., Wechsler, R., Weinberg, D., Yeche, C., and Zu, Y. (2019). The Dark Energy Spectroscopic Instrument (DESI). In *Bulletin of the American Astronomical Society*, volume 51, page 57.

Ma, C.-P. and Fry, J. N. (2000). Deriving the Nonlinear Cosmological Power Spectrum and Bispectrum from Analytic Dark Matter Halo Profiles and Mass Functions. *The Astrophysical Journal*, 543:503–513.

Magliocchetti, M. and Porciani, C. (2003). The halo distribution of 2dF galaxies. *Monthly Notices of the Royal Astronomical Society*, 346:186–198.

Mansfield, P. and Kravtsov, A. V. (2020). The three causes of low-mass assembly bias. *Monthly Notices of the Royal Astronomical Society*, 493(4):4763–4782.

Mao, Y.-Y., Williamson, M., and Wechsler, R. H. (2015). The Dependence of Subhalo Abundance on Halo Concentration. *The Astrophysical Journal*, 810:21.

Mao, Y.-Y., Zentner, A. R., and Wechsler, R. H. (2018). Beyond assembly bias: exploring secondary halo biases for cluster-size haloes. *Monthly Notices of the Royal Astronomical Society*, 474(4):5143–5157.

Marinacci, F., Vogelsberger, M., Pakmor, R., Torrey, P., Springel, V., Hernquist, L., Nelson, D., Weinberger, R., Pillepich, A., Naiman, J., and Genel, S. (2018). First results from the IllustrisTNG simulations: radio haloes and magnetic fields. *Monthly Notices of the Royal Astronomical Society*, 480:5113–5139.

McAlpine, S., Helly, J. C., Schaller, M., Trayford, J. W., Qu, Y., Furlong, M., Bower, R. G., Crain, R. A., Schaye, J., Theuns, T., Dalla Vecchia, C., Frenk, C. S., McCarthy, I. G., Jenkins, A., Rosas-Guevara, Y., White, S. D. M., Baes, M., Camps, P., and Lemson, G. (2016). The EAGLE simulations of galaxy formation: Public release of halo and galaxy catalogues. *Astronomy and Computing*, 15:72–89.

McBride, C., Berlind, A., Scoccimarro, R., Wechsler, R., Busha, M., Gardner, J., and van den Bosch, F. (2009). LasDamas Mock Galaxy Catalogs for SDSS. In *American Astronomical Society Meeting Abstracts #213*, volume 41 of *Bulletin of the American Astronomical Society*, page 253.

McCarthy, K. S., Zheng, Z., and Guo, H. (2019). The effects of galaxy assembly bias on the inference of growth rate from redshift-space distortions. *Monthly Notices of the Royal Astronomical Society*, 487(2):2424–2440.

McCarthy, K. S., Zheng, Z., Guo, H., Luo, W., and Lin, Y.-T. (2022). On the constraints of galaxy assembly bias in velocity space. *Monthly Notices of the Royal Astronomical Society*, 509(1):380–394.

McClelland, J. and Silk, J. (1977). The correlation function for density perturbations in an expanding universe. II - Nonlinear theory. *The Astrophysical Journal*, 217:331–352.

McCullagh, N., Norberg, P., Cole, S., Gonzalez-Perez, V., Baugh, C., and Helly, J. (2017). Revisiting HOD model assumptions: the impact of AGN feedback and assembly bias. *arXiv e-prints*.

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

McKinney, W. (2011). pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14.

Moffett, A. J., Kannappan, S. J., Berlind, A. A., Eckert, K. D., Stark, D. V., Hendel, D., Norris, M. A., and Grogin, N. A. (2015). ECO and RESOLVE: Galaxy Disk Growth in Environmental Context. *The Astrophysical Journal*, 812(2):89.

Moustakas, L. A. and Somerville, R. S. (2002). The Masses, Ancestors, and Descendants of Extremely Red Objects: Constraints from Spatial Clustering. *The Astrophysical Journal*, 577:1–10.

Naiman, J. P., Pillepich, A., Springel, V., Ramirez-Ruiz, E., Torrey, P., Vogelsberger, M., Pakmor, R., Nelson, D., Marinacci, F., Hernquist, L., Weinberger, R., and Genel, S. (2018). First results from the IllustrisTNG simulations: a tale of two elements - chemical evolution of magnesium and europium. *Monthly Notices of the Royal Astronomical Society*, 477:1206–1224.

Navarro, J. F., Frenk, C. S., and White, S. D. M. (1996). The Structure of Cold Dark Matter Halos. *The Astrophysical Journal*, 462:563.

Navarro, J. F., Frenk, C. S., and White, S. D. M. (1997). A Universal Density Profile from Hierarchical Clustering. *The Astrophysical Journal*, 490:493–508.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7:308–313.

Nelson, D., Pillepich, A., Genel, S., Vogelsberger, M., Springel, V., Torrey, P., Rodriguez-Gomez, V., Sijacki, D., Snyder, G. F., Griffen, B., Marinacci, F., Blecha, L., Sales, L., Xu, D., and Hernquist, L. (2015). The illustris simulation: Public data release. *Astronomy and Computing*, 13:12–37.

