Illuminating and Tabulating the Galaxy-Halo Connection

by

Alan Pearl

BS Physics, Rensselaer Polytechnic Institute, 2017

Submitted to the Graduate Faculty of

the Department of Physics and Astronomy in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF PHYSICS AND ASTRONOMY

This dissertation was presented

by

Alan Pearl

It was defended on

June 28, 2023

and approved by

Andrew Zentner, Prof., Physics & Astronomy, University of Pittsburgh

Rachel Bezanson, Associate Prof., Physics & Astronomy, University of Pittsburgh

Jeffrey Newman, Prof., Physics & Astronomy, University of Pittsburgh

Joe Boudreau, Prof., Physics & Astronomy, University of Pittsburgh

Rachel Mandelbaum, Prof., Astrophysics & Cosmology, Carnegie Melon University

Dissertation Director: Andrew Zentner, Prof., Physics & Astronomy, University of Pittsburgh

Illuminating and Tabulating the Galaxy-Halo Connection

Alan Pearl, PhD

University of Pittsburgh, 2023

In the near future, a new generation of massively multiplexed spectroscopic surveys like PFS, WAVES, and MOONS will enable detailed studies of galaxy evolution across cosmic timescales and connect galaxies to dark matter halos. I have generated realistic high-redshift mock catalogs for each of these three planned surveys to help quantify and optimize their scientific output. This uses a procedure I developed called Calibrating Light: Illuminating Mocks By Empirical Relations (CLIMBER), and is based on the UniverseMachine model and UltraVISTA photometry. I have compared different targeting strategies by varying the area and targeting completeness and quantified how these survey parameters affect the uncertainty of the two-point correlation function. Through mock observations, I have demonstrated that future extensions of the PFS and MOONS programs should primarily aim to reduce cosmic variance by surveying more uncorrelated sky areas. Additionally, I developed the `galtab` algorithm to enhance the efficiency of HOD inference by pretabulating populations of galaxies in simulated halo catalogs for rapid, quasi-deterministic estimation of counts-in-cells statistics. This methodology allows posterior probability distributions from Markov chains to converge much more quickly by reducing the required number of trial points by up to an order of magnitude, in addition to enabling even more drastic speedups due to its GPU portability. Leveraging early data from DESI, I have explored the galaxy-halo connection by supplementing number density and the two-point correlation function with `galtab`-accelerated counts-in-cylinders. My analysis tightly constrains characteristic halo masses and provides strong statistical evidence for positive assembly bias between the r-band luminosity and the halo concentration with up to 3σ significance. Looking ahead, as new methodologies and datasets continue to facilitate our understanding, I anticipate that future studies will soon shift their focus from mere detections of assembly bias to delving into its implications for galaxy formation and cosmology at greater depth.

Table of Contents

List of Tables

List of Figures

Preface

I am extremely fortunate to have had such strong support from my family, friends, and mentors. Their continual support and encouragement have provided me with the strength necessary to face the challenges of completing a Ph.D. degree.

To my loving family, thank you for setting me up for success. I believe that my accomplishments can be traced back to my parents' dedication to my education and their proud words of encouragement. My love of math may be thanks to playing schoolhouse with my sister Amy, who started teaching her three little brothers — Andrew, Joey, and me — from a young age. Thank you to my cousins, aunts, and uncles who have always been there for me. I want to thank my grandparents, Dommy and Bob — Bob's curiosity about physics, space, and artificial intelligence was formative in my decision to study astrophysics. Thank you to my partner Rachel for loving and supporting me every day.

I am grateful to my friends for making my six years here enjoyable. To all the old friends who have kept in touch, thank you for allowing me to visit at home and around the world. To my classmates, colleagues, post-docs, and other friends I've made in Pittsburgh, thank you for helping me survive through classes, research, and a pandemic. I have thoroughly enjoyed the countless entertaining discussions in the office and around the city. While I'm excited to start a new journey in Chicago this fall, I will miss the softball games, jam sessions, hikes in Frick Park, socially distant backyard drinks, Belvedere's indie dance nights, and Nico's karaoke (RIP). There are too many wonderful people to name, but you know who you are. Thank you for the memories, keep in touch, and I hope our paths cross again.

I could not have made it here without all of my mentors, from sports coaches and math teachers to my undergraduate research advisor, Heidi Newberg, who was monumental in my placement in this program. This chapter of my life has been no exception. I am enormously grateful to have not one but three professors directly advising my thesis. Andrew, Rachel, and Jeff — I don't take for granted the time you have each taken to teach and guide me over the years. Finally, I would like to thank the rest of my thesis committee, Joe Boudreau and Rachel Mandelbaum, for their valuable feedback and mentorship as well.

1.0 Introduction

1.1 Cosmology

One of the primary goals of modern astrophysics is to understand the initial conditions, evolution, and physical laws responsible for how the universe came to be the way it is today. The most successful model of cosmology to date, known as ΛCDM, is quite consistent with a multitude of observations spanning nearly the entire 13.8 billion years of evolution since the Big Bang. While ΛCDM has well-defined astrophysical implications, there remain unanswered questions as it remains agnostic to the precise type of substances responsible for dark energy (which is accounted for by cosmological constant, Λ, in Einstein's field equation) and cold dark matter (CDM). Note that, while ΛCDM has yet to be ruled out, there are several popular alternatives, such as models that include warm dark matter or time variations to the cosmological "constant" formulation of dark energy.

ΛCDM posits that the universe initially began expanding at the time of the Big Bang. At this time, the spatial distribution was nearly perfectly homogeneous and isotropic (an assumption known as the cosmological principle [46]), save for the minuscule quantum fluctuations that would eventually serve as the initial random seeds to form the complex structure present in the universe today. The evolution of this expansion can be derived from General Relativity [37] and is quantified as a function of time using a scale length, a, that starts at $a(0) = 0$ and ends at a $a(t_{\rm now}) = 1$ today, where $t_{\rm now}$ is the age of the universe (roughly 13.8 Gyr [89]). The functional form of $a(t)$ can be calculated given the relative composition of the universe assigned to matter, radiation, and dark energy. During much of the first 10^{-32} seconds after the Big Bang, the universe expanded extremely rapidly due to an unknown physical source we call inflation. For the next 47,000 years, the energy density of the universe was dominated by radiation (i.e., photons and other relativistic particles), resulting in a slight deceleration of the expansion (i.e., $d^2a/dt^2 < 0$). For approximately ten billion years following the radiation-domination period, the universe was dominated by matter, which continued to decelerate the expansion of the universe, but not enough for it to collapse back

in on itself. Following this time period, dark energy became dominant, which caused the expansion to begin exponentially increasing. The expansion of the universe continues to accelerate today, with no signs of halting.

Astronomical measurements have played a crucial role in validating the ΛCDM model and providing precise constraints on the energy composition and expansion history of the universe. These measurements predominantly rely on observations of light, which travels at a constant speed, allowing us to peer into the distant past. Determining the size of the universe at different times involves measuring the stretching (i.e., redshifting) of the wavelength of the light since it was emitted from its source. Independent measurements of distance (e.g., from the apparent brightness of Type Ia supernovae, which have a well-understood intrinsic luminosity [94]) and redshift have provided tight constraints on the evolution of the scale factor $a(t)$. However, even higher precision constraints arise from studying the cosmic microwave background (CMB [89]), which is radiation from a mere 380,000 years after the Big Bang when the universe was just 0.09% of its present size ($a = 0.0009$). The CMB reveals a significant clustering pattern known as baryonic acoustic oscillations (BAO), which can also be observed in later-universe galaxy populations, reinforcing our constraints. Collectively, these measurements have revealed that approximately 70% of the universe's energy budget is attributed to dark energy, whereas matter — predominantly dark matter — constitutes the remaining fraction.

Our precise understanding of the constituents and expansion history of the universe has enabled insight into the statistical nature of the large-scale matter distribution in the universe, which seeds the formation of galaxies, which are the birthing grounds of stars, which are capable of performing nucleosynthesis of the chemicals necessary for planets and life to exist. Amongst these processes, however, there still exist many mysteries and computational difficulties, particularly in fully understanding the evolutionary pathways of galaxies and what physical processes are responsible. Likewise, there is a wealth of cosmological information encoded into the spatial distribution of galaxies that trace the matter density field, and this is a primary driver of modern data-driven cosmological studies. Utilizing these data requires precise measurement of statistical observations and the development of accurate models of the statistical prescription for galaxy formation, which are the primary goals of

my thesis.

1.2 Gravity-Only Simulations

The quantum fluctuations present in the early universe present a statistical initial condition, where the density field at any given location is drawn from a Gaussian distribution. The density field proceeds to evolve as the overdensities grow and gravitationally attract one another. The abundance and statistical clustering (e.g., the two-point correlation function [32]) of overdensities can be derived analytically through linear perturbation theory, which does a good job of explaining large-scale clustering, where gravitational collapse evolves "linearly." However, in regimes of high gravitational attraction (e.g., high overdensities and small scales), this evolution becomes highly non-linear [28].

Thanks to the rapid advancement of modern supercomputers, it can be advantageous to forgo linear perturbation theory and instead simulate the gravitational evolution of a statistically large sample of massive particles exactly as dictated by Einstein's theory of general relativity [37] (which is usually approximated by Newtonian gravity). This approach is still limited by its ignorance of all fundamental forces except gravity, but it works well down to much smaller scales than linear perturbation theory because, in comparison, all other forces become negligible over sufficiently large separations. This assumption is further justified by the fact that dark matter — which accounts for 85% of all matter — is only known to interact through gravity.

In such gravity-only simulations, groups of particles often collapse into dense, gravitationally bound structures, known as halos. For the large-scale analyses that I perform in my thesis, halos can be approximated as spherically symmetric Navarro-Frenk-White (NFW [81]) profiles, although in reality, they are better fits to triaxial distributions [62]. These structures are of particular importance because they are the only locations with strong enough gravitational potentials for ordinary matter to collapse and form galaxies, which emit visible light that we can see from Earth. By analyzing very large simulations, we can quantify the statistics of these halos, such as their number density, clustering properties, and formation

histories. The most massive halos are formed hierarchically, through the merging of smaller progenitor halos. This leads to more massive halos being much rarer [91] and located in highly clustered environments [56, 76]. Since the abundance and clustering of halos are imprinted onto galaxies, we see similar trends where more massive galaxies are rarer and more clustered.

Abundance and clustering correlate not only with the mass of halos but also with other properties that vary for halos of different assembly histories. This phenomenon is known as halo assembly bias [43], which relies on the fact that a given halo's assembly history depends on the availability of accretion material in its local environment. For example, older halos (i.e., those that formed half of their mass earlier) tend to be more spatially clustered. The assembly history of a halo can also leave an imprint on other properties [98], such as spin (higher clustering produces faster spin) and concentration (higher clustering produces higher concentration, although this trend is reversed for the most massive halos).

Over the years, numerous gravity-only simulations have played a crucial role in advancing our understanding of the formation and evolution of cosmic structures. Conducted nearly two decades ago, the Millennium Simulation [105] was one of the pioneering simulations, which allowed for some of the first precise calculations of complex halo population statistics. The Millennium Simulation, with its cubical volume of $500h^{-1}$Mpc on a side, paved the way for the subsequent MultiDark and Bolshoi suites of simulations [61, 90, 60], each containing a similar number of particles to the Millennium Simulation. This large suite of simulations has given users the choice to analyze the differences between slightly different cosmological priors and a vast range of volumes (side lengths ranging from $160h^{-1}$Mpc to $4h^{-1}$Gpc), enabling analyses that require both higher resolution and larger sample sizes. The most recent state-of-the-art simulations, such as the Uchuu suite [53], have pushed the computational barrier even further, with a similar range of simulation volumes, but each simulation consists of hundreds to thousands of times more particles than the MultiDark simulations.

1.3 Galaxy Formation

Galaxies are the primary source of light that allows us to trace the cosmic web, and we now have observations going back. Therefore, they are an invaluable source of data that can help answer how the universe came to be the way it is today. Composed mainly of baryonic matter, galaxies undergo a series of physical processes that lead to the formation of gas clouds, stars, dust, and planets. Galaxies begin as dilute hydrogen gas clouds that approximately follow the same distribution as dark matter, until the density surpasses a critical threshold, triggering gravitational collapse. While dark matter also collapses gravitationally in dense regions, the collapse of baryonic gas is more efficient due to its ability to radiate away kinetic energy in the form of light. Gas clouds can continue to collapse until their cores become hot enough to fuse hydrogen into helium, and the gas cloud becomes a star. Stars, supernovae, and active galactic nuclei produce ionizing radiation that drives outflows and heats the surrounding gas, slowing down further star formation (i.e., star formation feedback).

To understand galaxy formation in the context of cosmological volumes, many research groups have developed hydrodynamic simulations such as IllustrisTNG [83], EAGLE [99, 29], FIRE [51], and SIMBA [31], just to name a few of the most recent examples. These simulations aim to capture much of the complex physics of galaxy formation. However, due to the vast range of scales and processes, each code relies on various approximations to account for the physics below the simulation resolution. While there are many high-resolution zoom-in simulations (e.g., [100, 70, 15]) that aim to justify the approximations made in those of larger volumes, modeling the physics of the formation of even a single galaxy from first principles remains computationally prohibitive. Therefore, empirical galaxy-halo connection models (as discussed in Section 1.4) play a crucial role in improving the reliability of hydrodynamic simulations and enabling cosmological inferences.

Whether empirically or physically motivated, the test of a good galaxy formation model is its ability to reproduce observational properties. This requires understanding how much light is emitted by the population of stars that have been formed in each galaxy. It is generally assumed that, whenever star formation occurs, there is a universal distribution of

masses of the newly formed stars known as the initial mass function (IMF [23]). The IMF predicts that low-mass stars form more frequently than high-mass stars, a trend originally observed in the empirical distribution of stars in the solar neighborhood [97]. Using the IMF, we can predict the light emitted by a stellar population created through an instantaneous burst of star formation. Since the massive, bright blue stars have shorter lifetimes, the light from the composite population gradually becomes redder over time. Therefore, blue galaxies indicate ongoing or recent star formation, while red galaxies are either composed entirely of older stellar populations or hide their young starlight behind optically thick dust clouds. Infrared observations of thermal dust emission can sometimes break this dust-age degeneracy [30].

To study the star formation history (SFH) of a galaxy, one employs a stellar population synthesis technique (e.g., [26, 22, 68]). By combining the light from stellar populations of different ages, the expected spectrum of a galaxy can be constructed. However, tightly constraining a given SFH is challenging without imposing strong priors on its functional form. Photometric data typically only offer a rough indication of a galaxy's age and redshift. However, high-quality observations covering a wide spectral range at high resolution can provide more meaningful constraints [54]. Such observations are valuable for studying the statistical dependence between galaxy assembly and the assembly of their underlying dark matter halos.

1.4 The Galaxy-Halo Connection

Thanks to the availability of dark matter halo catalogs from gravity-only simulations, it is possible to learn the statistical nature of galaxy formation and evolution. Models describing the galaxy-halo connection (GHC) aim to establish a statistical relationship between galaxies and their underlying dark matter halos, providing insights into these processes that cannot be simulated reliably. There are many types of GHC models, ranging from models that are physically motivated, like semi-analytic models (SAMs), to those that are purely empirical, like subhalo abundance matching (SHAM) or the halo occupation distribution (HOD). See

[114] and references therein for a detailed review of how these techniques are formulated and constrained.

One of the most commonly studied relationships is the stellar-mass-to-halo-mass relation [7]. This relationship is the most well-understood in the local universe, but it has been characterized all the way out to $z \sim 10$ [7]. Nearly all GHC models either produce or directly assume that there is a strong, monotonic correlation between stellar mass and halo mass. SAMs describe physical processes, such as gas cooling, star formation, and feedback to link the growth of galaxies to that of their host halos. SHAM and HOD models directly assign galaxies to halos based on their stellar mass, incorporating scatter to reproduce observed statistics. It should be noted that recent studies have suggested that the stellar mass of a galaxy may correlate better with other properties like the halo's gravitational potential depth, which is dependent on mass, but also introduces some dependence on the halo's concentration.

It is more challenging to understand the connection between halo mass and secondary galaxy properties, such as color, star-formation rate (SFR), size, or morphology. One of the challenges lies in the fact that these properties depend on stellar mass, which is itself correlated with halo mass, making it difficult to disentangle additional correlations. At fixed stellar mass, it is known that redder, lower-SFR, elliptical galaxies (which have more compact sizes than spirals) tend to live in higher-density environments [108], suggesting that they live in larger halo masses. In fact, several low-redshift studies [77, 128, 95] have found very strong evidence that galaxy color correlates with halo mass, especially high stellar masses ($\gtrsim 10^{11} M_{\odot}$).

Recently, there has been much research into whether the properties of a galaxy depend on its halo properties beyond the mass or gravitational potential (i.e., galaxy assembly bias). For example, many physically motivated galaxy formation theories (e.g., [104, 116]) predict a correlation between the accretion rate of dark matter and that of gas — the latter serving as the fuel for star formation. This would likely imply a correlation between the SFR of the galaxy and the mass accretion rate of the halo (i.e., assembly correlation), as is assumed by the UniverseMachine [7], an empirical model of the GHC. Several studies have presented evidence both supporting [49, 12, 6, 96] and challenging [84] this correlation across different

galaxy samples. Maintaining consistency with these studies, the best-fit UniverseMachine model prefers stronger assembly correlation for low-mass, low-redshift galaxies. However, stronger statistical constraints on this relationship are needed to infer which physical processes are responsible for driving and impeding the correlation between matter accretion and star formation.

Assembly correlation is one example of galaxy assembly bias, but it is possible that assembly bias could affect other properties of the galaxy, such as stellar mass or luminosity. This is important because if there is indeed a correlation between stellar mass and secondary halo properties, then the excess clustering due to halo assembly bias will be imprinted in the galaxy clustering signal. Under the standard HOD assumption that galaxies occupy halos as a function of halo mass alone, assembly bias could lead to biased results [125]. Some previous studies have found evidence that galaxy assembly bias is needed to explain observations [67, 124, 112], while others still claim that observations can be explained equally well using mass-only models [129]. This thesis presents additional evidence for assembly bias from a novel spectroscopic dataset.

1.5 Observational Signatures

Observations of various summary statistics play a crucial role in constraining the GHC, providing empirical data that can inform and validate our models. The most important observational constraint is that the number density of galaxies must match the number density of their host halos, which are known from simulations. Accurately measuring the number density of galaxies alone, as a function of their stellar mass or luminosity, can yield a fairly precise stellar-to-halo-mass relation. However, this requires assuming a perfectly monotonic SHMR, which works better at the high-mass end, where the halo mass function is very steep. Allowing for scatter in this relationship requires additional information.

Since the clustering of halo populations is also well characterized by simulations, galaxy clustering measurements are another essential tool for understanding the GHC. The primary metric used for clustering analyses is the two-point correlation function (2PCF [32]). The

2PCF quantifies the excess probability of finding pairs of galaxies at different separations compared to a random distribution. Its information can be marginalized in various ways to optimize its information into fewer degrees of freedom. Many analyses have used line-of-sight projected correlation functions, angular correlation functions, and the monopole and quadrupole moments of the 2PCF, to name a few. These measurements provide insights into galaxies' spatial distribution and clustering properties, allowing for inferences about the spread of halo masses and satellite occupations.

Beyond two-point metrics, higher-order functions, such as the three-point correlation function (3PCF [44]), offer additional insights into the galaxy distribution. The 3PCF captures the configurations of each combination of three galaxies, capable of indicating non-Gaussian features in the large-scale structure, such as the alignment of galaxies along filaments. Due to the high complexity and dimensionality of high-order correlation functions, it is common to turn to other summary statistics to probe higher-order clustering features. For example, the void probability function characterizes the probability of finding empty regions in the galaxy distribution. Counts-in-cells is a metric that counts the number of neighboring galaxies within a given region around each galaxy in the observational sample [119]. Some studies go beyond spatial clustering metrics — for example, satellite kinematics or gravitational lensing — because galaxies' dark matter environment directly affects their dynamical motion and the gravitational distortion of light.

In recent years, there has been a growing interest in moving beyond summary statistics and adopting field-level inference approaches [33]. These methods utilize models that must reproduce the entire galaxy field, rather than focusing on a set of summary statistics, which inevitably cannot convey all of the information available in the observed spatial distribution of galaxies. By directly analyzing this distribution and leveraging machine learning and likelihood-free techniques, these approaches may one day provide a more comprehensive understanding of the GHC and its uncertainties. However, many challenges remain in our ability to train such models in a computationally feasible, but fair way. Therefore, in my thesis, I primarily focus on analyzing summary statistics like number density, the projected correlation function, and counts-in-cells.

1.6 Survey Data

A galaxy survey is a program that observes the location and light output of a representative sample of a given galaxy population. In a given survey, our ability to measure the aforementioned three-dimensional spatial clustering statistics depends on how precisely we can recover distance, which is approximately given by the redshift of each galaxy spectrum. In this section, I highlight a few of the major surveys that are the most valuable to the work of my thesis, starting with those that have contributed significantly to our current understanding of the GHC, and ending with those coming in the near future.

The Sloan Digital Sky Survey (SDSS [1]) is one of the most influential surveys in observational astronomy. It has provided multi-color photometric and spectroscopic data for millions of galaxies, enabling GHC studies primarily from low-redshift spectroscopic samples. Along with spectroscopic redshifts from the 2-degree Field Galaxy Redshift Survey (2dFGRS [25]) and 6-degree Field Galaxy Survey (6dFGS [55]), SDSS has been instrumental in measuring the stellar-to-halo-mass relation in the local universe. While it is commonly assumed that this relation does not evolve strongly with redshift, this is a major test for the coming generation of SDSS-like surveys pushing to higher redshifts.

The Dark Energy Spectroscopic Instrument (DESI [34]) is an ongoing large-scale spectroscopic survey that aims to obtain spectra for tens of millions of galaxies, thanks to its slightly wider spectroscopic coverage and automated fiber placement mechanism, which allows for much more efficient use of time than the manual fiber placements required by SDSS. For individual objects, its spectra are also of higher quality and span slightly wider wavelength ranges compared to previous surveys, improving the statistical accuracy of redshift measurements. These incremental improvements will allow DESI to push to slightly higher redshifts in its luminosity-complete samples and will focus on other tracers of the density field at higher redshifts. However, this leaves room for future surveys to continue to fill in the gaps.

The Prime Focus Spectrograph (PFS [107]) is a massively multiplexed spectrograph that has been installed on the Subaru Telescope. In the coming years, the PFS Galaxy Evolution Survey will conduct a wide-field spectroscopic survey of millions of precise galaxy properties,

such as stellar masses, ages, and redshifts, which will enable galaxy clustering measurements into many subdivided galaxy populations at intermediate redshifts from $0.7 < z < 1.7$.

The Wide-Area VISTA Extra-galactic Survey (WAVES [36]) is an upcoming near-infrared spectroscopic survey using the 4-meter Multi-Object Spectrograph Telescope (4MOST) and the Visible and Infrared Survey Telescope for Astronomy (VISTA). Covering a large area of the sky, the WAVES-Deep survey will provide deep imaging of luminosity-complete samples out to moderately higher redshift than the comparable DESI sample. By probing the properties and clustering of this population, WAVES will play a crucial role by filling the redshift gap from $0.2 < z < 0.8$.

The Multi-Object Optical and Near-infrared Spectrograph (MOONS [73]) is a future spectroscopic survey that will operate on the Very Large Telescope. The MOONRISE survey aims to provide high-resolution spectroscopy for a large number of galaxies at high redshifts, from $0.9 < z < 2.6$. MOONS will offer valuable insights into the physical processes governing galaxy formation by probing galaxy kinematics and chemical properties at a very early, active time in cosmic history, enabled by very deep spectra.

This is by no means a complete list of all the observational programs that are foundational to our understanding of the GHC, but just a small selection of the most relevant ones to my thesis work, which I have shown provide measurements of diverse and complementary galaxy populations. My fiducial models are primarily informed by previous analyses of SDSS, and I have utilized early DESI data to validate and improve these models against brand new galaxy samples. Finally, I have forecasted the constraining power that we should achieve in the future with PFS, WAVES, and MOONS.

1.7 Dissertation Overview

The goal of my thesis is to improve our understanding of the galaxy-halo connection by (1) identifying optimal strategies for near-future high-redshift spectroscopic surveys to efficiently collect the most constraining datasets and (2) analyzing early DESI with a new CiC methodology to increase our galaxy clustering information content accurately and com-

putationally feasibly. I have made several significant contributions that I present in the form of two journal publications (one accepted and one in collaboration review). In the first publication, I introduced a tool, CLIMBER, for illuminating mock galaxies derived from the UniverseMachine, a state-of-the-art empirical model, and apply these mocks to the upcoming PFS, WAVES, and MOONS programs. In the second publication, I introduced a pretabulation-accelerated code, `galtab`, which drastically increases the efficiency with which we can perform HOD inference via counts-in-cells statistics, and applied it to early DESI data to find one of the most significant detections of galaxy assembly bias to date.

My dissertation is organized as follows: Chapter 1 provides a broad introduction of background material that aims to aid the reader in understanding the context and significance of my thesis. Chapter 2 is taken from my peer-reviewed publication detailing my methods for generating mock galaxy catalogs using the CLIMBER methodology. Chapter 3 is taken from my late-stage draft, currently in peer revision within the DESI collaboration, that describes `galtab` and my DESI HOD analysis that it has enabled. Finally, Chapter 4 summarizes the conclusions and accomplishments of my thesis and poses open research questions that may guide the future work of myself and others investigating the galaxy-halo connection.

2.0 CLIMBER: Galaxy-Halo Connection Constraints from Next-Generation Surveys

This chapter (Pearl et al. 2023 [85]) is published in the AstroPhysical Journal.

2.1 CLIMBER Introduction

The ΛCDM model of cosmology has been widely accepted for decades, and its parameters are now known to quite high precision [89]. Within this framework, it is assumed that galaxies form inside dark matter halos [120, 17]. While halos can be accurately modeled through gravity-only simulations (e.g., Bolshoi-Planck [60]), the fine details of galaxy formation are strongly influenced by baryonic physics, which poses a serious challenge for theoretical models of galaxy evolution to tackle.

In recent years, several ongoing projects have made great advancements to include baryonic physics in hydrodynamic simulations of galaxies in a cosmological context. These projects include but are not limited to IllustrisTNG [83], Evolution and Assembly of GaLaxies and their Environments (EAGLE [99, 29]), Feedback In Realistic Environments (FIRE [51]), and Simba [31]. However, it is still computationally prohibitive to resolve the small scales needed to simulate the processes that regulate star formation. Therefore, all of these hydrodynamic simulations still include analytic approximations for these small-scale processes. It is possible to approximate the rates of processes like gas cooling and star formation using semi-analytic models (SAMs; e.g., [121, 104]) which trace dark matter halos through gravity-only simulations and map baryonic physics into these halos using analytic scaling relations. SAMs have contributed significantly to our knowledge of galaxy formation, despite challenges disentangling various physical processes that produce degenerate observations.

Alternatively, many studies of galaxy evolution and cosmology use empirical models to populate galaxies on top of dark matter halos (i.e., the galaxy-halo connection; see [114] for an extensive review). Methods such as the halo occupation distribution (HOD; e.g.,

[10, 127, 50]) or abundance-matching (e.g., [63, 49]) are most commonly used to statistically match the stellar masses of galaxies to the masses of their host halos. Since the number density and clustering of halos are strongly dependent on halo mass [91, 56, 18, 76, 123], these models are informed through observations of the stellar mass function and the two-point correlation function [122, 92, 113].

Through analytic empirical models, we can infer the stellar-to-halo mass relation (SHMR). The SHMR indicates halo masses that are the most efficient at forming stellar mass. At its peak, the stellar mass can account for up to roughly 5% of the total mass of Milky Way-mass halos, which is approximately 30% of the cosmic baryon mass fraction. However, this star formation efficiency drops dramatically at both lower and higher halo masses. This is caused by various processes causing star formation to shut off (i.e., quench) through heating or removing the gas that was fueling the star formation. Low-mass quenching is often attributed to stellar feedback [41] and satellite stripping [45], while high-mass quenching is primarily attributed to active galactic nucleus (AGN) feedback [38]. However, its dependence on redshift is poorly constrained by existing datasets.

Most of our constraints on the galaxy-halo connection come from low-redshift surveys such as the Sloan Digital Sky Survey (SDSS [16]). While it is commonly assumed that the SHMR does not evolve strongly with redshift, it is particularly difficult to probe the same range of halos at high redshifts because these surveys quickly lose faint, low-mass galaxies and massive galaxies are rare. Extending our empirical constraints on the galaxy-halo connection to the high-redshift universe, where star formation rates (SFRs) were higher and galaxy populations were rapidly evolving, should have profound implications on our knowledge of galaxy evolution.

With the advent of highly multiplexed spectrographs being used on large telescopes, thousands of spectra will be simultaneously measured, a number of spectroscopic surveys will begin to map the distant universe to an unparalleled degree over the next decade. These surveys will probe the evolution of the precise statistical distribution of galaxies at earlier cosmological times than previously possible. Interpreting these types of datasets, however, is particularly challenging due to the systematic sampling that is more easily avoidable in the nearby universe. Utilizing this new information to place constraints on the galaxy-halo

connection will require careful planning of survey designs and new theoretical frameworks.

In this paper, we present a procedure for mapping photometric properties (flux and colors) onto physical properties from the UniverseMachine empirical model [7]. We refer to this procedure as Calibrating Light: Illuminating Mocks By Empirical Relations (CLIMBER). We use this procedure to construct mock galaxy catalogs, which we use to investigate the mass completeness and statistical constraints that will be available from several future massively multiplexed spectroscopic galaxy surveys: the Prime Focus Spectrograph Galaxy Evolution Survey (PFS [107]), the Guaranteed Time Observation Extragalactic Survey of the Multi-Object Optical and Near-infrared Spectrograph for the Very Large Telescope (MOONS [73]), and the Wide Area Vista Extragalactic Survey-Deep (WAVES [36]). For each survey, we quantify its performance and make recommendations about future extensions to improve its constraining power on the galaxy-halo connection.

This paper is organized as follows: Section 2.2 explains the procedure we followed to construct our mock galaxy catalog from the fiducial UniverseMachine model (with more details in Section 2.7) and discusses selection functions that we place to construct mock surveys of various galaxy populations. In Section 2.3, we formulate our conservative HOD model. In Section 2.4, we present mock measurements of number density and the two-point correlation function, which are the primary constraints of this model. In Section 2.5, we present projected constraints of the two-point correlation function and HOD models through Markov chain Monte Carlo (MCMC) fits, for a variety of survey parameters. We give a brief discussion of our conclusions in Section 2.6.

The cosmological assumptions used in each step of generating our mock catalog were made self-consistently. Bolshoi-Planck, the UniverseMachine, and all of our following calculations use a Planck-tuned ΛCDM cosmology with parameters given in Table 1. Although the stellar masses and SFRs from UltraVISTA [79] assumed a slightly different cosmology, their dependence on h has been corrected to match our assumption. Note that all halo masses refer to the virial mass of the halo, and we do not use h-scaled units with the exception of h^{-1} Mpc for distance.

Table 1: Cosmological parameters

Parameter	Value	Description
h	0.678	Hubble parameter
Ω_Λ	0.693	density parameter for dark energy
Ω_m	0.307	density parameter for total matter
Ω_b	0.048	density parameter for baryonic matter
n_s	0.96	normalization of the Power spectrum
σ_8	0.823	amplitude of mass density fluctuation

2.2 Building the Empirically Calibrated Mock Surveys

To construct a realistic mock galaxy catalog, we start from the UniverseMachine [7] empirical model, which is calibrated to reliably reproduce a very large number of statistics of galaxy populations from $0 < z < 10$. However, this model lacks a crucial element needed to test empirical selection functions: the apparent brightness of each galaxy in the observed-frame wavelengths of photometric filters. We, therefore, calibrate these model galaxies to photometry from the UltraVISTA survey [79] using a combination of abundance matching and random forest mapping to calculate observed mass-to-light ratios (see Section 2.2.2).

2.2.1 UniverseMachine

The UniverseMachine [7] is a sophisticated empirical galaxy-halo connection model of 44 parameters, which were iteratively fit to 1069 observed data points across a redshift range of $0 < z < 10$. It traces each dark matter halo in a gravity-only simulation and assigns a SFR to the galaxy at the center of the halo, assuming that star formation correlates with dark matter assembly. The model thereby tracks the accumulation of stellar mass of each galaxy over its entire formation history. The SFR of each galaxy is drawn from an empirically motivated distribution, which is the sum of two log-normal distributions, representing a

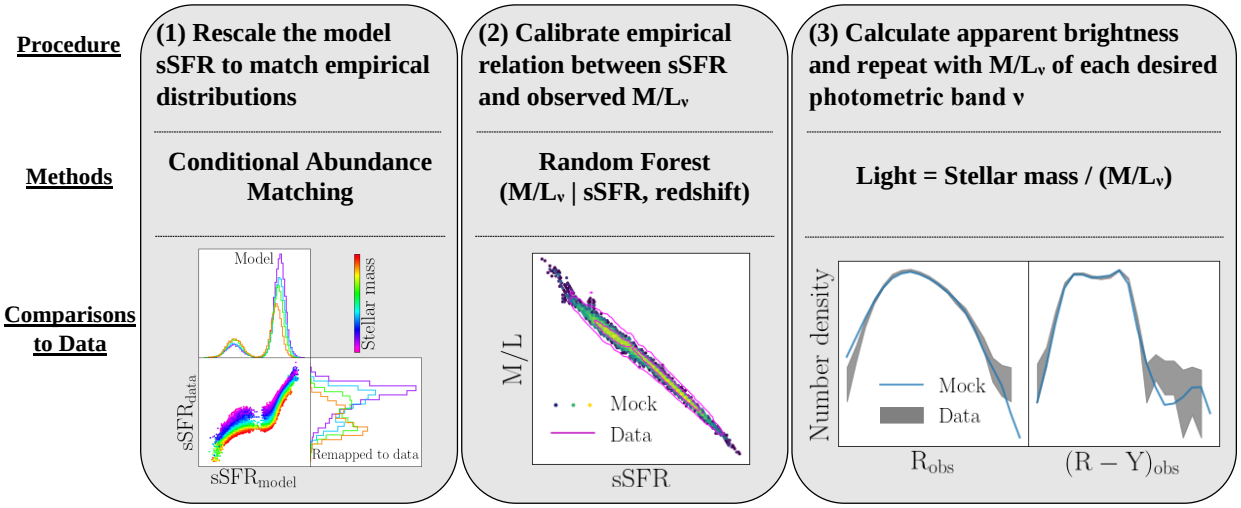

Figure 1: Visualization of our calibration procedure (CLIMBER) developed to assign the brightness and color of each mock galaxy taken from the UniverseMachine empirical model. Note that we abbreviate specific SFR (sSFR; i.e., SFR divided by stellar mass) and stellar mass-to-light ratio (M/L_ν), where ν represents the effective observed-frame frequency of a photometric band.

quenched and star-forming population. The quenched fraction, as well as the center and width of the star-forming distribution, are parameterized by analytic expressions dependent on halo mass and redshift. In order to impose some assembly correlation, the SFR is not randomly drawn from the model distribution, but instead weighted such that higher SFRs are more likely to be assigned to halos with greater mass accretion rates. For an excellent visual summary of this procedure, see Figure 1 of [7].