Nelson, D., Pillepich, A., Springel, V., Weinberger, R., Hernquist, L., Pakmor, R., Genel, S., Torrey, P., Vogelsberger, M., Kauffmann, G., Marinacci, F., and Naiman, J. (2018). First results from the IllustrisTNG simulations: the galaxy colour bimodality. *Monthly Notices of the Royal Astronomical Society*, 475:624–647.

Neyman, J. and Scott, E. L. (1952). A Theory of the Spatial Distribution of Galaxies. *The Astrophysical Journal*, 116:144.

Norberg, P., Baugh, C. M., Gaztañaga, E., and Croton, D. J. (2009). Statistical analysis of galaxy surveys - I. Robust error estimation for two-point clustering statistics. *Monthly Notices of the Royal Astronomical Society*, 396(1):19–38.

Padilla, N., Contreras, S., Zehavi, I., Baugh, C. M., and Norberg, P. (2019). The effect of assembly bias on redshift-space distortions. *Monthly Notices of the Royal Astronomical Society*, 486(1):582–595.

Parejko, J. K., Sunayama, T., Padmanabhan, N., Wake, D. A., Berlind, A. A., Bizyaev, D., Blanton, M., Bolton, A. S., van den Bosch, F., Brinkmann, J., Brownstein, J. R., da Costa, L. A. N., Eisenstein, D. J., Guo, H., Kazin, E., Maia, M., Malanushenko, E., Maraston, C., McBride, C. K., Nichol, R. C., Oravetz, D. J., Pan, K., Percival, W. J., Prada, F., Ross, A. J., Ross, N. P., Schlegel, D. J., Schneider, D., Simmons, A. E., Skibba, R., Tinker, J., Tojeiro, R., Weaver, B. A., Wetzel, A., White, M., Weinberg, D. H., Thomas, D., Zehavi, I., and Zheng, Z. (2013). The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: the low-redshift sample. *Monthly Notices of the Royal Astronomical Society*, 429(1):98–112.

Peacock, J. A. and Smith, R. E. (2000). Halo occupation numbers and galaxy bias. *Monthly Notices of the Royal Astronomical Society*, 318:1144–1156.

Peebles, P. J. E. (1974). The Nature of the Distribution of Galaxies. *Astronomy & Astrophysics*, 32:197.

Penzias, A. A. and Wilson, R. W. (1965). A Measurement of Excess Antenna Temperature at 4080 Mc/s. *The Astrophysical Journal*, 142:419–421.

Pérez, F. and Granger, B. E. (2007). IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29.

Perlmutter, S., Aldering, G., Goldhaber, G., Knop, R. A., Nugent, P., Castro, P. G., Deustua, S., Fabbro, S., Goobar, A., Groom, D. E., Hook, I. M., Kim, A. G., Kim, M. Y., Lee, J. C., Nunes, N. J., Pain, R., Pennypacker, C. R., Quimby, R., Lidman, C., Ellis, R. S., Irwin, M., McMahon, R. G., Ruiz-Lapuente, P., Walton, N., Schaefer, B., Boyle, B. J., Filippenko, A. V., Matheson, T., Fruchter, A. S., Panagia, N., Newberg, H. J. M., Couch, W. J., and Project, T. S. C. (1999). Measurements of Ω and Λ from 42 High-Redshift Supernovae. *The Astrophysical Journal*, 517(2):565–586.

Pillepich, A., Nelson, D., Hernquist, L., Springel, V., Pakmor, R., Torrey, P., Weinberger, R., Genel, S., Naiman, J. P., Marinacci, F., and Vogelsberger, M. (2018a). First results from the IllustrisTNG simulations: the stellar mass content of groups and clusters of galaxies. *Monthly Notices of the Royal Astronomical Society*, 475:648–675.

Pillepich, A., Springel, V., Nelson, D., Genel, S., Naiman, J., Pakmor, R., Hernquist, L., Torrey, P., Vogelsberger, M., Weinberger, R., and Marinacci, F. (2018b). Simulating galaxy formation with the IllustrisTNG model. *Monthly Notices of the Royal Astronomical Society*, 473(3):4077–4106.