The UniverseMachine DR1 derives its halo catalog from the Bolshoi-Planck cosmological N-body simulation [60]. The UniverseMachine then provides mock galaxy properties such SFR and stellar mass (note that this is the "live" stellar mass, which is the integral of the star formation history subtracted by the mass returned to the interstellar medium) into each snapshot of this simulation. By piecing together these snapshots, the UniverseMachine has been tuned to reproduce many observables, such as stellar mass functions, two-point correlation functions, the star-forming main sequence, quenched fractions, environmental quenching, and more.

Before using the UniverseMachine to generate mock surveys, one needs to define the empirical properties of each mock galaxy to impose selection functions. This has previously been done by performing stellar population synthesis over each star formation history to fit the UniverseMachine to UV luminosity functions and UVJ quenching classifications. However, the UniverseMachine is only tuned to reproduce global star formation histories. Because our goal is to apply targeting strategies, the distribution of colors and fluxes must be reliable, and we, therefore, empirically calibrate a mapping from stellar mass and SFR (the `obs_sm` and `obs_sfr` columns) to the brightness in various photometric filters, as explained in Section 2.2.2.

2.2.2 CLIMBER

Most spectroscopic galaxy surveys have well-defined selection functions based on previously taken photometric data. Therefore, we need a method of predicting the observed light of UniverseMachine galaxies at multiple wavelengths to understand the representation of properties of the targeted galaxies (e.g., mass completeness) of these surveys. For this

reason, we have developed a procedure to assign mock apparent magnitudes informed by an observational dataset. We refer to this procedure as Calibrating Light: Illuminating Mocks By Empirical Relations (CLIMBER).

In CLIMBER, we utilize the tight correlation between the mass-to-light ratio and color of a galaxy (e.g., [9]). Analogously, we map sSFR to the mass-to-light ratio for each mock galaxy via random forest regression. This can be trained by any observational dataset with the desired mass, redshift, and photometric coverage. The data shown in this paper have been calibrated to UltraVISTA photometry [79], but CLIMBER is a generalizable, flexible procedure that can be applied to any dataset that includes mappings between empirical fluxes and physical galaxy properties.

While the scaling relation between total SFR and stellar mass of star-forming galaxies (the star-forming main sequence) has been measured out to $z \sim 5$ [57, 117, 109, 69], the evolution of SFRs for galaxies that fall off this relation is poorly constrained due to the difficulty of measuring low SFRs [68]. For this reason, the UniverseMachine only evolves the star-forming SFR distribution with redshift. In contrast, the specific SFR (sSFR) of each quiescent galaxy in the UniverseMachine was simply drawn from a non-evolving log-normal distribution of $10^{-11.8}$ yr$^{-1} \pm 0.36$ dex. While this empirically matches the local universe, it likely underestimates SFRs of high-redshift quiescent populations. This assumption does not greatly influence the accumulation of stellar mass modeled in the UniverseMachine, but it presents a problem for assigning the luminosities of quiescent galaxies. We solve this by rescaling the sSFR distributions via conditional abundance matching, as discussed in greater detail in Section 2.7.

The most crucial decision one needs to make before running CLIMBER is in choosing an sSFR proxy from the calibration dataset. This proxy must (1) approximately conserve rank-ordering with true sSFR (at fixed stellar mass), (2) produce a tight, negative correlation with mass-to-light ratio, and (3) have a high detection fraction in the full galaxy population – quenched and star-forming galaxies alike. From UltraVISTA, we chose to use the specific ultraviolet SFR ($\mathrm{sSFR_{UV}}$), which is the SFR inferred from ultraviolet bands, divided by stellar mass derived from SED fitting. These SEDs were fit by Fitting and Assessment of Synthetic Templates (FAST [64]) using the Chabrier [23] initial mass function, an exponen-

tially declining star formation history, and the Bruzual & Charlot [21] stellar population synthesis model. Other sSFR proxies may be useful in different datasets. For example, we considered using sSFRs directly from SED fits, but the grid-based values fit by FAST were sampled too sparsely and therefore provide a poor mapping between physical properties and flux.

See Figure 1 for a flow chart visualization of the CLIMBER procedure. To summarize, we first perform conditional abundance matching from the model sSFR to match the empirical sSFR distribution. Then, we train the mapping from sSFR to an observed mass-to-light ratio using random forest. Finally, we convert the UniverseMachine stellar mass values to luminosities via the predicted mass-to-light ratios. For further details and analysis of our procedure, see Section 2.7.

2.2.3 Mock Survey Selections

The product of CLIMBER is a mock realization of the universe in the form of a light cone, which can be iterated over random origins and orientations in the Bolshoi-Planck cube. We conduct mock surveys over these light cones by performing cuts that imitate the selection functions of several next-generation surveys. We then analyze many realizations of each mock survey to try to determine the uncertainty of the number density and two-point correlation function (see Sections 2.4.1 and 2.4.2) that will be measured.

In this work, we ignore the intricate details of survey geometries, overlapping pointings, and fiber collisions, which would require targeting strategies that are not yet finalized (however, our mock catalogs will be an extremely useful tool for running targeting simulations and analyzing the systematics they produce). We define our survey geometry by a square in angular coordinates and remove a random subsample to account for incompleteness primarily due to fiber collisions. Our two survey parameters are thus sky area and completeness fraction.

We implement this selection using $|\alpha| < \alpha_{\rm max}$ and $|\delta| < \delta_{\rm max}$ where $\alpha_{\rm max} = \delta_{\rm max}$. For small angles, the solid angle area is approximately $\Omega \approx 4\alpha_{\rm max}\delta_{\rm max}$ (in radians/steradians),

Table 2: Survey parameters

Name	Area (sq. deg)	Completeness	Redshift	Magnitude limits	References
WAVES	66	95%	$0.2 < z < 0.8$	$m_z < 21.25$	[36]
PFS ($z < 1$)	12	70%	$0.7 < z < 1.0$	$m_Y < 22.5$ & $m_J < 22.8$	[107]
PFS ($z > 1$)	12	70%	$1.0 < z < 1.7$	$m_J < 22.8$	[107]
MOONS ($z \sim 1$)	4	72.5%	$0.9 < z < 1.1$	$m_H < 23.0$	[73]
MOONS ($z \sim 1.5$)	4	72.5%	$1.2 < z < 1.7$	$m_H < 23.5$	[73]
MOONS ($z \sim 2$)	4	72.5%	$2.0 < z < 2.6$	$m_H < 24.0$	[73]

but to be precise, we calculate $\alpha_{\rm max}$ and $\delta_{\rm max}$ by inverting Equation 1.

$$\begin{aligned} \Omega &= \int_{-\alpha_{\rm max}}^{\alpha_{\rm max}} d\alpha \int_{-\delta_{\rm max}}^{\delta_{\rm max}} d\delta \cos(\delta) \\ &= 4(\alpha_{\rm max}) \sin(\delta_{\rm max}) \end{aligned} \tag{1}$$

To perform any type of scientific study on a sample of galaxies, its selection function must be well understood in terms of physical properties. By imposing the published magnitude limits (see Table 2) of the WAVES, PFS, and MOONS surveys on galaxies in our mock, we can test the fraction of galaxies at a given mass that is included in the selection function to test how well-represented they will be in the survey. For each survey, we show the 90% and 99% mass-completeness limit as a function of redshift in Figure 2. The color-coded bands in this figure enclose the three galaxy populations that we further analyze in this paper by calculating mock observables (Section 2.4) to constrain our HOD model (Section 2.3). The mass thresholds and effective redshifts of these samples are listed in Table 3. These cuts are almost entirely above the respective 99% mass-completeness limits, which means these surveys should observe representative samples.

Figure 2 additionally shows the comoving area probed by each survey as a function of redshift, in comparison to that of the Bolshoi-Planck simulation, which is a periodic cube of side length 250 h^{-1} Mpc. Note that WAVES reaches a slightly larger comoving area at the high-redshift end, which may cause the cosmic variance in our mocks to be slightly underestimated, due to the high probability of resampling the same galaxies across realizations. While this should not affect our primary conclusions, a more precise analysis of the cosmic variance in WAVES should use a larger simulation than Bolshoi-Planck.

Table 3: Galaxy samples

Name	Mass threshold ($M_\odot$)	Mass completeness	Redshift range	Effective redshift	Mean sample size
WAVES	10^{11}	99.903%	$0.5 < z < 0.8$	0.647	33,583
PFS	$10^{10.5}$	99.997%	$0.8 < z < 1.2$	0.979	61,307
MOONS	10^{10}	99.744%	$1.2 < z < 1.6$	1.367	41,661

2.3 HOD Formulation

2.3.1 The HOD

In this paper, we will predict the level of constraints that several upcoming surveys will place on the galaxy-halo connection. To quantify these constraints, we use the halo occupation distribution (HOD), which has been a standard way to measure the galaxy-halo connection in magnitude limited surveys for nearly two decades [10].

The HOD prescribes the mean number of galaxies above a mass or luminosity threshold per halo. This formalism is very popular due to its simplicity and utility for galaxy clustering predictions. We use the HOD parameter convention introduced by [127]. Under this formalism, we describe the expected number of central and satellite galaxies per halo above a stellar mass threshold, $M_{*\mathrm{thresh}}$, as

$$\langle N_{\mathrm{cen}} \rangle = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\log\left(M_h/M_{\mathrm{min}}\right)}{\sigma}\right)\right) \tag{2}$$

and

$$\langle N_{\mathrm{sat}} \rangle = \left(\frac{M_h - M_0}{M_1}\right)^{\alpha}, \tag{3}$$

where we do not assume any functional forms for the redshift and $M_{*\mathrm{thresh}}$ dependence of the free parameters M_{min}, σ, M_0, M_1, and α. Instead, we fit the HOD independently to each galaxy population of interest.

We plot these equations in Figure 3 using the fiducial parameters for our PFS sample and demonstrate how varying these parameters varies the number of galaxies per halo.

The parameters controlling $\langle N_{\mathrm{cen}} \rangle$ are the characteristic halo mass M_{min} and the characteristic spread σ. The parameters controlling $\langle N_{\mathrm{sat}} \rangle$ are the minimum halo mass M_0, the characteristic halo mass M_1, and the power-law slope α.

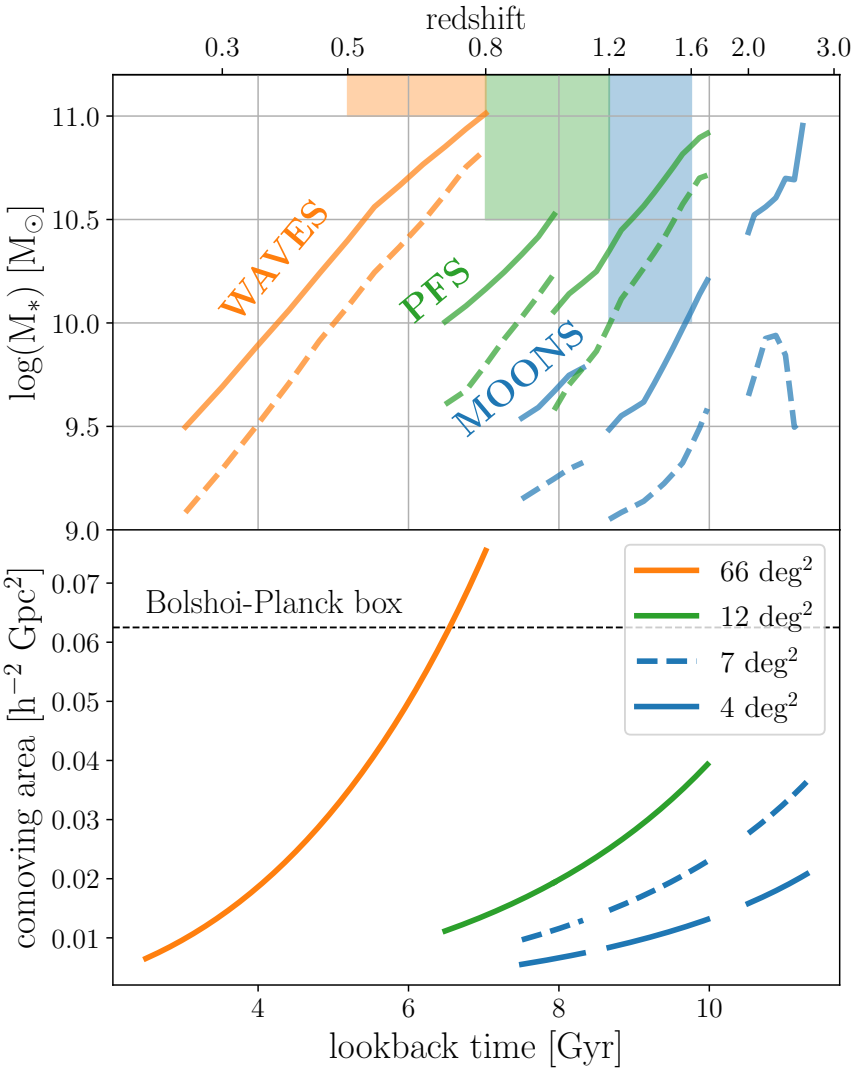

Figure 2: Mass completeness (upper panel) and field size (lower panel) for the targeting strategies (given in Table 2) of PFS, WAVES, and MOONS as a function of redshift. In the upper panel, we include 99% (solid lines) and 90% (dashed lines) completeness limits. These values are averaged over 25 mock catalog realizations. Color-coded bands indicate the mass-complete samples used in this analysis (see Table 3). In the lower panel, we plot the comoving area of each field, with sky areas taken from Table 2.

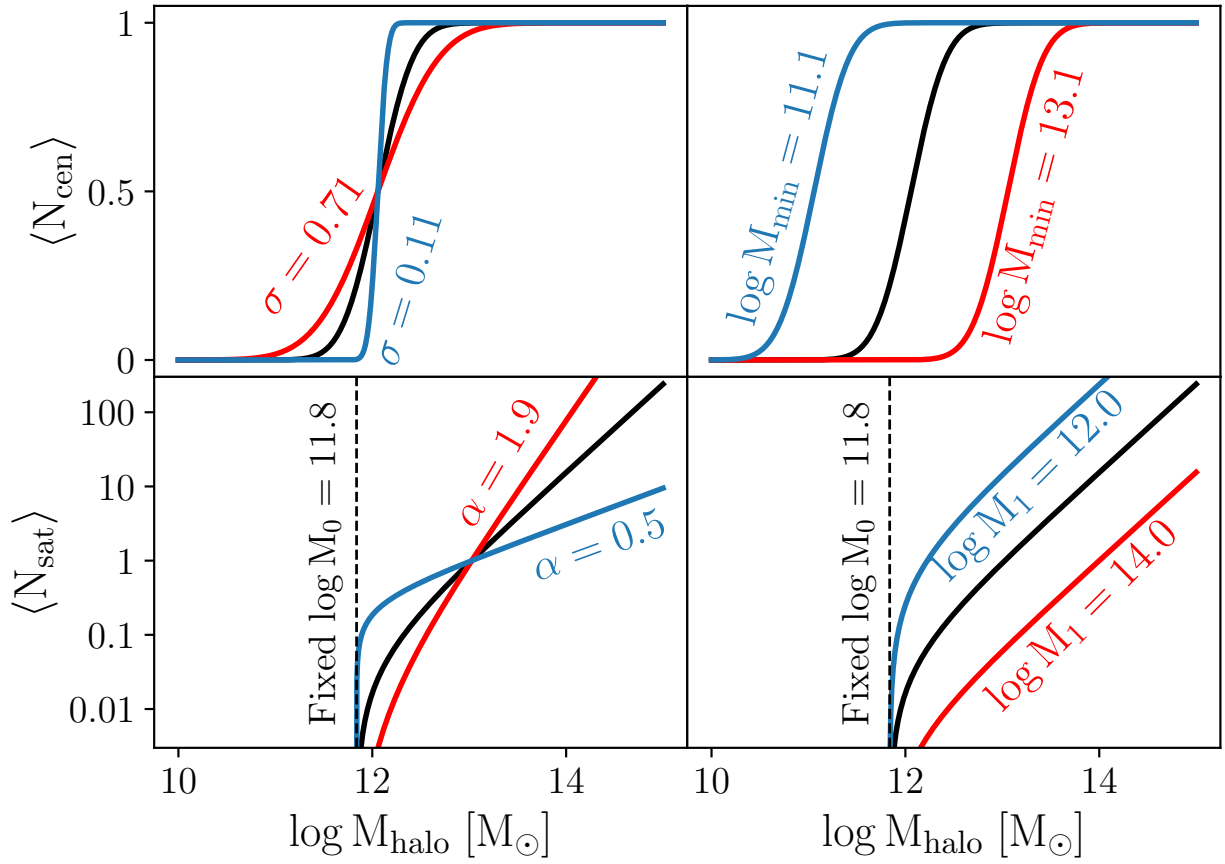

Figure 3: Mean occupation functions of centrals (Equation 2; top panels) and satellites (Equation 3; bottom panels) per halo in our HOD model. The total number of galaxies per halo is the sum of $N_{\rm cen}$ and $N_{\rm sat}$. Black curves are plotted with our fiducial set of parameters for our PFS galaxy sample. We demonstrate the effect of each model parameter by varying them one at a time, as labeled. Note that our conservative HOD model (Section 2.3.2) always maintains a constant $M_{\rm min}/M_1$ ratio, and conserves total number density by automatically updating $\log M_{\rm min}$ and $\log M_1$ to account for any change in σ or α.