Piscionere, J. A., Berlind, A. A., McBride, C. K., and Scoccimarro, R. (2015). The Spatial Distribution of Satellite Galaxies within Halos: Measuring the Very Small Scale Angular Clustering of SDSS Galaxies. *The Astrophysical Journal*, 806:125.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Armitage-Caplan, C., Arnaud, M., Ashdown, M., Atrio-Barand ela, F., Aumont, J., Baccigalupi, C., Banday, A. J., Barreiro, R. B., Bartlett, J. G., Battaner, E., Benabed, K., Benoît, A., Benoit-Lévy, A., Bernard, J. P., Bersanelli, M., Bielewicz, P., Bobin, J., Bock, J. J., Bonaldi, A., Bond, J. R., Borrill, J., Bouchet, F. R., Bridges, M., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cappellini, B., Cardoso, J. F., Catalano, A., Challinor, A., Chamballu, A., Chary, R. R., Chen, X., Chiang, H. C., Chiang, L. Y., Christensen, P. R., Church, S., Clements, D. L., Colombi, S., Colombo, L. P. L., Couchot, F., Coulais, A., Crill, B. P., Curto, A., Cuttaia, F., Danese, L., Davies, R. D., Davis, R. J., de Bernardis, P., de Rosa, A., de Zotti, G., Delabrouille, J., Delouis, J. M., Désert, F. X., Dickinson, C., Diego, J. M., Dolag, K., Dole, H., Donzelli, S., Doré, O., Douspis, M., Dunkley, J., Dupac, X., Efstathiou, G., Elsner, F., Enßlin, T. A., Eriksen, H. K., Finelli, F., Forni, O., Frailis, M., Fraisse, A. A., Franceschi, E., Gaier, T. C., Galeotta, S., Galli, S., Ganga, K., Giard, M., Giardino, G., Giraud-Héraud, Y., Gjerløw, E., González-Nuevo, J., Górski, K. M., Gratton, S., Gregorio, A., Gruppuso, A., Gudmundsson, J. E., Haissinski, J., Hamann, J., Hansen, F. K., Hanson, D., Harrison, D., Henrot-Versillé, S., Hernández-Monteagudo, C., Herranz, D., Hildebrand t, S. R., Hivon, E., Hobson, M., Holmes, W. A., Hornstrup, A., Hou, Z., Hovest, W., Huffenberger, K. M., Jaffe, A. H., Jaffe, T. R., Jewell, J., Jones, W. C., Juvela, M., Keihänen, E., Keskitalo, R., Kisner, T. S., Kneissl, R., Knoche, J., Knox, L., Kunz, M., Kurki-Suonio, H., Lagache, G., Lähteenmäki, A., Lamarre, J. M., Lasenby, A., Lattanzi, M., Laureijs, R. J., Lawrence, C. R., Leach, S., Leahy, J. P., Leonardi, R., León-Tavares, J., Lesgourgues, J., Lewis, A., Liguori, M., Lilje, P. B., Linden-Vørnle, M., López-Caniego, M., Lubin, P. M., Macías-Pérez, J. F., Maffei, B., Maino, D., Mand olesi, N., Maris, M., Marshall, D. J., Martin, P. G., Martínez-González, E., Masi, S., Massardi, M., Matarrese, S., Matthai, F., Mazzotta, P., Meinhold, P. R., Melchiorri, A., Melin, J. B., Mendes, L., Menegoni, E., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschênes, M. A., Moneti, A., Montier, L., Morgante, G., Mortlock, D., Moss, A., Munshi, D., Murphy, J. A., Naselsky, P., Nati, F., Natoli, P., Netterfield, C. B., Nørgaard-Nielsen, H. U., Noviello, F., Novikov, D., Novikov, I., O'Dwyer, I. J., Osborne, S., Oxborrow, C. A., Paci, F., Pagano, L., Pajot, F., Paladini, R., Paoletti, D., Partridge, B., Pasian, F., Patanchon, G., Pearson, D., Pearson, T. J., Peiris, H. V., Perdereau, O., Perotto, L., Perrotta, F., Pettorino, V., Piacentini, F., Piat, M., Pierpaoli, E., Pietrobon, D., Plaszczynski, S., Platania, P., Pointecouteau, E., Polenta, G., Ponthieu, N., Popa, L., Poutanen, T., Pratt, G. W., Prézeau, G., Prunet, S., Puget, J. L., Rachen, J. P., Reach, W. T., Rebolo, R., Reinecke, M., Remazeilles, M., Renault, C., Ricciardi, S., Riller, T., Ristorcelli, I., Rocha, G., Rosset, C., Roudier, G., Rowan-Robinson, M., Rubiño-Martín, J. A., Rusholme, B., Sandri, M., Santos, D., Savelainen, M., Savini, G., Scott, D., Seiffert, M. D., Shellard, E. P. S., Spencer, L. D., Starck, J. L., Stolyarov, V., Stompor, R., Sudiwala, R., Sunyaev, R., Sureau, F., Sutton, D., Suur-Uski, A. S., Sygnet, J. F., Tauber, J. A., Tavagnacco, D., Terenzi, L., Toffolatti, L., Tomasi, M., Tristram, M., Tucci, M., Tuovinen, J., Türler, M., Umana, G., Valenziano, L., Valiviita, J., Van Tent, B., Vielva, P., Villa, F., Vittorio, N., Wade, L. A., Wandelt, B. D., Wehus, I. K., White, M., White, S. D. M., Wilkinson, A., Yvon, D., Zacchei, A., and Zonca, A. (2014). Planck 2013 results. XVI. Cosmological parameters. *Astronomy & Astrophysics*, 571:A16.