Table 4: Fiducial HOD parameters (UniverseMachine "truths")

(Parameter type)	*(Fixed)*	*(Fixed)*	*(Free)*	*(Free)*	*(Derived)*	*(Derived)*	*(Fixed)*	*(Derived)*
WAVES	9.992×10^{-4}	5.229	0.736	1.291	12.956	13.674	12.176	0.190
PFS	4.838×10^{-3}	8.691	0.407	1.188	12.058	12.997	11.838	0.221
MOONS	7.619×10^{-3}	9.132	0.220	1.196	11.740	12.701	11.655	0.222

While the mean occupation function is a deterministic function of the HOD parameters, note that the number of galaxies assigned to each halo is stochastically drawn from a distribution around this mean. We draw from a Bernoulli distribution (1 or 0) for central galaxies and a Poisson distribution for satellites. Therefore, this induces some stochasticity in the number density and correlation function when populating a simulation of finite volume according to our HOD.

The HOD is constrained by quantities that probe the mass of the underlying halo population: abundance and clustering. To quantify clustering, we measure the two-point correlation function, $w_{\rm p}(r_{\rm p})$ (see Section 2.4.2). We quantify abundance with the number density of galaxies above the stellar mass threshold, n, which is related to the SMF (see Section 2.4.1) by

$$n = \int_{M_{*\rm thresh}}^{\infty} \Phi(M_*) dM_*. \tag{4}$$

The HOD can be calculated directly from the UniverseMachine by counting the average number of galaxies above the threshold in each halo. We fit Equations 2 and 3 to this calculation in narrow M_h bins to obtain fiducial HOD parameters for our WAVES, PFS, and MOONS galaxy samples. We list each fiducial HOD parameter, as well as the average number density n and satellite fraction $f_{\rm sat}$ in Table 4.

2.3.2 The Conservative HOD Model

Combining information from multiple sources, we expect very strong empirical constraints on the number density of most galaxy populations. Even in our mock surveys alone, we measure n to the precision of 1 to 4%, which is an order of magnitude smaller than the fractional error of $w_{\rm p}$ at large scales. Therefore, if allowed to freely vary, the number

density causes a near-degeneracy between the HOD parameters it is sensitive to, increasing the difficulty of calculating constraints through MCMC with little gain. We, therefore, set the fiducial value of n as a hard prior and only consider the HOD parameter-space that conserves the number density from the UniverseMachine. To do this, we integrate Equations 2 and 3 with the halo mass function $\Phi(M_h)$ to obtain

$$n_{\rm cen} = \int_0^\infty \langle N_{\rm cen} \rangle \Phi(M_h) dM_h \tag{5}$$

and

$$n_{\rm sat} = \int_0^\infty \langle N_{\rm sat} \rangle \Phi(M_h) dM_h, \tag{6}$$

where

$$n = n_{\rm cen} + n_{\rm sat}. \tag{7}$$

The parameter $M_{\rm min}$ primarily sets the number density for the centrals and M_1 for the satellites. Since we have removed one degree of freedom by holding the total number density fixed, we are free to combine these two parameters into a single parameter: $M_1/M_{\rm min}$. This ratio is directly influenced by the ratio of central to satellite dark matter halos predicted by dark matter simulations. Since this is decided by our cosmological prior, we choose to hold constant the $M_1/M_{\rm min}$ parameter measured from the UniverseMachine. Then, once a value is chosen for each free parameter, we can individually derive M_1 and $M_{\rm min}$ by numerically inverting Equation 7.

We remove another degree of freedom in our model by holding the fiducial value of M_0 fixed. This is common practice because the observables that we examine are not sensitive to large changes in this parameter. Therefore, we only tune two free parameters in our conservative HOD model: σ and α. We demonstrate the effect these parameters have on $w_{\rm p}(r_{\rm p})$ predictions in Figure 5. Increasing σ decreases clustering at all scales, while increasing α increases clustering, especially at small scales. Pushing to smaller-scale measurements will therefore be greatly beneficial in breaking the degeneracy of these parameters.

2.4 Constraints on the HOD

Following the standard methodology, we constrain the HOD by empirical measurements of number density and clustering of the galaxy population. Therefore, in this section, we present mock measurements of the stellar mass function and the projected two-point correlation function.

2.4.1 Stellar Mass Function

The stellar mass function (SMF), $\Phi(M_*)$, measures the number density of a galaxy population subdivided into bins of stellar mass. This is one of the most direct measurements of the efficiency of galaxy evolution, tracking the overall growth of galaxies over cosmic times from the accumulation of star formation. In terms of the galaxy-halo connection, the SMF provides an estimate for the SHMR, if we assume a good rank-order correlation between stellar and halo mass, as is done by abundance matching models.

Spectroscopic surveys like PFS, WAVES, and MOONS will improve stellar mass estimates of galaxies from the respective epochs they are probing due to the significantly increased precision via spectroscopic redshifts, as well as tighter constraints on mass-to-light ratios from stellar ages obtained by stellar population synthesis. This will substantially improve our certainty on the distribution of stellar masses [80], while abundance measurements will be further solidified by larger photometric surveys. Therefore, from this combination of data sources, we expect tight constraints on the SMF for a wide range of redshifts, especially at $z > 1$, in the coming years.

In each of our redshift samples, we measure the SMF over an ensemble of 25 mock survey realizations to quantify the uncertainty of a single survey. We present the mock SMFs of each survey in the bottom panel of Figure 4. Note that each mock SMF agrees with the truth down to the indicated mass threshold, which is a good sign that the survey samples will be representative of the true galaxy populations. Additionally, the upper panel of Figure 4 shows the satellite fraction as a function of stellar mass, which is also in good agreement down to the completeness limit for each sample.

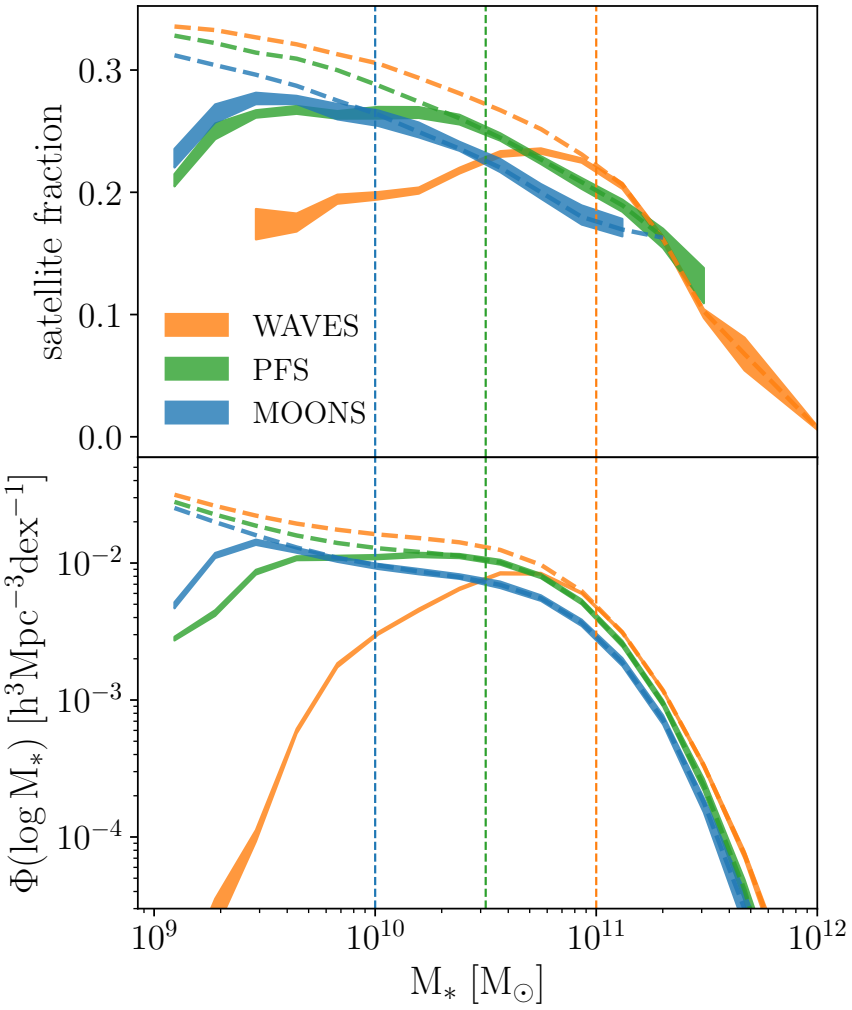

Figure 4: Mock measurements of the stellar mass function (bottom) and satellite fraction (top). We plot WAVES ($0.5 < z < 0.8$) in orange, PFS ($0.8 < z < 1.2$) in green, and MOONS ($1.2 < z < 1.6$) in blue. Each band represents the 1σ confidence region of each mock measurement, inferred by independent realizations (note that the satellite fraction here assumes perfect central/satellite assignment). Vertical dashed lines represent the stellar mass threshold we impose on each survey, while the thick dashed curves represent the true functions without photometric target selection. The observed functions agree very well with the true functions above the respective mass thresholds.

As discussed in Section 2.2.3, we are able to quantify the mass completeness of each sample by selecting all mock galaxies over the mass threshold and calculating the fraction of them which are under the survey's magnitude limits. Over the entire redshift range, each sample is well over 99% complete down to its mass threshold (see Table 3).

Note that the SMF is typically used as a direct constraint for the HOD through Equation 4. However, in our conservative model, the SMF is used as a hard prior because we do not allow the total number density of galaxies to vary (see Section 2.3.2).

2.4.2 Two-Point Correlation Function

The two-point correlation function, $\xi(r)$, is a canonical constraint on the HOD because it measures the clustering strength of the galaxy population, which is indicative of the clustering strength of the underlying halo population. Since halo clustering is a strong function of halo mass, the two-point correlation function is very sensitive to the typical mass of the halo population [122, 92, 113]. We adopt the projected correlation function, which is defined as

$$w_{\rm p}(r_{\rm p}) = 2 \int_0^{\pi_{\rm max}} \xi(r_{\rm p}, \pi) d\pi, \tag{8}$$

where we choose $\pi_{\rm max} = 50\ h^{-1}$ Mpc. The projected correlation function conveniently integrates out most of the dependence on redshift-space distortions. This is desired because our galaxy-halo connection model has no dependence on velocity dispersion, and we do not want our observable to be sensitive to that level of detail.

We perform this computation in six $r_{\rm p}$ bins with logarithmically spaced edges from 1-27 h^{-1} Mpc. Using a relatively small number of bins here helps reduce the number of realizations needed to calculate the covariance matrix, which must be very precise for our analysis. Due to the systematic sampling caused by fiber collisions in multiplexed spectroscopic surveys, we don't attempt to calculate the two-point correlation function below a scale of 1 h^{-1} Mpc (i.e., 106 arcsec at $z = 0.8$, 78.7 arcsec at $z = 1.2$, and 65.2 arcsec at $z = 1.6$). The fiber positioner patrol diameters of the instruments used by WAVES, PFS, and MOONS will likely be similarly sized, so fiber collisions should not dominate our uncertainty and the effect will be mostly mitigated by revisiting fields multiple times.

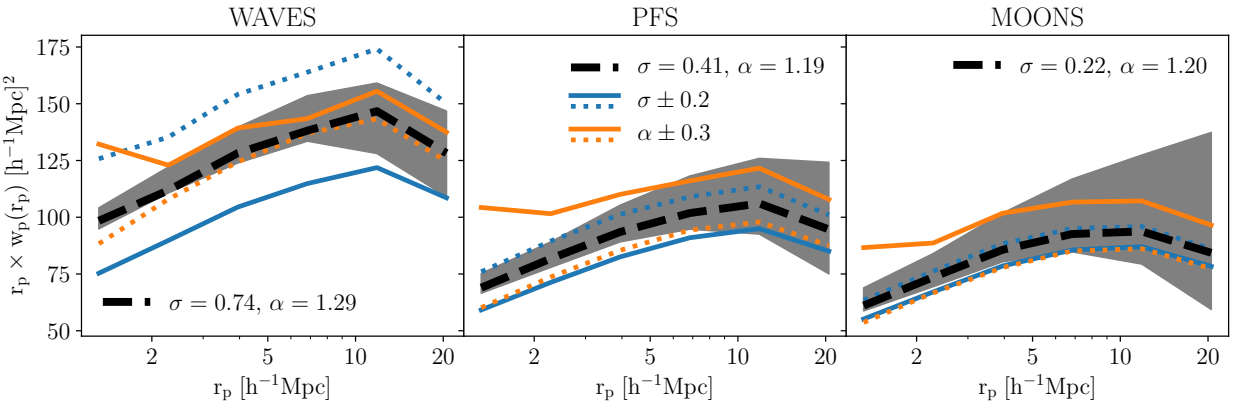

Figure 5: The projected two-point correlation function. Mock 1σ constraints from each survey are given by grey shaded regions. The UniverseMachine "truth" HOD model is represented by the thick black dashed line, while the colored solid and dotted lines show the effect of increasing or decreasing one parameter at a time, respectively. High-mass samples like WAVES are much more sensitive to changes in σ, and low-mass samples like MOONS are more sensitive to changes in α. These parameters produce similar observational effects, but this degeneracy can be broken by probing smaller scales.

We calculate the projected two-point correlation function using the Landy-Szalay [65] estimator implemented in the `DDrppi`, `DDrppi_mocks`, and `convert_rp_pi_counts_to_wp` functions from the `Corrfunc` package [102]. We measure $w_{\rm p}(r_{\rm p})$ from each mock survey in Figure 5, and compare it to various predictions of our conservative HOD model (see Section 2.3.2). These measurements are sensitive to fairly small variations in σ or α, as can be seen in this figure. However, note that typically higher mass samples (e.g., WAVES) are less sensitive to α and lower mass samples (e.g., MOONS) are less sensitive to σ. This is because halo bias increases more rapidly as a function of mass at higher masses than lower masses, causing variations in high mass halo occupation to be more sensitive to the two-point correlation function. Conversely, at lower masses, there is less sensitivity to clustering signals except by varying the number of satellites, which dominate the two-point correlation function.