Planck Collaboration, Aghanim, N., Akrami, Y., Ashdown, M., Aumont, J., Baccigalupi, C., Ballardini, M., Banday, A. J., Barreiro, R. B., Bartolo, N., Basak, S., Battye, R., Benabed, K., Bernard, J. P., Bersanelli, M., Bielewicz, P., Bock, J. J., Bond, J. R., Borrill, J., Bouchet, F. R., Boulanger, F., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cardoso, J. F., Carron, J., Challinor, A., Chiang, H. C., Chluba, J., Colombo, L. P. L., Combet, C., Contreras, D., Crill, B. P., Cuttaia, F., de Bernardis, P., de Zotti, G., Delabrouille, J., Delouis, J. M., Di Valentino, E., Diego, J. M., Doré, O., Douspis, M., Ducout, A., Dupac, X., Dusini, S., Efstathiou, G., Elsner, F., Enßlin, T. A., Eriksen, H. K., Fantaye, Y., Farhang, M., Fergusson, J., Fernandez-Cobos, R., Finelli, F., Forastieri, F., Frailis, M., Fraisse, A. A., Franceschi, E., Frolov, A., Galeotta, S., Galli, S., Ganga, K., Génova-Santos, R. T., Gerbino, M., Ghosh, T., González-Nuevo, J., Górski, K. M., Gratton, S., Gruppuso, A., Gudmundsson, J. E., Hamann, J., Handley, W., Hansen, F. K., Herranz, D., Hildebrandt, S. R., Hivon, E., Huang, Z., Jaffe, A. H., Jones, W. C., Karakci, A., Keihänen, E., Keskitalo, R., Kiiveri, K., Kim, J., Kisner, T. S., Knox, L., Krachmalnicoff, N., Kunz, M., Kurki-Suonio, H., Lagache, G., Lamarre, J. M., Lasenby, A., Lattanzi, M., Lawrence, C. R., Le Jeune, M., Lemos, P., Lesgourgues, J., Levrier, F., Lewis, A., Liguori, M., Lilje, P. B., Lilley, M., Lindholm, V., López-Caniego, M., Lubin, P. M., Ma, Y. Z., Macías-Pérez, J. F., Maggio, G., Maino, D., Mandolesi, N., Mangilli, A., Marcos-Caballero, A., Maris, M., Martin, P. G., Martinelli, M., Martínez-González, E., Matarrese, S., Mauri, N., McEwen, J. D., Meinhold, P. R., Melchiorri, A., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschênes, M. A., Molinari, D., Montier, L., Morgante, G., Moss, A., Natoli, P., Nørgaard-Nielsen, H. U., Pagano, L., Paoletti, D., Partridge, B., Patanchon, G., Peiris, H. V., Perrotta, F., Pettorino, V., Piacentini, F., Polastri, L., Polenta, G., Puget, J. L., Rachen, J. P., Reinecke, M., Remazeilles, M., Renzi, A., Rocha, G., Rosset, C., Roudier, G., Rubiño-Martín, J. A., Ruiz-Granados, B., Salvati, L., Sandri, M., Savelainen, M., Scott, D., Shellard, E. P. S., Sirignano, C., Sirri, G., Spencer, L. D., Sunyaev, R., Suur-Uski, A. S., Tauber, J. A., Tavagnacco, D., Tenti, M., Toffolatti, L., Tomasi, M., Trombetti, T., Valenziano, L., Valiviita, J., Van Tent, B., Vibert, L., Vielva, P., Villa, F., Vittorio, N., Wandelt, B. D., Wehus, I. K., White, M., White, S. D. M., Zacchei, A., and Zonca, A. (2020). Planck 2018 results. VI. Cosmological parameters. *Astronomy & Astrophysics*, 641:A6.

Pujol, A. and Gaztañaga, E. (2014). Are the halo occupation predictions consistent with large-scale galaxy clustering? *Monthly Notices of the Royal Astronomical Society*, 442:1930–1941.

Pujol, A., Hoffmann, K., Jiménez, N., and Gaztañaga, E. (2017). What determines large scale galaxy clustering: halo mass or local density? *Astronomy & Astrophsics*, 598:A103.

Ragagnin, A., Dolag, K., Moscardini, L., Biviano, A., and D'Onofrio, M. (2019). Dependency of halo concentration on mass, redshift and fossilness in Magneticum hydrodynamic simulations. *Monthly Notices of the Royal Astronomical Society*, 486(3):4001–4012.

Ragagnin, A., Saro, A., Singh, P., and Dolag, K. (2021). Cosmology dependence of halo masses and concentrations in hydrodynamic simulations. *Monthly Notices of the Royal Astronomical Society*, 500(4):5056–5071.

Riess, A. G., Filippenko, A. V., Challis, P., Clocchiatti, A., Diercks, A., Garnavich, P. M., Gilliland, R. L., Hogan, C. J., Jha, S., Kirshner, R. P., Leibundgut, B., Phillips, M. M., Reiss, D., Schmidt, B. P., Schommer, R. A., Smith, R. C., Spyromilio, J., Stubbs, C., Suntzeff, N. B., and Tonry, J. (1998). Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *The Astronomical Journal*, 116(3):1009–1038.

Rubin, V. C., Ford, W. K., J., and Thonnard, N. (1980). Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 (R=4kpc) to UGC 2885 (R=122kpc). *The Astrophysical Journal*, 238:471–487.

Salcedo, A. N., Maller, A. H., Berlind, A. A., Sinha, M., McBride, C. K., Behroozi, P. S., Wechsler, R. H., and Weinberg, D. H. (2018). Spatial clustering of dark matter haloes: secondary bias, neighbour bias, and the influence of massive neighbours on halo properties. *Monthly Notices of the Royal Astronomical Society*, 475:4411–4423.

Salcedo, A. N., Zu, Y., Zhang, Y., Wang, H., Yang, X., Wu, Y., Jing, Y., Mo, H., and Weinberg, D. H. (2020). Elucidating Galaxy Assembly Bias in SDSS. *arXiv e-prints*, page arXiv:2010.04176.

Sawala, T., Frenk, C. S., Crain, R. A., Jenkins, A., Schaye, J., Theuns, T., and Zavala, J. (2013). The abundance of (not just) dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 431:1366–1382.

Scannapieco, C., Wadepuhl, M., Parry, O. H., Navarro, J. F., Jenkins, A., Springel, V., Teyssier, R., Carlson, E., Couchman, H. M. P., Crain, R. A., Dalla Vecchia, C., Frenk, C. S., Kobayashi, C., Monaco, P., Murante, G., Okamoto, T., Quinn, T., Schaye, J., Stinson, G. S., Theuns, T., Wadsley, J., White, S. D. M., and Woods, R. (2012). The Aquila comparison project: the effects of feedback and numerical methods on simulations of galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 423:1726–1749.

Schaller, M., Frenk, C. S., Bower, R. G., Theuns, T., Jenkins, A., Schaye, J., Crain, R. A., Furlong, M., Dalla Vecchia, C., and McCarthy, I. G. (2015). Baryon effects on the internal structure of ΛCDM haloes in the EAGLE simulations. *Monthly Notices of the Royal Astronomical Society*, 451:1247–1267.

Schaye, J., Crain, R. A., Bower, R. G., Furlong, M., Schaller, M., Theuns, T., Dalla Vecchia, C., Frenk, C. S., McCarthy, I. G., Helly, J. C., Jenkins, A., Rosas-Guevara, Y. M., White, S. D. M., Baes, M., Booth, C. M., Camps, P., Navarro, J. F., Qu, Y., Rahmati, A., Sawala, T., Thomas, P. A., and Trayford, J. (2015). The EAGLE project: simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the Royal Astronomical Society*, 446:521–554.

Scherrer, R. J. and Bertschinger, E. (1991). Statistics of primordial density perturbations from discrete seed masses. *The Astrophysical Journal*, 381:349–360.

Scherrer, R. J. and Weinberg, D. H. (1998). Constraints on the Effects of Locally Biased Galaxy Formation. , 504(2):607–611.

Scoccimarro, R. (1998). Transients from initial conditions: a perturbative analysis. *Monthly Notices of the Royal Astronomical Society*, 299(4):1097–1118.

Scoccimarro, R., Sheth, R. K., Hui, L., and Jain, B. (2001). How Many Galaxies Fit in a Halo? Constraints on Galaxy Formation Efficiency from Spatial Clustering. *The Astrophysical Journal*, 546:20–34.

Seljak, U. (2000). Analytic model for galaxy and dark matter clustering. *Monthly Notices of the Royal Astronomical Society*, 318:203–213.

Seljak, U. and Zaldarriaga, M. (1996). A Line-of-Sight Integration Approach to Cosmic Microwave Background Anisotropies. *The Astrophysical Journal*, 469:437.

Sheth, R. K., Hui, L., Diaferio, A., and Scoccimarro, R. (2001). Linear and non-linear contributions to pairwise peculiar velocities. *Monthly Notices of the Royal Astronomical Society*, 325:1288–1302.

Sinha, M., Berlind, A. A., McBride, C. K., Scoccimarro, R., Piscionere, J. A., and Wibking, B. D. (2018). Towards accurate modelling of galaxy clustering on small scales: testing the standard ΛCDM + halo model. *Monthly Notices of the Royal Astronomical Society*, 478:1042–1064.

Sinha, M. and Garrison, L. (2017). Corrfunc: Blazing fast correlation functions on the CPU. Astrophysics Source Code Library.

Sinha, M. and Garrison, L. H. (2019). Corrfunc — A Suite of Blazing Fast Correlation Functions on the CPU. *Monthly Notices of the Royal Astronomical Society*, page 2750.

Spergel, D. N., Verde, L., Peiris, H. V., Komatsu, E., Nolta, M. R., Bennett, C. L., Halpern, M., Hinshaw, G., Jarosik, N., Kogut, A., Limon, M., Meyer, S. S., Page, L., Tucker, G. S., Weiland, J. L., Wollack, E., and Wright, E. L. (2003). First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters. *The Astrophysical Journal Supplement*, 148(1):175–194.

Springel, V. (2005). The cosmological simulation code GADGET-2. *Monthly Notices of the Royal Astronomical Society*, 364:1105–1134.

Springel, V., Pakmor, R., Pillepich, A., Weinberger, R., Nelson, D., Hernquist, L., Vogelsberger, M., Genel, S., Torrey, P., Marinacci, F., and Naiman, J. (2018). First results from the IllustrisTNG simulations: matter and galaxy clustering. *Monthly Notices of the Royal Astronomical Society*, 475:676–698.

Springel, V., White, S. D. M., Jenkins, A., Frenk, C. S., Yoshida, N., Gao, L., Navarro, J., Thacker, R., Croton, D., Helly, J., Peacock, J. A., Cole, S., Thomas, P., Couchman, H., Evrard, A., Colberg, J., and Pearce, F. (2005). Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, 435(7042):629–636.