Our model predictions for the projected correlation function are calculated using a periodic box, for which we use the Bolshoi-Planck snapshot whose redshift is closest to the effective redshift for the given sample, as listed in Table 3. We weight our effective redshift (Equation 9) by pair counts, which scales with number times number density (Equation 10).

$$z_{\rm eff} = \frac{\int_{z_{\rm min}}^{z_{\rm max}} zW(z)dz}{\int_{z_{\rm min}}^{z_{\rm max}} W(z)dz} \tag{9}$$

where

$$W(z) = (dN/dz)n = \frac{(dN/dz)^2}{dV/dz} \tag{10}$$

2.4.3 MCMC Fits

We perform the measurement of $w_{\rm p}(r_{\rm p})$ (see Section 2.4.2) on 600 independent realizations of each mock survey, seeded by randomized orientations and origins in the Bolshoi-Planck box, using the `lightcone` code provided in the UniverseMachine package. We then calculate the mean and covariance matrix from these samples and define a six-dimensional multivariate normal likelihood distribution for our six $r_{\rm p}$ bins of $w_{\rm p}$. We then use the `emcee` package to sample the posterior probability distribution of our HOD parameter-space: $\{\sigma, \alpha\}$, with a uniform prior confined to $10^{-5} < \sigma < 5$ and $0.1 < \alpha < 3$. We initialize our

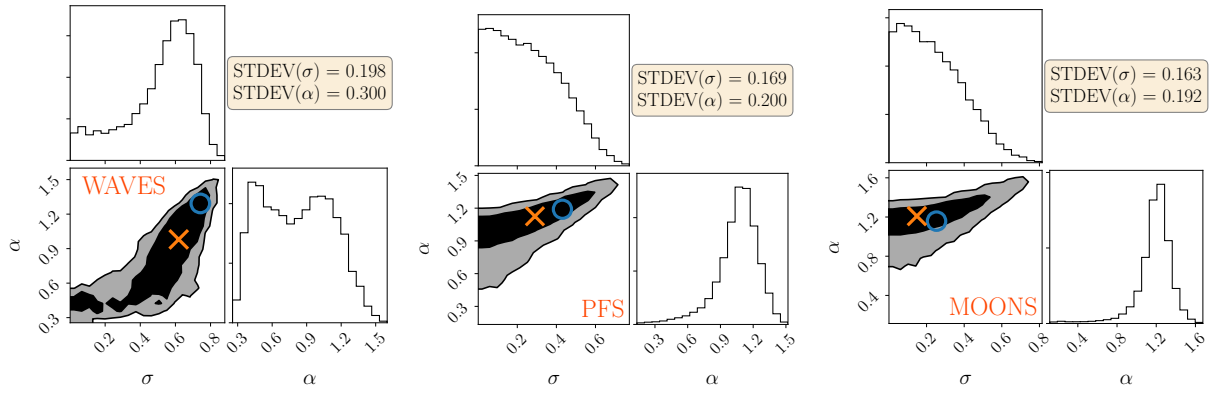

Figure 6: HOD posterior probability distribution measured in WAVES (left), PFS (center), and MOONS (right). Measured by MCMC sampling of our two-parameter conservative HOD model at the effective redshift of each sample, and comparing the predicted $w_{\rm p}(r_{\rm p})$ to 600 mock realizations. For each sample, the UniverseMachine truth value is marked with a blue circle and best-fit parameters are marked with an orange X.

MCMC chains very close to the corresponding fiducial parameters given in Table 4, but allow them to run many autocorrelation lengths to ensure they are well converged, as discussed in Section 2.5.1.

Note that the $w_{\rm p}(r_{\rm p})$ measured by a survey could be obtained more realistically by drawing one of the 600 realizations, rather than using the mean. However, from tests of additional MCMC runs, we confirm that there is no strong bias in the constraining power from using individual realizations instead of using the mean value. Therefore, we define our likelihood using the mean, which is more stable and the only fair comparison between measurements using various survey parameters. Note that cosmic variance and measurement error is still incorporated through the covariance matrix.

2.5 Results: Predictions for Next-Generation Surveys

2.5.1 Forecasts for WAVES, PFS, and MOONS

Thanks to the small number of free parameters in our model, we obtain favorably high acceptance rates ($\sim 50\%$) and low autocorrelation lengths (~ 50), which helps reduce the time required to run our MCMC chains. To ensure we are not biased by our initial guess, we removed a burn-in of 250 iterations from the beginning of each chain, although this has a very small effect due to the long length of our chains. In each MCMC, we sample 150,000 trial points, yielding posteriors of very high resolution. We present a corner plot of the posterior measured in each mock survey in Figure 6. These posteriors show that our method does a good job of constraining our HOD to a fairly small region in parameter space. The largest difficulty is constraining the α parameter in the WAVES sample due to the high mass centrals dominating both the number density and clustering signal, and a large covariance between σ and α.

Putting together the information from the posteriors from WAVES, PFS, and MOONS, we will be able to constrain the HOD across a wide range of mass and redshift. We compile the predicted constraints on the evolution of these HOD parameters in Figure 7. Certain parameters will still be very poorly constrained using this type of analysis; for example, α in the WAVES sample. This is primarily because the two-point correlation function is most sensitive to σ at high masses and α at low masses, but additional metrics may be able to provide more information (see Section 2.8).

Note that the WAVES sample produces very poor constraints on α (the satellite occupation slope parameter), whereas the MOONS sample produces particularly strong constraints on α. This demonstrates a fundamental difficulty of using constraints only from number density and the two-point correlation function. Since the two-point correlation function is primarily sensitive to the most clustered data, it is more informative for satellites at lower mass thresholds and centrals at high mass thresholds (as seen in Figure 5). However, for the WAVES sample, note that σ and α are nearly degenerate, which results in relatively poor constraints for both parameters.

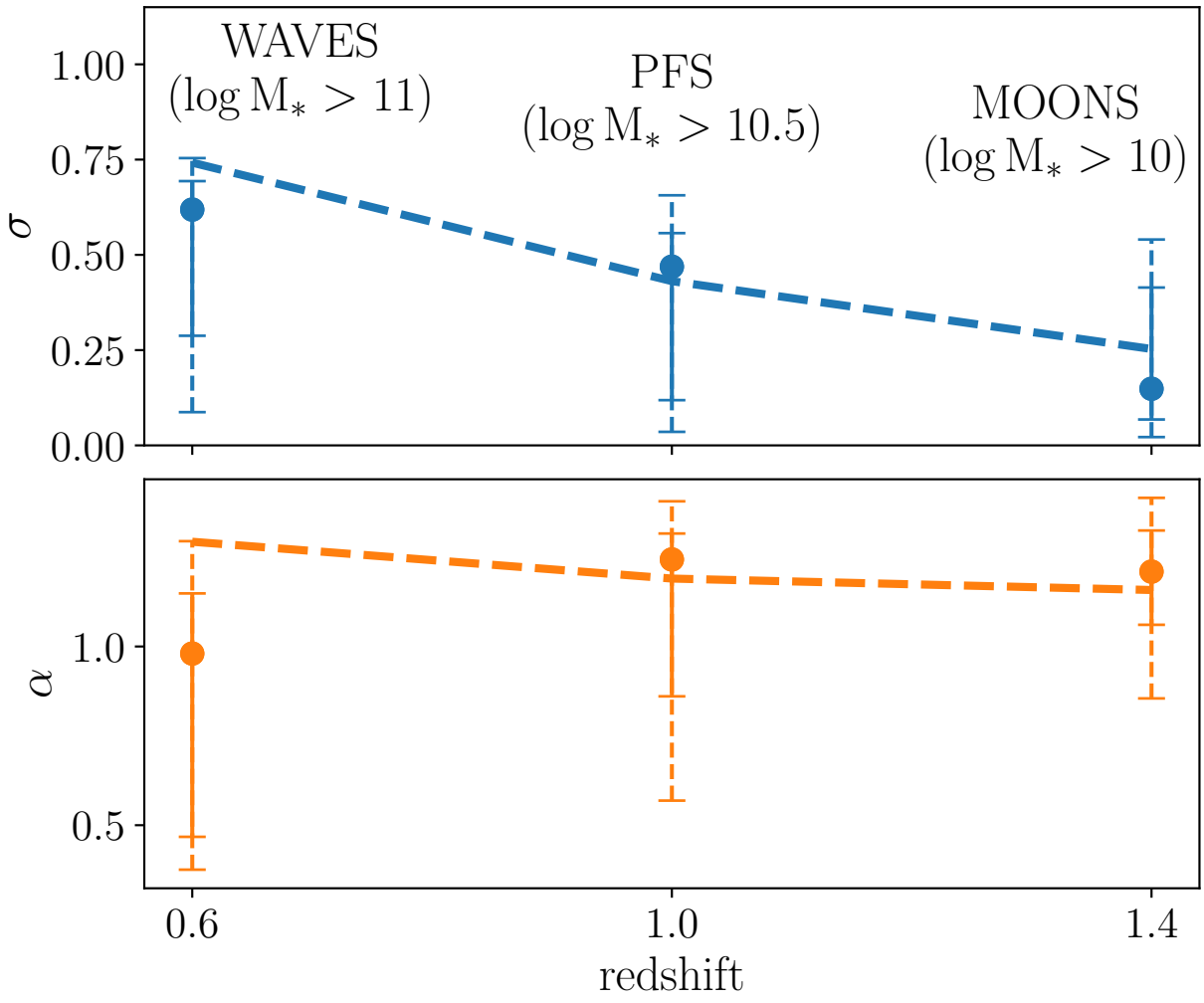

Figure 7: Predicted constraints on the evolution of the HOD. We compile best fit measurements along with 5th, 16th, 84th, and 95th percentile of the posterior (see Figure 6) of σ (upper panel) and α (lower panel). We use a prior of $10^{-5} < \sigma < 5$, so MOONS provides essentially no constraints at all on σ. UniverseMachine "truth" values are connected by dashed lines.

It should be noted that improving our techniques may lead to tighter constraints on these regions of difficulty. First of all, it is possible to partially break the degeneracy between central and satellite clustering by measuring the two-point correlation function to smaller scales (sub-Mpc) using fiber collision corrections. Additionally, alternative statistics like counts-in-cylinders tend to be more sensitive to certain galaxy populations, and therefore provide excellent complementary information to the two-point correlation function [113]. This will be important for analyzing the real data, but will come at a greater computational cost, particularly because this increases the size of the covariance matrix, and will likely require many more mock realizations to calculate accurately.

2.5.2 Measurement Error vs. Survey Parameters

Telescope time is typically the limiting factor in survey design. To first order, the amount of time a survey requires is roughly proportional to the number of objects observed. It is therefore possible to either scale up the sky area in exchange for a decrease in completeness fraction or vice versa. Increasing the sky area of a survey would increase the volume and therefore decrease the cosmic sample variance; on the other hand, high completeness helps mitigate the uncertainty of small-scale pair counts, especially when accounting for fiber collisions [14]. Although this is not tested in this work, it should also be noted that higher completeness fractions should be favored if the goal is to increase the accuracy of identifying central galaxies in group reconstruction [71].

Using our mock catalogs to estimate uncertainties, we vary these survey parameters to quantify their effects on constraining power. In Figure 8, we present the dependence of survey parameters on the uncertainty of the projected two-point correlation function. For PFS- and MOONS-like surveys, the only way to greatly reduce uncertainties is by increasing the area (see top row). Increasing the number of targets in the same area makes much more modest improvements (see bottom row). In other words, smaller area surveys like PFS and MOONS are dominated by cosmic variance, not shot noise. Wider surveys like WAVES (left panel) typically find that the two-point correlation function uncertainty is dominated by cosmic variance on large scales (9-27 h^{-1} Mpc) and shot noise on small scales (1-3 h^{-1} Mpc).

To roughly quantify the effect survey parameters have on correlation function uncertainties, we calculate the percent decrease in $w_{\rm p}$ error at $r_{\rm p}$ = 3-9 h^{-1} Mpc from the true survey parameters for two cases: (1) doubled area and half the completeness (conserving the number of targets) and (2) doubled completeness and the same area. Since we can't actually calculate double the area and double the number of targets for each survey, we simply use linear regression to fit the slope of the orange lines in Figure 8 to extrapolate these numbers. We find that doubling the survey area decreases the uncertainty of the correlation function at intermediate scales by 3%, 24%, and 22% for our WAVES, PFS, and MOONS samples, respectively. For a fixed survey area, doubling the number of targets improves the same constraints by 14%, 5%, and 2%. These numbers reinforce that the samples probed by PFS and MOONS are dominated by cosmic variance and can only be improved significantly by increasing the observing area.

Given the covariance matrix of $w_{\rm p}(r_{\rm p})$ calculated for each set of survey parameters, we also calculate HOD constraints via our MCMC method. We present the HOD constraining power as a function of survey parameters by varying completeness fraction and area in Figure 9. In the WAVES survey, we only find very small changes in constraints as we vary the survey parameters. For PFS and MOONS, the difference between Figures 9 and 8 indicates that the covariance of the two-point correlation function with respect to various regions of the universe is significantly different from the covariance due to varying HOD parameters. It appears that this allows the HOD to be somewhat more robust to cosmic variance than $w_{\rm p}$. Throughout, the constraints on α may be slightly more dependent on survey parameters than the constraints on σ.

2.5.3 Comparisons to Past Surveys

Surveys like WAVES, PFS, and MOONS will be monumental in pushing measurements of the two-point correlation function to higher redshifts because they will provide the precise spectroscopic measurements necessary to perform those calculations. There are currently no existing spectroscopic datasets that are comparable in size to the $z > 1$ samples we will obtain from PFS and MOONS.

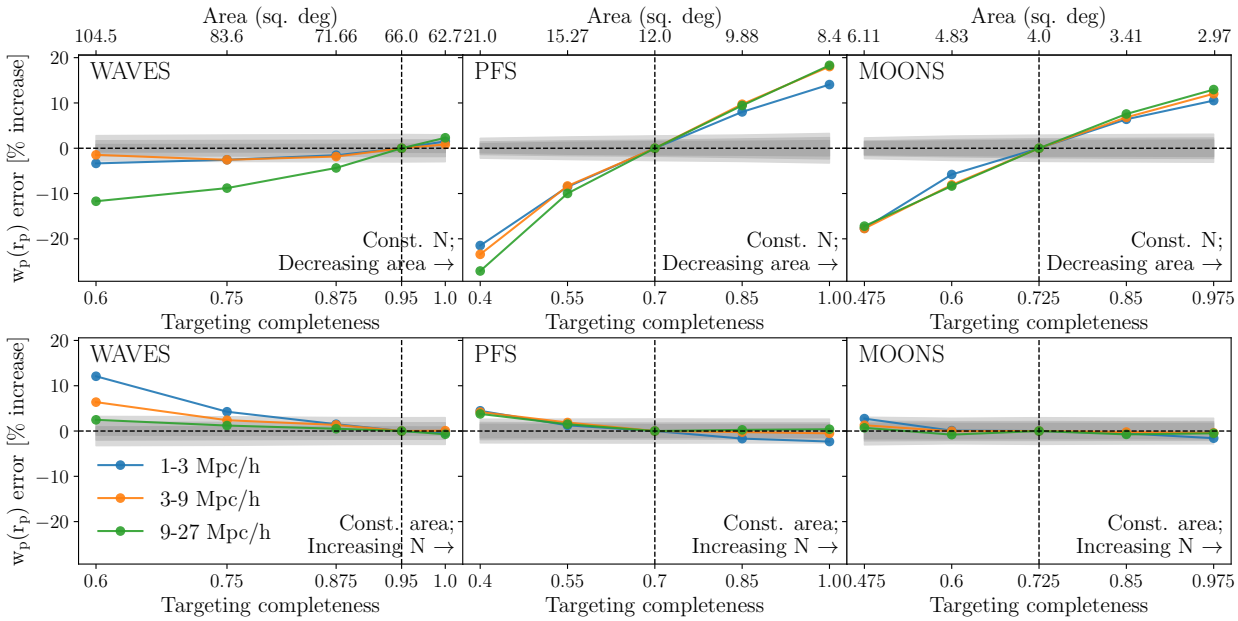

Figure 8: Precision of mock w_p measurements, color-coded by r_p scale, as a function of completeness fraction. Characteristic jackknife uncertainties are shown with grey bands. In the top panels, the sky area is varied to conserve the total number of targets (given in Table 3). In the bottom panels, the sky area is conserved at the true value for each survey, and therefore the number of targets increases with completeness. Particularly for the PFS and MOONS samples, there are significantly stronger trends in the top panels. This suggests that the uncertainties in $w_p(r_p)$ are dominated by cosmic variance, as opposed to shot noise.

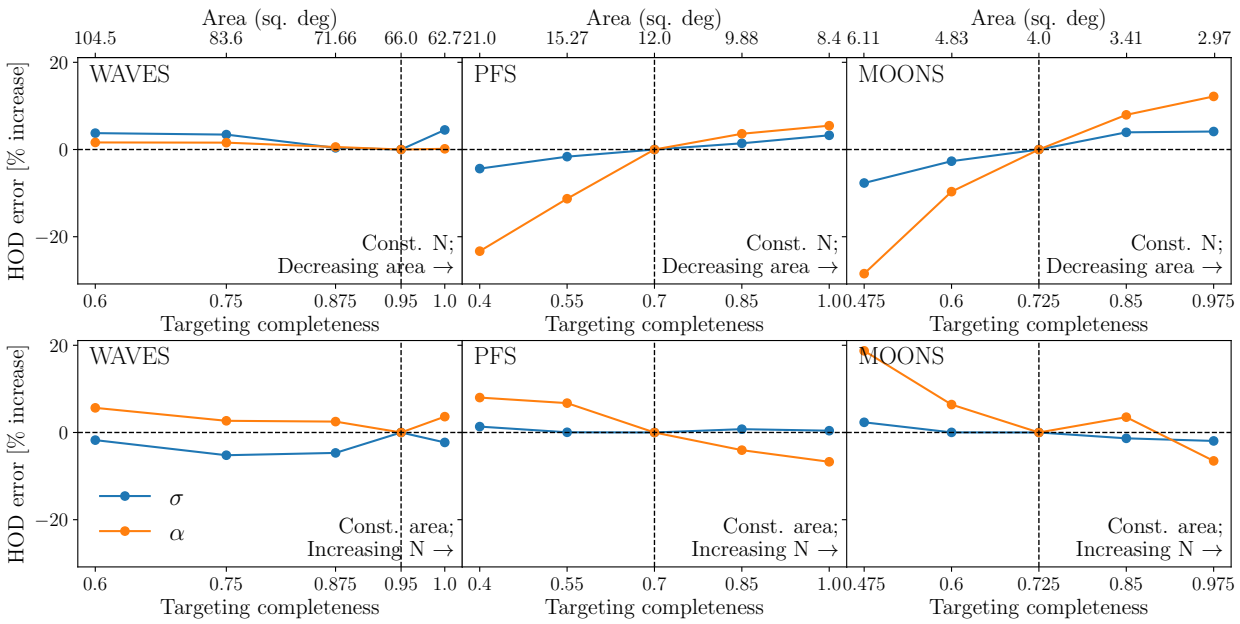

Figure 9: Precision of HOD parameters σ (blue) and α (orange), as a function of completeness fraction. The panels are arranged analogously to Figure 8. These uncertainties do not appear to be dominated by cosmic variance as strongly as the two-point correlation function. They are affected just as strongly by shot noise. This demonstrates the importance of propagating uncertainties from the full covariance matrix, rather than just the diagonal components shown in Figure 8.