Szewciw, A. O., Beltz-Mohrmann, G. D., Berlind, A. A., and Sinha, M. (2022). Toward Accurate Modeling of Galaxy Clustering on Small Scales: Constraining the Galaxy-halo Connection with Optimal Statistics. *The Astrophysical Journal*, 926(1):15.

The EAGLE team (2017). The EAGLE simulations of galaxy formation: Public release of particle data. *arXiv e-prints*.

Tinker, J. L., Conroy, C., Norberg, P., Patiri, S. G., Weinberg, D. H., and Warren, M. S. (2008). Void Statistics in Large Galaxy Redshift Surveys: Does Halo Occupation of Field Galaxies Depend on Environment? *The Astrophysical Journal*, 686:53–71.

Tinker, J. L., Wechsler, R. H., and Zheng, Z. (2010). Interpreting the Clustering of Distant Red Galaxies. *The Astrophysical Journal*, 709:67–76.

Tinker, J. L., Weinberg, D. H., and Warren, M. S. (2006a). Cosmic Voids and Galaxy Bias in the Halo Occupation Framework. *The Astrophysical Journal*, 647:737–752.

Tinker, J. L., Weinberg, D. H., and Zheng, Z. (2006b). Redshift-space distortions with the halo occupation distribution - I. Numerical simulations. *Monthly Notices of the Royal Astronomical Society*, 368(1):85–108.

Tinker, J. L., Weinberg, D. H., Zheng, Z., and Zehavi, I. (2005). On the Mass-to-Light Ratio of Large-Scale Structure. *The Astrophysical Journal*, 631:41–58.

Tonegawa, M., Park, C., Zheng, Y., Park, H., Hong, S. E., Hwang, H. S., and Kim, J. (2020). Cosmological Information from the Small-scale Redshift-space Distortion. *The Astrophysical Journal*, 897(1):17.

Vakili, M. and Hahn, C. (2019). How Are Galaxies Assigned to Halos? Searching for Assembly Bias in the SDSS Galaxy Clustering. *The Astrophysical Journal*, 872:115.

van Daalen, M. P., Schaye, J., McCarthy, I. G., Booth, C. M., and Dalla Vecchia, C. (2014). The impact of baryonic processes on the two-point correlation functions of galaxies, subhaloes and matter. *Monthly Notices of the Royal Astronomical Society*, 440(4):2997–3010.

Van den Bosch, F. C., Weinmann, S. M., Yang, X., Mo, H. J., Li, C., and Jing, Y. P. (2005). The phase-space parameters of the brightest halo galaxies. *Monthly Notices of the Royal Astronomical Society*, 361:1203–1215.

Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30.

Velliscig, M., van Daalen, M. P., Schaye, J., McCarthy, I. G., Cacciato, M., Le Brun, A. i. M. C., and Dalla Vecchia, C. (2014). The impact of galaxy formation on the total mass, mass profile and abundance of haloes. *Monthly Notices of the Royal Astronomical Society*, 442(3):2641–2658.

Villarreal, A. S., Zentner, A. R., Mao, Y.-Y., Purcell, C. W., van den Bosch, F. C., Diemer, B., Lange, J. U., Wang, K., and Campbell, D. (2017). The immitigable nature of assembly bias: the impact of halo definition on assembly bias. *Monthly Notices of the Royal Astronomical Society*, 472(1):1088–1105.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and Contributors, S. . (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Vogelsberger, M., Genel, S., Springel, V., Torrey, P., Sijacki, D., Xu, D., Snyder, G., Bird, S., Nelson, D., and Hernquist, L. (2014a). Properties of galaxies reproduced by a hydrodynamic simulation. *Nature*, 509:177–182.

Vogelsberger, M., Genel, S., Springel, V., Torrey, P., Sijacki, D., Xu, D., Snyder, G., Nelson, D., and Hernquist, L. (2014b). Introducing the Illustris Project: simulating the coevolution of dark and visible matter in the Universe. *Monthly Notices of the Royal Astronomical Society*, 444:1518–1547.

Walsh, K. and Tinker, J. (2019). Probing Galaxy Assembly Bias in BOSS Galaxies Using Void Probabilities. *Monthly Notices of the Royal Astronomical Society*.

Wang, K., Mao, Y.-Y., Zentner, A. R., van den Bosch, F. C., Lange, J. U., Schafer, C. M., Villarreal, A. S., Hearin, A. P., and Campbell, D. (2019). How to optimally constrain galaxy assembly bias: supplement projected correlation functions with count-in-cells statistics. *Monthly Notices of the Royal Astronomical Society*, 488(3):3541–3567.

Wang, Y., Vogelsberger, M., Xu, D., Mao, S., Springel, V., Li, H., Barnes, D., Hernquist, L., Pillepich, A., Marinacci, F., Pakmor, R., Weinberger, R., and Torrey, P. (2020). Early-type galaxy density profiles from IllustrisTNG - I. Galaxy correlations and the impact of baryons. *Monthly Notices of the Royal Astronomical Society*, 491(4):5188–5215.

Watson, D. F., Berlind, A. A., McBride, C. K., Hogg, D. W., and Jiang, T. (2012). The Extreme Small Scales: Do Satellite Galaxies Trace Dark Matter? *The Astrophysical Journal*, 749:83.

Wechsler, R. H., Zentner, A. R., Bullock, J. S., Kravtsov, A. V., and Allgood, B. (2006). The Dependence of Halo Clustering on Halo Formation History, Concentration, and Occupation. *The Astrophysical Journal*, 652:71–84.

Weinberger, R., Springel, V., Hernquist, L., Pillepich, A., Marinacci, F., Pakmor, R., Nelson, D., Genel, S., Vogelsberger, M., Naiman, J., and Torrey, P. (2017). Simulating galaxy formation with black hole driven thermal and kinetic feedback. *Monthly Notices of the Royal Astronomical Society*, 465(3):3291–3308.

White, M., Hernquist, L., and Springel, V. (2001). The Halo Model and Numerical Simulations. *The Astrophysical Journal Letters*, 550:L129–L132.

Xu, X. and Zheng, Z. (2018). Galaxy assembly bias of central galaxies in the Illustris simulation. *arXiv e-prints*.

Xu, X. and Zheng, Z. (2020). Galaxy assembly bias of central galaxies in the Illustris simulation. *Monthly Notices of the Royal Astronomical Society*, 492(2):2739–2754.

York, D. G., Adelman, J., Anderson, Jr., J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J. A., Barkhouser, R., Bastian, S., Berman, E., Boroski, W. N., Bracker, S., Briegel, C., Briggs, J. W., Brinkmann, J., Brunner, R., Burles, S., Carey, L., Carr, M. A., Castander, F. J., Chen, B., Colestock, P. L., Connolly, A. J., Crocker, J. H., Csabai, I., Czarapata, P. C., Davis, J. E., Doi, M., Dombeck, T., Eisenstein, D., Ellman, N., Elms, B. R., Evans, M. L., Fan, X., Federwitz, G. R., Fiscelli, L., Friedman, S., Frieman, J. A., Fukugita, M., Gillespie, B., Gunn, J. E., Gurbani, V. K., de Haas, E., Haldeman, M., Harris, F. H., Hayes, J., Heckman, T. M., Hennessy, G. S., Hindsley, R. B., Holm, S., Holmgren, D. J., Huang, C.-h., Hull, C., Husby, D., Ichikawa, S.-I., Ichikawa, T., Ivezić, Ž., Kent, S., Kim, R. S. J., Kinney, E., Klaene, M., Kleinman, A. N., Kleinman, S., Knapp, G. R., Korienek, J., Kron, R. G., Kunszt, P. Z., Lamb, D. Q., Lee, B., Leger, R. F., Limmongkol, S., Lindenmeyer, C., Long, D. C., Loomis, C., Loveday, J., Lucinio, R., Lupton, R. H., MacKinnon, B., Mannery, E. J., Mantsch, P. M., Margon, B., McGehee, P., McKay, T. A., Meiksin, A., Merelli, A., Monet, D. G., Munn, J. A., Narayanan, V. K., Nash, T., Neilsen, E., Neswold, R., Newberg, H. J., Nichol, R. C., Nicinski, T., Nonino, M., Okada, N., Okamura, S., Ostriker, J. P., Owen, R., Pauls, A. G., Peoples, J., Peterson, R. L., Petravick, D., Pier, J. R., Pope, A., Pordes, R., Prosapio, A., Rechenmacher, R., Quinn, T. R., Richards, G. T., Richmond, M. W., Rivetta, C. H., Rockosi, C. M., Ruthmansdorfer, K., Sandford, D., Schlegel, D. J., Schneider, D. P., Sekiguchi, M., Sergey, G., Shimasaku, K., Siegmund, W. A., Smee, S., Smith, J. A., Snedden, S., Stone, R., Stoughton, C., Strauss, M. A., Stubbs, C., SubbaRao, M., Szalay, A. S., Szapudi, I., Szokoly, G. P., Thakar, A. R., Tremonti, C., Tucker, D. L., Uomoto, A., Vanden Berk, D., Vogeley, M. S., Waddell, P., Wang, S.-i., Watanabe, M., Weinberg, D. H., Yanny, B., Yasuda, N., and SDSS Collaboration (2000). The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, 120:1579–1587.

Zaldarriaga, M. and Seljak, U. (2000). CMBFAST for Spatially Closed Universes. *The Astrophysical Journal Supplement*, 129(2):431–434.

Zaldarriaga, M., Seljak, U., and Bertschinger, E. (1998). Integral Solution for the Microwave Background Anisotropies in Nonflat Universes. *The Astrophysical Journal*, 494(2):491–502.