For the slightly lower redshifts probed by WAVES, the closest comparison would be prism surveys such as the PRIsm MUlti-object Survey (PRIMUS [24]) or the Carnegie-Spitzer-IMACS (CSI [59]). PRIMUS, the larger of these two surveys, has measured the spectra of galaxies out to $z \sim 1.2$ with a redshift precision of $\sigma_z/(1+z) \sim 0.005$. Incorporating all fields that overlap with imaging from the Galaxy Evolution Explorer (GALEX [74]), Spitzer Space Telescope [115], Infrared Array Camera (IRAC [39]), and various ground-based surveys, PRIMUS has an area of 5.5 deg^2 and is complete down to similar stellar masses as WAVES (see [78]).

We compare mock measurements of the projected two-point correlation function of our WAVES galaxy sample ($0.5 < z < 0.8$ and $\log(M_*/M_\odot) > 11$) for the survey parameters of WAVES and PRIMUS in Figure 10. This plot illustrates the significantly increased precision we can expect to obtain from this sample of galaxies. Additionally, unlike PRIMUS, we assume that the redshift uncertainties in WAVES will be negligible compared to redshift distortions. The additional redshift error from PRIMUS does not greatly contribute to the errorbars of the two-point correlation function, but this does produce a small systematic offset which may further reduce the correlation function's sensitivity to HOD parameters.

However, most of our current understanding of the galaxy-halo connection comes from studies of surveys that either span lower redshifts or rely on photometric redshifts. For example, [128] use photometric redshifts from SDSS [16] and [66] from COSMOS [101]. In Figure 11, we present key findings of these past studies in terms of the SHMR, absolute bias, and satellite fraction of several stellar mass threshold galaxy samples and compare to the same measurements derived from our HOD projections of the WAVES, PFS, and MOONS surveys. These upcoming surveys will push to significantly deeper redshifts than past surveys, with comparable uncertainties.

Measuring the HOD for various surveys at distinct redshifts and stellar mass thresholds has been and will continue to be a powerful tool for studying galaxy evolution as measured by mean halo mass, satellite fraction, and galaxy bias (see Figure 11). Currently, we have little evidence for significant redshift evolution in most of these properties. However, more precise and higher redshift measurements of these parameters will give us a much clearer picture of how they may evolve with the age of the universe. Each of these metrics is highly

sensitive to various models of galaxy formation, so these new measurements will have a large impact on how we think about the important processes driving the shut down of rapid star formation that occurred at cosmic noon.

2.6 CLIMBER Conclusions

In this paper, we present the CLIMBER procedure, which we use to calibrate photometry into the UniversemMachine and similar models. This procedure performs well at reproducing a broad range of properties simultaneously. Twenty-five realizations of the $0.7 < z < 1.7$ mock catalog used for the PFS and MOONS samples in this work are available at `https://alanpearl.github.io/#data`. Alternatively, you may install the utilities we used to construct these catalogs at `https://github.com/AlanPearl/mocksurvey`.

We have used our mock catalogs to test the forthcoming generation of massively multiplexed spectroscopic galaxy surveys, which will likely change our understanding of galaxy formation, the galaxy-halo connection, and possibly even cosmology in profound ways. The high-redshift samples being probed will provide new constraints on theories and models which predict how populations evolve with redshift. The UniverseMachine will face new scrutiny in its ability to reproduce clustering and environmental quenching signals in the distant universe. The constraints on parameters of interest in the UniverseMachine will likely tighten significantly to match the new observations. It may even be possible that the UniverseMachine will require reparameterization in order to achieve a good fit to the new data. Either way, analysis of these new surveys will greatly impact our understanding of the evolution of the galaxy population's connection to its dark matter environment over the history of the universe.

Using the two-point correlation function, we have found that surveys such as WAVES, PFS, and MOONS will place new constraints on the galaxy-halo connection. We characterize constraints on the central term of the HOD with the parameters σ and $\log M_{\rm min}$. The precision to which we measure these parameters is displayed in Figure 6. We have found that studies of lower mass galaxies, like the MOONS sample, will not achieve strong constraints

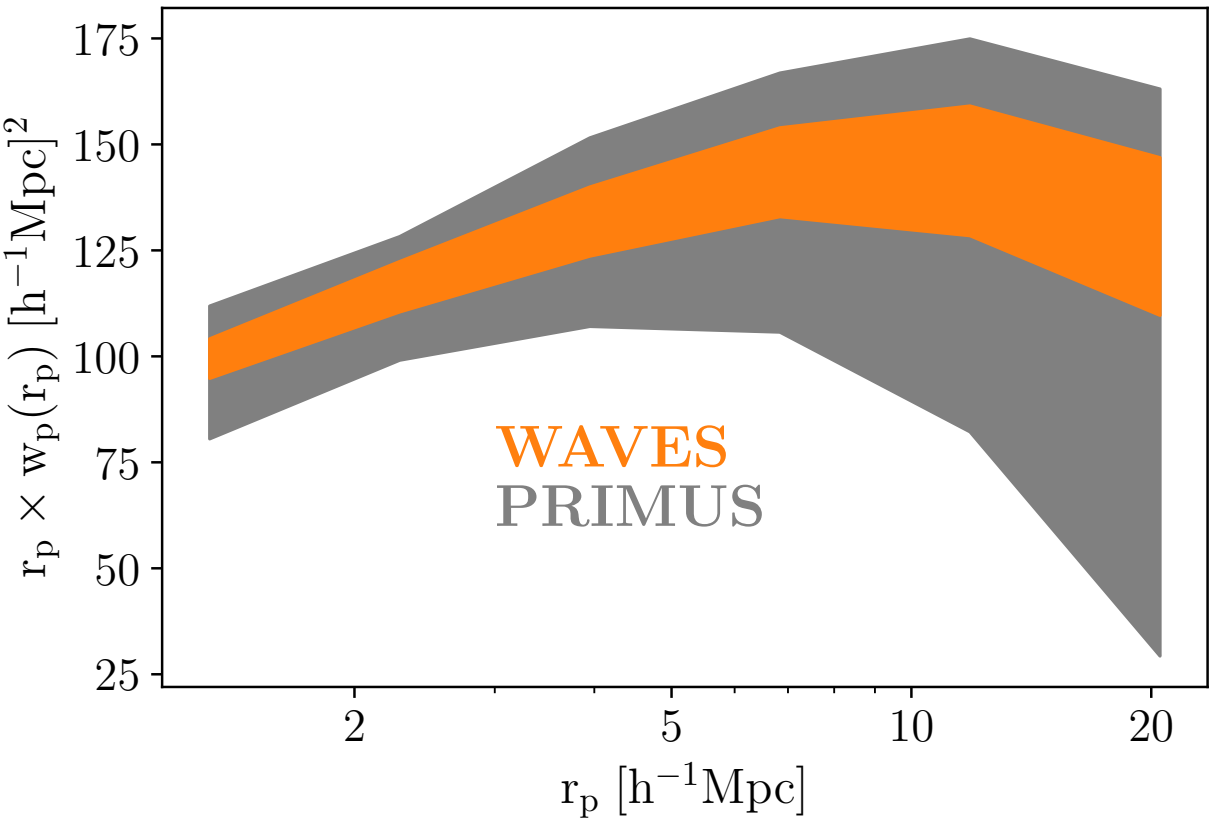

Figure 10: Mock measurements of the projected two-point correlation function for WAVES vs. PRIMUS. Note that we only used 25 independent realizations to quantify the observational error of the PRIMUS measurement, compared to the 600 used for WAVES. The increased precision from WAVES is due to its area (66 deg^2) which is over 10 times that of PRIMUS (5.5 deg^2). A small systematic offset is driven by redshift errors in PRIMUS, which we assume will be negligible compared to velocity distortions in WAVES.

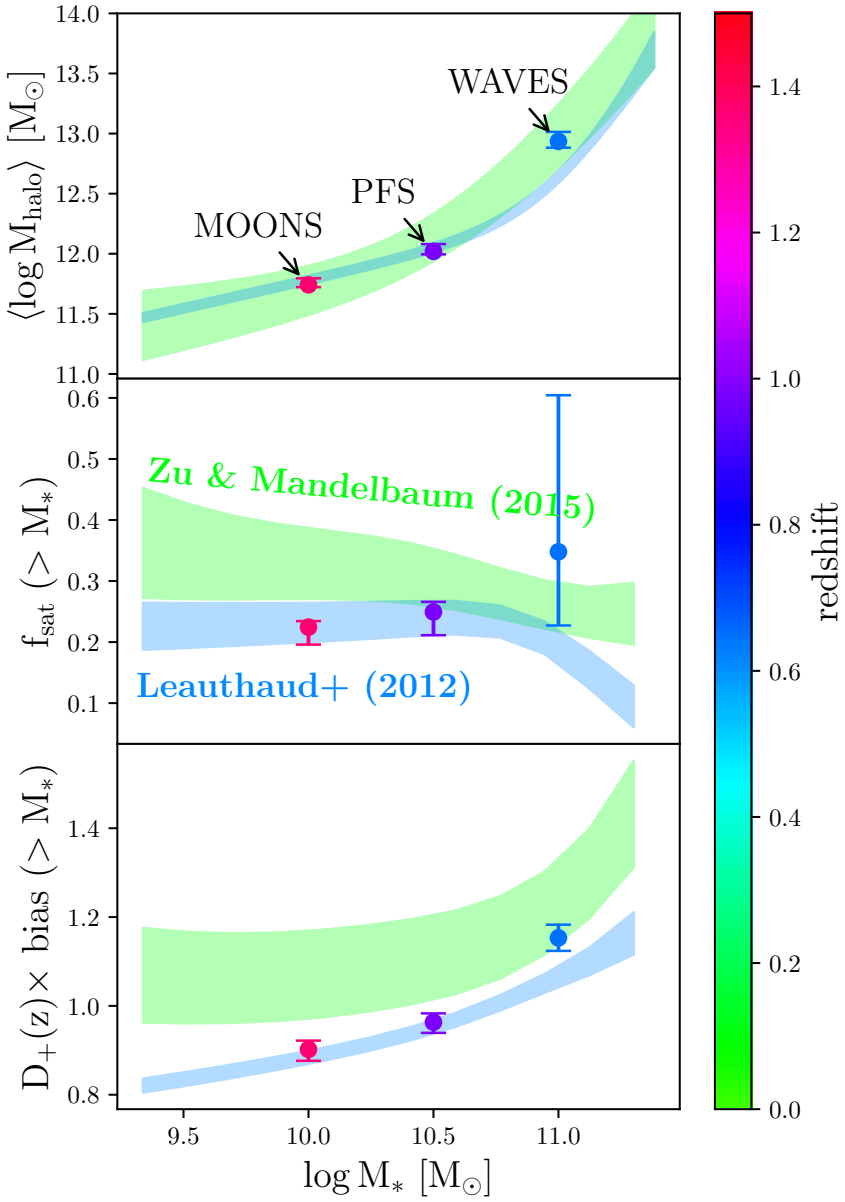

Figure 11: Compiled measurements of the galaxy-halo connection. As a function of stellar mass, we show several studies that have measured the mean halo mass (top panel), satellite fraction above the stellar mass threshold (middle panel), and absolute bias above the stellar mass threshold (bottom panel). Projected HOD analysis from this work for WAVES, PFS, and MOONS is shown by the colored points, which push to significantly higher redshifts with comparable uncertainties to previous studies.

from $w_{\rm p}(r_{\rm p})$ alone and will likely need to use other counts-in-cells statistics that are more sensitive to the weak clustering signal of low-mass centrals.

We characterize constraints on the satellite term of the HOD with the parameters α and $\log M_1$. The precision of these measurements is also displayed in Figure 6. We find that the satellite term of the HOD will be most poorly constrained in high-mass samples, where the clustering signal is dominated by centrals. In this regime, small-scale correlation function measurements are the most sensitive to the satellite occupation of halos. Smaller-scale measurements will be possible from these surveys using fiber collision corrections, but those will likely come at the cost of large shot noise. Therefore, it is still important for follow-up surveys to improve the completeness of these galaxy samples and reduce the effect of fiber collisions.

The achievable constraining power of these parameters is dependent on survey parameters, such as completeness (which reduces shot noise) and area (which reduces cosmic variance). Our conclusions can be summarized by the following key points:

- The two-point correlation function measurements from PFS and MOONS are both primarily dominated by cosmic variance, rather than shot noise. We have shown that, with fixed sample size, increasing their survey area drastically reduces this uncertainty.
- The HOD constraints from PFS and MOONS are less dominated by cosmic variance. This demonstrates the importance of using the full covariance matrix to calculate HOD constraints.
- From WAVES, there is a more balanced combination of shot noise, which is dominant on small scales ($1 - 3\ h^{-1}$ Mpc) and cosmic variance, which is dominant on large scales ($9 - 27\ h^{-1}$ Mpc). The resulting HOD constraints are not strongly affected by small changes in survey parameters.

Another important survey parameter, which has not been explored in this work, is the number of independent fields. PFS and MOONS are both planning on dividing their survey into several fields, which will slightly mitigate some of their large cosmic variance. Additionally, our predicted future constraints could be improved by supplementing the correlation function with counts-in-cylinders, which is more sensitive to the weak clustering of low-mass

centrals. Simulating detailed targeting strategies may also be important for more precise optimizations of constraining power, as this is necessary to calculate pair counts corrections at very small scales due to fiber collisions. This is important for unbiased estimates of both the two-point correlation function and counts-in-cylinders.

There are sure to be many discoveries in store thanks to this new generation of surveys. This new data will be investigated from all angles of the galaxy-halo connection: empirical models, SAMs, and hydrodynamic simulations alike will be put to the test. We hope that through this combined effort and publicly available tools like our mock catalogs, we will be able to utilize this data to its full potential.

2.7 CLIMBER Appendix - CLIMBER Details

The goal of Calibrating Light: Illuminating Mocks By Empirical Relations (CLIMBER) is to estimate the luminosity of each mock galaxy in any observed photometric band. Since star-forming galaxies host more young blue stars, color is a smooth function of specific SFR (sSFR). Both of these quantities correlate strongly with the mass-to-light ratio, as shown by [9]. The relationship between mass-to-light ratio in each band and sSFR is approximately a power law that can be empirically calibrated.

However, we first need to ensure self-consistency between the model and empirical parameters. In the model, SFRs for the star-forming population are drawn from mass-dependent distribution (the star-forming main sequence) which evolves with redshift. The star-forming main sequence is matched to empirical distributions at similar redshifts to ensure the values are physical. However, for the quiescent population, the SFRs are drawn from a non-evolving log-normal distribution centered around an sSFR of $10^{-11.8}$ yr^{-1}. Given that this is inconsistent with empirical assumptions past $z \sim 0$, we adopt only their rank-ordering. For each mock galaxy, we first map its sSFR to an empirically calibrated value of ultraviolet sSFR (sSFR$_{\rm UV}$). We choose sSFR$_{\rm UV}$ due to its tight correlation with mass-to-light ratio and reliable measurements in both quiescent and star-forming galaxies.

It would be even better to use the total sSFR by adding infrared (IR) sSFR, but this is not

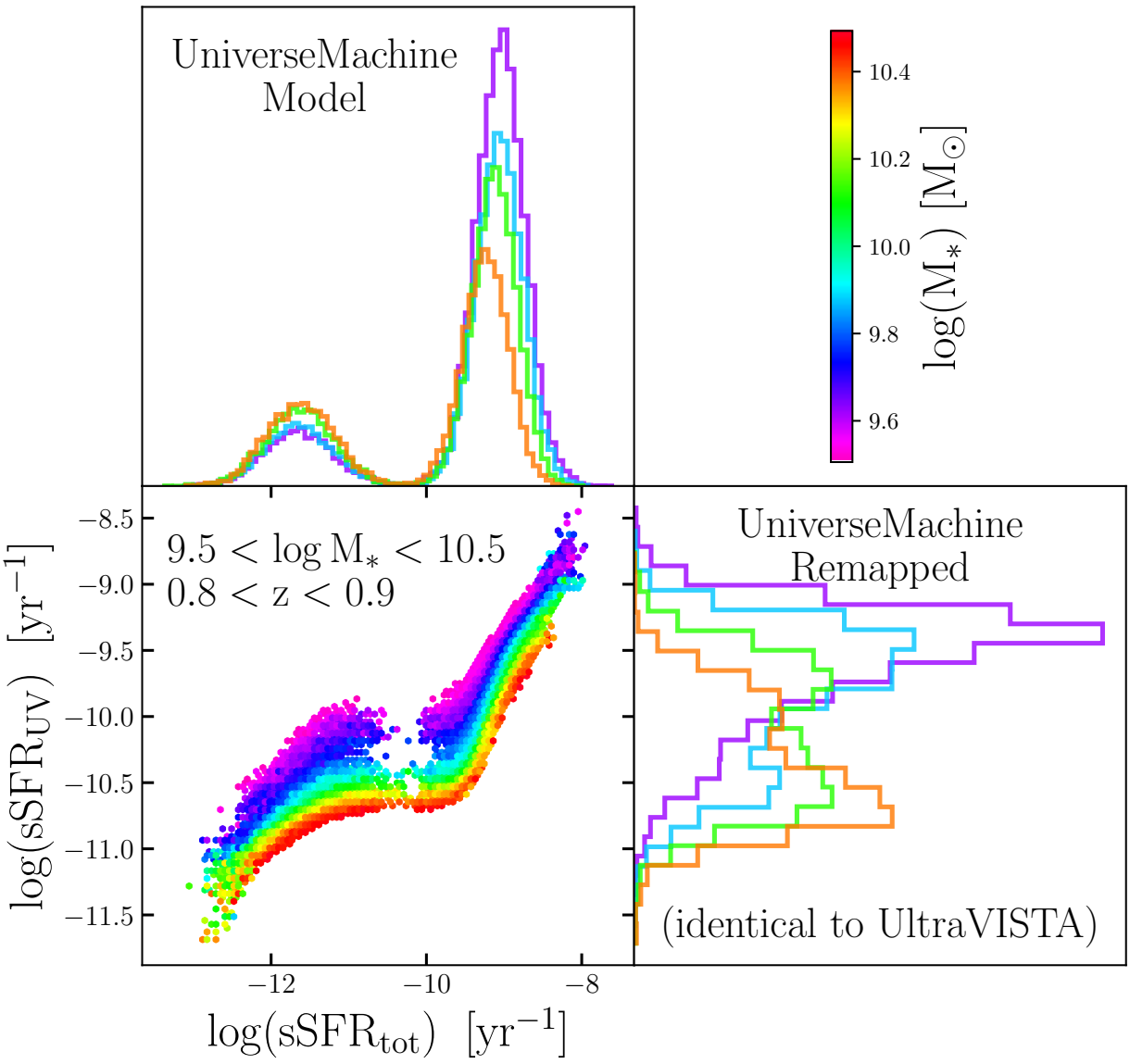

Figure 12: Conditional abundance-matching mapping from $\mathrm{sSFR} \rightarrow \mathrm{sSFR_{UV}}$ for UniverseMachine mock galaxies for a range of stellar masses at the redshift slice $0.8 < z < 0.9$. After being remapped, the distribution is forced to be identical to that of UltraVISTA at fixed stellar mass and redshift. Note that there is a near one-to-one mapping for star-forming galaxies, but the very low sSFRs of quiescent UniverseMachine galaxies are shifted up significantly.

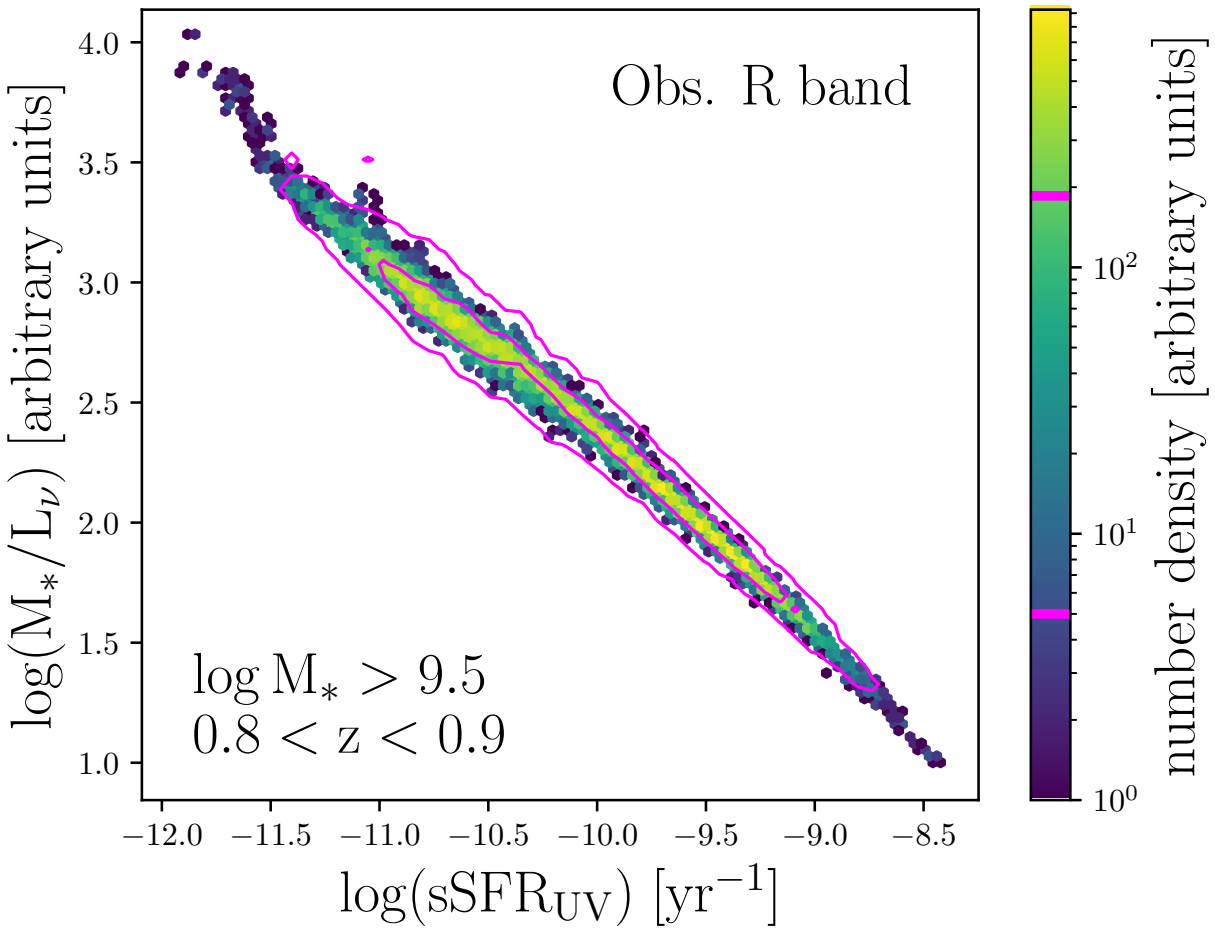

Figure 13: The relation between UV specific star formation rate and mass-to-light ratio in the *observed* R band at the redshift slice $0.8 < z < 0.9$. Magenta lines represent logarithmic contours of the UltraVISTA training data, while the colored points represent logarithmic counts of UniverseMachine mock galaxies which were fit via the Random Forest method described in Section 2.7. We present the same relation in all other available photometric bands in Figure 16.

possible for the UltraVISTA dataset because 53% of our training data have no detection in the IR. This creates an artificial discontinuity between IR detections and non-detections, thereby forming two distinct populations in the sSFR-M/L plane (see Figure 14). Including this discontinuity would greatly increase the difficulty in mapping from sSFR to M/L, particularly because it would require knowing the stellar mass to know which population a galaxy is a part of (this is primarily an observational effect in which lower mass galaxies at the same sSFR are less likely to have infrared detections). Including the stellar mass as a feature in our random forest is undesirable because it could have unknown consequences when extrapolating to lower stellar masses. Due to our decision to map sSFR to $\mathrm{sSFR_{UV}}$, we warn that our mocks may underpredict the spread in the distribution of M/L at a given sSFR. However, we believe that this is the best option due to its continuity and accuracy in reproducing magnitude and color distributions.

The mapping of $\mathrm{sSFR} \to \mathrm{sSFR_{UV}}$ is not uniform because UV flux decreases with higher dust obscuration, which is strongly dependent on stellar mass [118]. Therefore, we map $\mathrm{sSFR} \to \mathrm{sSFR_{UV}}$ through conditional abundance-matching (CAM) using the `halotools` package, which preserves the rank-ordering at fixed stellar mass (tolerance of $\sim$0.05 dex). We iterate this method in fuzzy redshift bins (width of $\sim$0.1) using the code `fuzzy_digitize` from the `halotools` package. We match the distribution to identical photometric redshift bins of the UltraVISTA survey [79]. This CAM mapping from the UniverseMachine sSFR to UltraVISTA-calibrated $\mathrm{sSFR_{UV}}$ is shown in Figure 12.

We train the mapping from $\mathrm{sSFR_{UV}}$ to mass-to-light ratio using the photometry and FAST stellar masses from UltraVISTA. We plot the feature-space of the UltraVISTA training data in Figure 15. These 140,472 training data leave very few missing regions of feature-space for $z < 3$, making it an ideal training set for our purposes. The random forest regression (the `RandomForestRegressor` class from the `scikit-learn` package [86]) is then used to predict $\log(M_*/L)$ from the two features $\{\log \mathrm{sSFR_{UV}}, z\}$. The advantages of this approach are its simplicity, flexibility, and sufficient accuracy in predicting this relation and its intrinsic scatter (see Figure 13 and 16). Additionally, this method automatically includes covariance between mass-to-light ratios of different photometric bands, which is important to accurately capture distributions of color as well as multivariate color-magnitude distri-

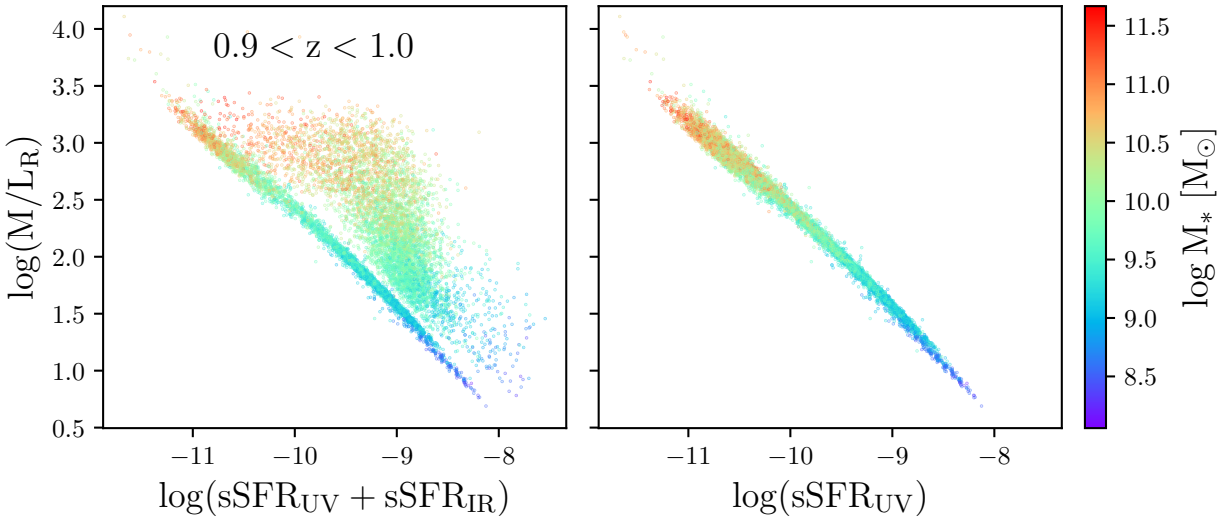

Figure 14: The relation between specific star formation rate and mass-to-light ratio for UltraVISTA galaxies in the redshift slice $0.9 < z < 1.0$. In the left panel, we use total sSFR (UV + IR), and in the right panel, we only use UV sSFR. The two plots share in common all galaxies without any IR detection, but the galaxies with IR detection form a cloud to the right due to their higher total sSFR (removing galaxies without IR detection removes the remaining narrow distribution entirely). We choose not to include $\mathrm{sSFR_{IR}}$ in our mass-to-light calibration due to this discontinuity between detections and non-detections.

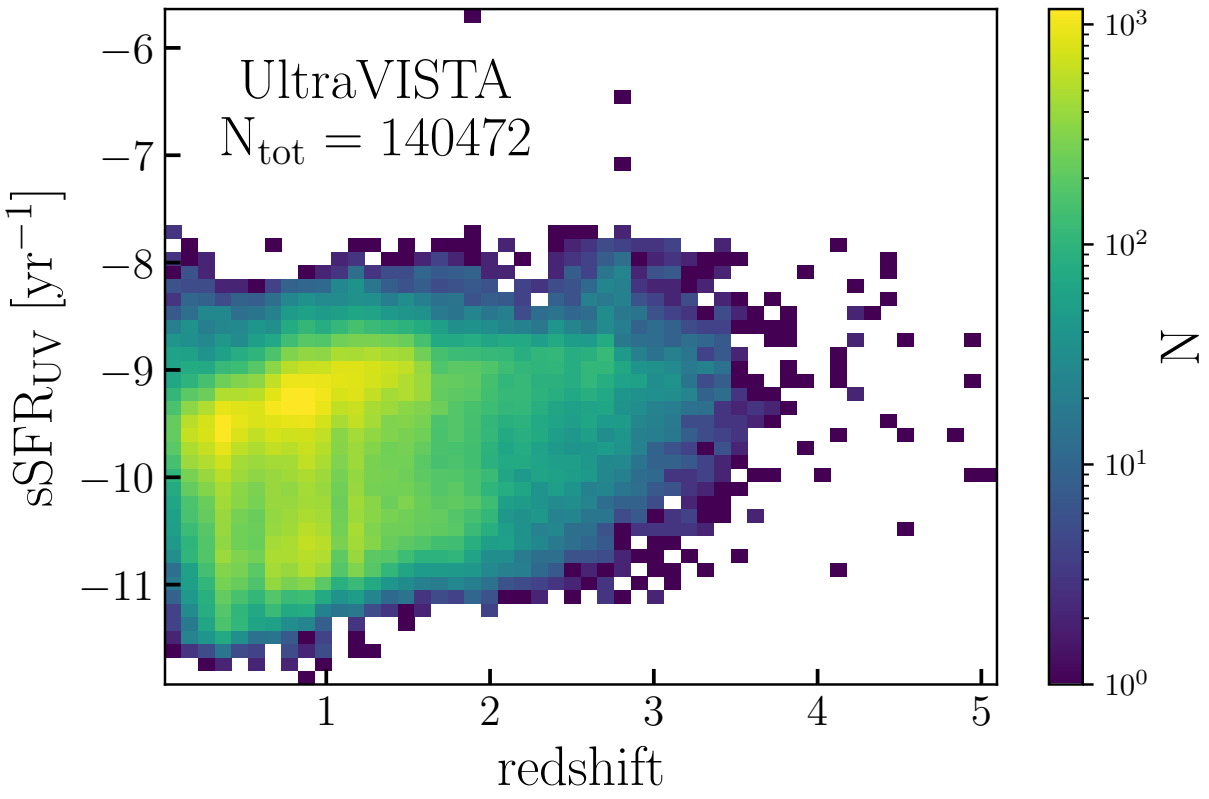

Figure 15: 2D histogram of the UltraVISTA data used to train mass-to-light ratios in our random forest. There are very few regions of parameter space without any data, which makes this an ideal dataset up to $z < 3$.

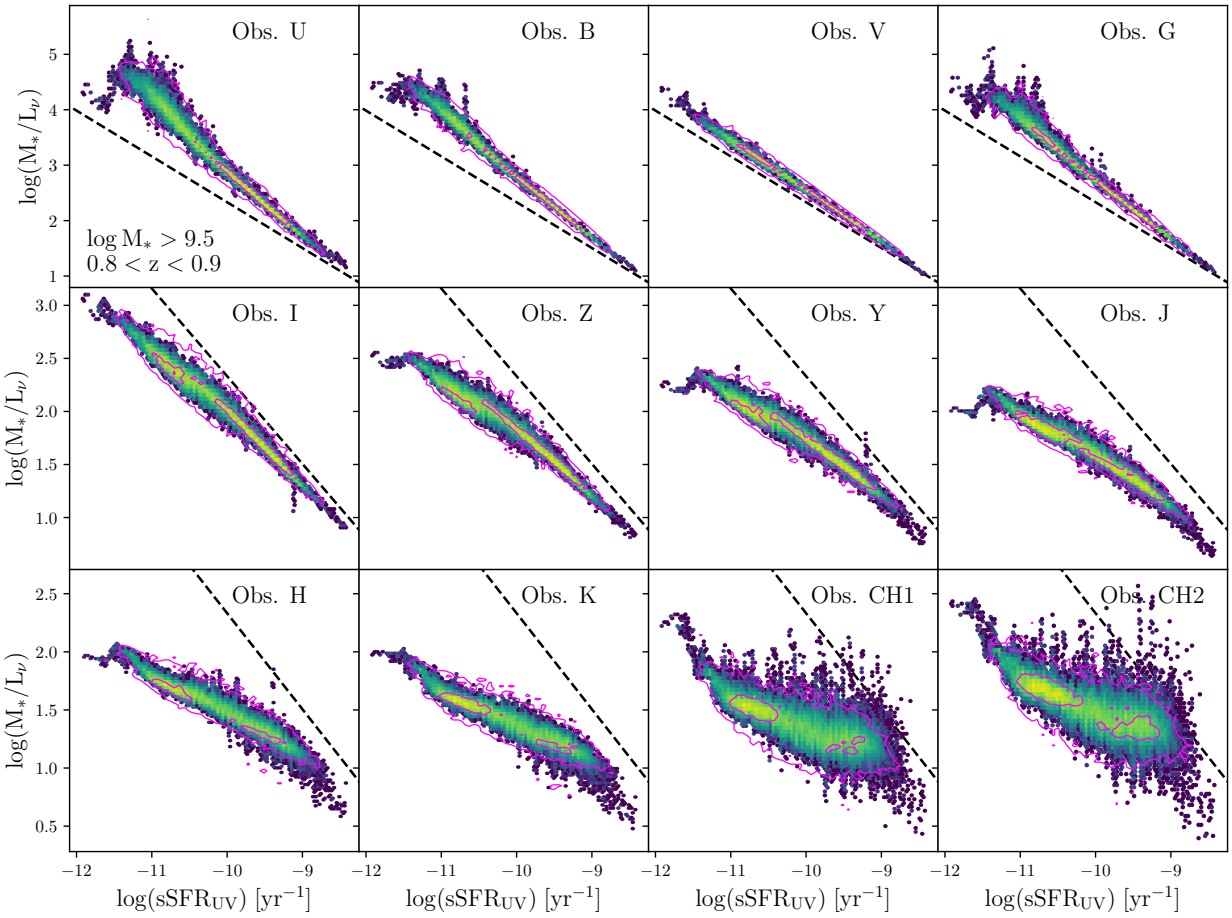

Figure 16: The relation between UV specific star formation rate and mass-to-light ratio in the redshift slice $0.8 < z < 0.9$. Same as Figure 13 but the y-axes represent the mass-to-light ratio in all other available photometric bands. The dashed line in each panel is the best-fit line for the R-band as a reference point.