Zehavi, I., Blanton, M. R., Frieman, J. A., Weinberg, D. H., Mo, H. J., Strauss, M. A., Anderson, S. F., Annis, J., Bahcall, N. A., Bernardi, M., Briggs, J. W., Brinkmann, J., Burles, S., Carey, L., Castander, F. J., Connolly, A. J., Csabai, I., Dalcanton, J. J., Dodelson, S., Doi, M., Eisenstein, D., Evans, M. L., Finkbeiner, D. P., Friedman, S., Fukugita, M., Gunn, J. E., Hennessy, G. S., Hindsley, R. B., Ivezić, Ž., Kent, S., Knapp, G. R., Kron, R., Kunszt, P., Lamb, D. Q., Leger, R. F., Long, D. C., Loveday, J., Lupton, R. H., McKay, T., Meiksin, A., Merrelli, A., Munn, J. A., Narayanan, V., Newcomb, M., Nichol, R. C., Owen, R., Peoples, J., Pope, A., Rockosi, C. M., Schlegel, D., Schneider, D. P., Scoccimarro, R., Sheth, R. K., Siegmund, W., Smee, S., Snir, Y., Stebbins, A., Stoughton, C., SubbaRao, M., Szalay, A. S., Szapudi, I., Tegmark, M., Tucker, D. L., Uomoto, A., Vanden Berk, D., Vogeley, M. S., Waddell, P., Yanny, B., and York, D. G. (2002). Galaxy Clustering in Early Sloan Digital Sky Survey Redshift Data. *The Astrophysical Journal*, 571:172–190.

Zehavi, I., Contreras, S., Padilla, N., Smith, N. J., Baugh, C. M., and Norberg, P. (2018). The Impact of Assembly Bias on the Galaxy Content of Dark Matter Halos. *The Astrophysical Journal*, 853:84.

Zehavi, I., Weinberg, D. H., Zheng, Z., Berlind, A. A., Frieman, J. A., Scoccimarro, R., Sheth, R. K., Blanton, M. R., Tegmark, M., Mo, H. J., Bahcall, N. A., Brinkmann, J., Burles, S., Csabai, I., Fukugita, M., Gunn, J. E., Lamb, D. Q., Loveday, J., Lupton, R. H., Meiksin, A., Munn, J. A., Nichol, R. C., Schlegel, D., Schneider, D. P., SubbaRao, M., Szalay, A. S., Uomoto, A., York, D. G., and SDSS Collaboration (2004). On Departures from a Power Law in the Galaxy Correlation Function. *The Astrophysical Journal*, 608:16–24.

Zehavi, I., Zheng, Z., Weinberg, D. H., Blanton, M. R., Bahcall, N. A., Berlind, A. A., Brinkmann, J., Frieman, J. A., Gunn, J. E., Lupton, R. H., Nichol, R. C., Percival, W. J., Schneider, D. P., Skibba, R. A., Strauss, M. A., Tegmark, M., and York, D. G. (2011). Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity. *The Astrophysical Journal*, 736:59.

Zehavi, I., Zheng, Z., Weinberg, D. H., Frieman, J. A., Berlind, A. A., Blanton, M. R., Scoccimarro, R., Sheth, R. K., Strauss, M. A., Kayo, I., Suto, Y., Fukugita, M., Nakamura, O., Bahcall, N. A., Brinkmann, J., Gunn, J. E., Hennessy, G. S., Ivezić, Ž., Knapp, G. R., Loveday, J., Meiksin, A., Schlegel, D. J., Schneider, D. P., Szapudi, I., Tegmark, M., Vogeley, M. S., York, D. G., and SDSS Collaboration (2005). The Luminosity and Color Dependence of the Galaxy Correlation Function. *The Astrophysical Journal*, 630:1–27.

Zel'Dovich, Y. B. (1970). Reprint of 1970A&A.....5...84Z. Gravitational instability: an approximate theory for large density perturbations. *Astronomy & Astrophysics*, 500:13–18.

Zentner, A. R., Hearin, A., van den Bosch, F. C., Lange, J. U., and Villarreal, A. (2019). Constraints on assembly bias from galaxy clustering. *Monthly Notices of the Royal Astronomical Society*, 485:1196–1209.

Zentner, A. R., Hearin, A. P., and van den Bosch, F. C. (2014). Galaxy assembly bias: a significant source of systematic error in the galaxy-halo relationship. *Monthly Notices of the Royal Astronomical Society*, 443:3044–3067.

Zheng, Z. (2004). Interpreting the Observed Clustering of Red Galaxies at z ˜ 3. *The Astrophysical Journal*, 610:61–68.

Zheng, Z., Berlind, A. A., Weinberg, D. H., Benson, A. J., Baugh, C. M., Cole, S., Davé, R., Frenk, C. S., Katz, N., and Lacey, C. G. (2005). Theoretical Models of the Halo Occupation Distribution: Separating Central and Satellite Galaxies. *The Astrophysical Journal*, 633:791–809.

Zheng, Z., Coil, A. L., and Zehavi, I. (2007). Galaxy Evolution from Halo Occupation Distribution Modeling of DEEP2 and SDSS Galaxy Clustering. *The Astrophysical Journal*, 667:760–779.

Zheng, Z. and Weinberg, D. H. (2007). Breaking the Degeneracies between Cosmology and Galaxy Bias. *The Astrophysical Journal*, 659:1–28.

Zu, Y. and Mandelbaum, R. (2018). Mapping stellar content to dark matter haloes - III. Environmental dependence and conformity of galaxy colours. *Monthly Notices of the Royal Astronomical Society*, 476(2):1637–1653.