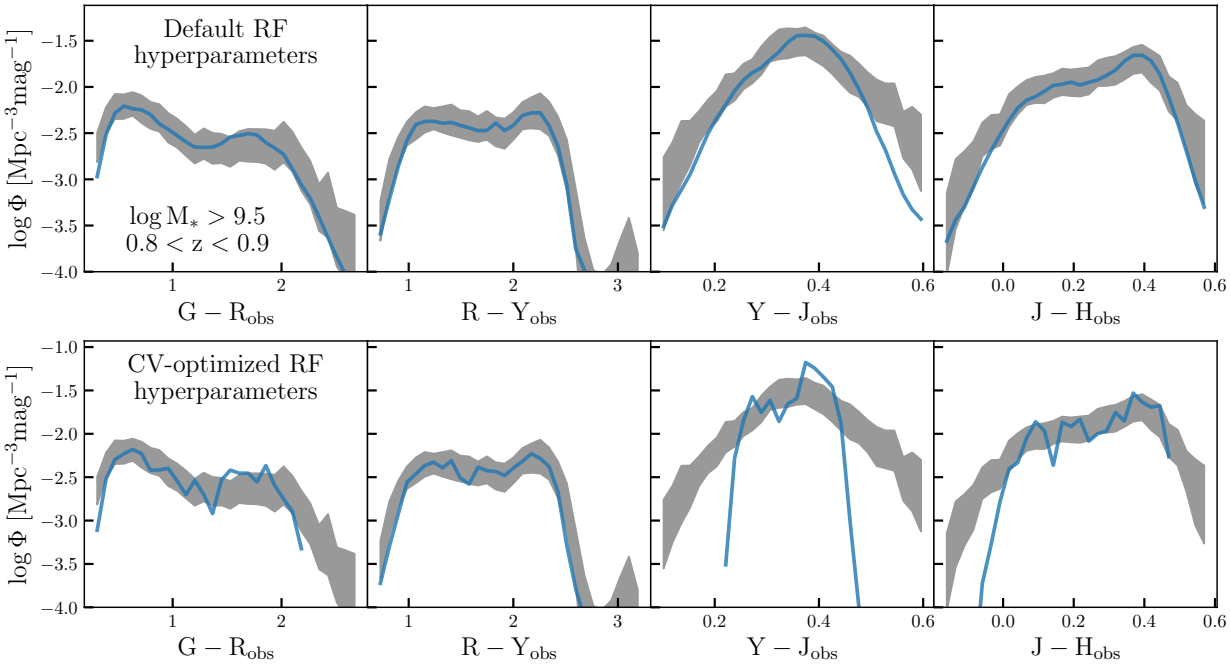

Figure 17: Comparison of the color distributions between UniverseMachine mock galaxies (blue lines) and the UltraVISTA data (grey bands) that they were fit to at the redshift slice $0.8 < z < 0.9$. The fits in the top panel use the default scikit-learn hyperparameters, while those in the bottom panel use cross-validation-optimized hyperparameters. By construction, the optimized hyperparameters predict colors with a lower mean absolute deviation. However, we adopt the default parameters due to their significantly better performance at reproducing the distributions as a whole.

butions. While random forests are not particularly good at extrapolation, our mock does not need to extrapolate in feature-space. Since we do not use stellar mass as a predictor of M/L (except indirectly through conditional abundance matching), it generalizes reasonably to masses slightly below the completeness limit of UltraVISTA, although predictions for galaxies far below this limit should be used with care.

To assess our mock catalog's performance at matching the UltraVISTA photometry, we compare the predicted distributions for a variety of properties, including the colors shown in Figure 17. The mass function and luminosity functions match very well by construction, but we also want to accurately reproduce color distributions, as the selection functions in real surveys will often use color cuts in addition to magnitude cuts. We have tested a variety of choices for the random forest hyperparameters used, but found that the default scikit-learn values performed best at reproducing a broad range of properties simultaneously. The hyperparameter values we use are therefore: `n_estimators` = 10, `bootstrap` = `True`, `max_depth` = `None`, `max_features` = `"auto"`, `min_samples_leaf` = 1, and `min_samples_split` = 2.

We have applied a number of common machine learning validation methods including 5-fold cross-validation testing and learning curve analysis, using the mean absolute deviation (MAD) of colors as our loss function, and find that these hyperparameter values yield a model that is modestly overfitted (cross-validated MAD value of 0.150, versus a training score of 0.058). We calculated an alternative set of optimized hyperparameters by minimizing the mean absolute deviation of the colors (via a random hyperparameter search followed by a smaller, but exhaustive, grid search), which yielded a more converged learning curve, indicating less overfitting (cross-validated MAD value of 0.130, versus a training score of 0.126). However, while the optimized hyperparameters perform better for predictions of individual colors, the M/L values at fixed mass concentrate closely around the mean value rather than capturing the full distribution. This results in worse color distributions (as seen in the bottom panel of Figure 17). Since we are in a regime where the feature-space distribution of the training set is nearly identical to that of the mock (by construction via conditional abundance matching), the overfitting when using the default parameters actually helps us, since it guarantees that the magnitude and color distributions in the mock match those of the training data.

2.8 CLIMBER Appendix - Additional Metrics

While the two-point correlation function provides quite good HOD constraints, it is also very sensitive to cosmology; particularly the σ_8 parameter, which controls the linear bias of halos. It has been shown [103] that the three-point correlation function would break this degeneracy, at least on linear scales. Calculating the three-point correlation function with existing public codes is too expensive to run at each iteration of an MCMC chain in our analysis. If a more efficient implementation becomes available, we would be interested in including this statistic to compare the added constraining power. This statistic will be especially important for joint analyses of cosmology and the galaxy-halo connection.

Additionally, using the two-point correlation function alone can miss out on important clustering information, as it is primarily driven by the most clustered galaxies, which is almost always satellites except at the highest masses. One can explicitly measure a signal from central galaxies singling them out via isolation criteria or group catalog reconstruction. One metric in particular that is capable of measuring assembly bias signals is the number of satellites vs. central stellar mass, split into quenched and star-forming populations (shown in private communication with Rodríguez-Puebla). It is unclear how accurately group catalogs can be reconstructed from surveys like PFS and MOONS, which are incomplete due to fiber collisions, and require the use of photometric redshifts to fill in the gaps. This may prove to be a strong argument for extensions to these surveys prioritizing increasing the completeness over the area, contrary to the cosmic variance tests herein.

Alternatively, it is possible to supplement the two-point correlation function with other statistics such as counts-in-cylinders (CIC). By tuning the radius of the cylinder and inner annulus, CIC can effectively separate the signal of centrals from satellites, without requiring full group catalog reconstruction. [113] demonstrate that CIC + $w_{\rm p}(r_{\rm p})$ may be sufficient to constrain assembly bias as well. However, further testing is required for samples affected by fiber collisions like PFS and MOONS.

3.0 Galtab: Assembly Bias Evidence from Low-Redshift Counts-in-Cylinders Measurements in the DESI One-Percent Survey

This chapter (Pearl et al. 2023 in prep) will be submitted to the AstroPhysical Journal following DESI collaboration-wide review, which is ongoing at the time of thesis submission.

3.1 Galtab Introduction

The large-scale distribution of galaxies in the universe is a powerful probe of cosmological models (e.g., [13, 5, 2]). This is because galaxies trace the dark matter distribution, whose distribution is set by cosmological parameters and is well-characterized by modern simulations (e.g., [60, 53]). However, for accurate cosmological inference, it is necessary to marginalize over the possible relationships between observational probes and the theoretical matter distribution. Therefore, leveraging large-scale structure to constrain cosmology requires flexible models of the galaxy-halo connection, and necessitates incorporating as much empirical information as possible to tightly constrain such flexible models.

Central and satellite galaxies are thought to form at the dense centers of halo and subhalo potential wells, respectively. Therefore, the spatial clustering of most galaxy samples can be described well by a halo occupation distribution (HOD; e.g., [10, 127]), which probabilistically connects the average number of central and satellite galaxies a dark matter halo hosts to its mass. This formalism can be extended through additional parameters that lead to correlations between galaxy abundance and secondary halo properties (i.e., assembly bias [50]), which can improve fit quality. As the data improve, further extensions to HOD models may be warranted, e.g., by relaxing the assumption of a log-normal stellar-to-halo-mass relation or of a spatially isotropic Navarro-Frenk-White (NFW [81]) distribution of satellite galaxies.

The most common observables used to constrain the galaxy-halo connection via spectroscopic galaxy samples are the number density and the projected two-point correlation

function $w_{\rm p}(r_{\rm p})$ (e.g., [122, 92]). However, [113] has shown that the counts-in-cylinders (CiC) distribution $P(N_{\rm CiC})$ offers significant complementary information on the parameters of interest – particularly those that control satellite occupation and assembly bias. As demonstrated by [106], it is also possible to quantify clustering information beyond the two-point function using the underdensity probability function and the density-marked correlation function. These studies highlight that even with existing datasets, incorporating different measurements of the large-scale structure can help optimize model fitting.

In this paper, we extend previous analyses by incorporating a novel spectroscopic dataset; implementing a new, more efficient CiC prediction framework; and demonstrating the gain these provide. We leverage data from the Dark Energy Spectroscopic Instrument (DESI [34]), which will ultimately obtain spectroscopic redshifts of 40 million galaxies in an effort to precisely map the large-scale structure of a large volume of the observable universe. While the full dataset is still being collected, this work utilizes redshift measurements for more than 40,000 galaxies obtained by the Survey Validation 3 (SV3) component of the DESI early data release [35].

We approximately adopt the best-fit flat-universe cosmology from [89]. The relevant cosmological parameters that we use are as follows: $h = 0.6777$, $\Omega_{m,0} = 0.30712$, $\Omega_{b,0} = 0.048252$, and $T_{\rm CMB} = 2.7255$ K. However, we scale all distance and distance-dependent values to units equivalent to setting the Hubble parameter to $h = 1$ (e.g., h^{-1}Mpc).

This paper is organized as follows. We describe the DESI and simulation data in Section 3.2. We outline the summary statistics used in our analysis in Section 3.3. We detail our methodology for measuring and predicting CiC, through the `galtab` package, in Section 3.4. We present our resulting constraints on the HOD in Section 3.5, and discuss our conclusions in Section 3.6.

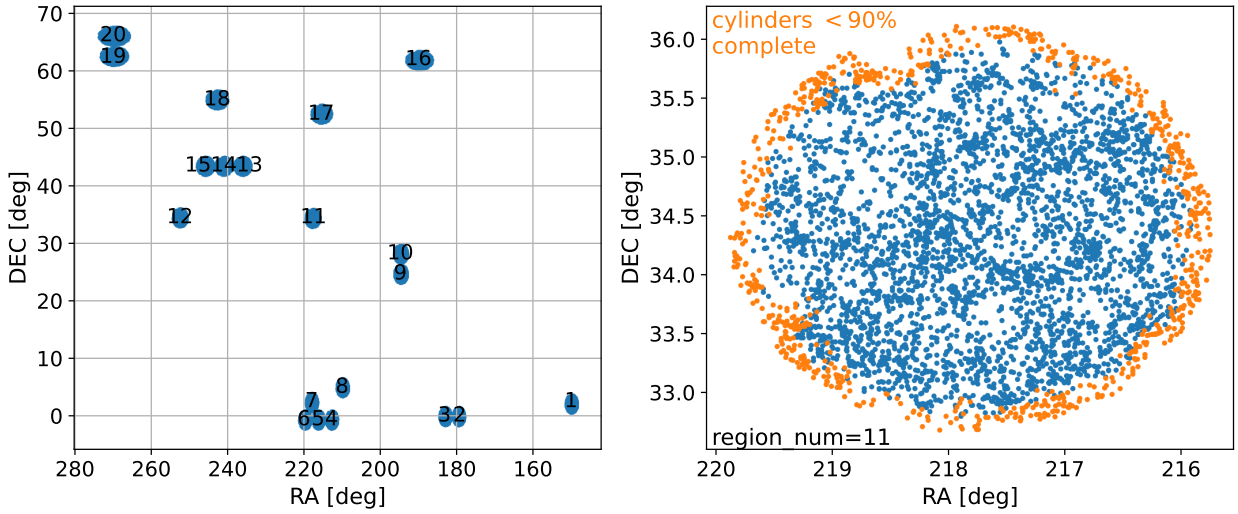

Figure 18: Footprint of the DESI Survey Validation 3 (SV3). The left panel displays the entire survey, broken up into twenty regions that are for the most part spatially isolated from each other. The right panel presents a close-up of the region labeled by the number 11 in the left panel. The points shown in orange, which are located primarily near the edge of the region, indicate objects excluded as cylinder centers in our CiC measurement, as described in Section 3.4.2

.

3.2 Data

3.2.1 DESI BGS

The DESI Bright Galaxy Survey (BGS) is a highly complete magnitude-limited spectroscopic survey of $z < 0.5$ galaxies, which aims to target galaxies over at least 14,000 square degrees down to a limit roughly two magnitudes fainter than the Sloan Digital Sky Survey (SDSS [1]). Our analyses only use the BGS Bright sample, which is complete down to an apparent r-band magnitude of $m_r < 19.5$. Because the DESI survey is still in progress at the time of this writing, we analyze only data from the Survey Validation 3 (SV3 [35]) dataset (also known as the One-Percent Survey as it contains approximately 1% of the anticipated

volume of DESI). These data were obtained in over twenty sky regions totaling an area of 173.3 sq deg, as shown in Figure 18. A significantly higher fraction of potential targets was observed in the SV3 fields than will be the case for typical DESI survey data due to the use of a denser tiling strategy, simplifying the corrections needed for our analysis.

We specifically use the SV3 Large Scale Structure (LSS) catalogs, which only include sources with secure spectroscopic redshift measurements, as described in [35]. These catalogs are well suited for clustering measurements since they are paired with 18 random realization files, each containing 2500 objects per deg^2 of sky coverage, and weights from 128 fiber assignment realizations. We also utilize r-band absolute magnitude measurements from `fastspecfit` (Moustakas et al. in prep.[1]), which are computed for an SDSS r-band response curve K-corrected to the $z = 0.1$ reference frame. Note that all references to absolute magnitudes in this paper, M_r, are scaled to $h = 1$ units; therefore, they are equivalent to $M_r - 5 \log h$ for all other values of the Hubble parameter.

We break this data into three volume-limited samples which each cover the redshift range $0.1 < z < 0.2$, constructed with absolute r-band absolute magnitude limits of $M_r < -20.0$, -20.5, and -21.0. We also define a fourth sample covering a slightly higher redshift range of $0.2 < z < 0.3$ with limit $M_r < -21.0$. We plot each sample cut in Figure 19 and summarize these samples in Table 5. Unless otherwise specified, all observational measurements in this paper are measured from one of these samples.

3.2.2 Small MultiDark Planck

To study the galaxy-halo connection, we must compare DESI galaxy clustering data to an assumed distribution of underlying dark matter halos. For this halo distribution prior, we adopt the Small MultiDark Planck simulation (SMDPL [60]), which uses the same Planck cosmology that we assume in this work. This simulation was performed with 3840^3 particles, but our analysis is based only upon the halo catalogs produced by applying the Rockstar halo finder [8]. We adopt the virial mass from Rockstar as our halo mass, M_h.

SMDPL covers a $400h^{-1}$ Mpc periodic cube, which is over ten times the volume of our

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

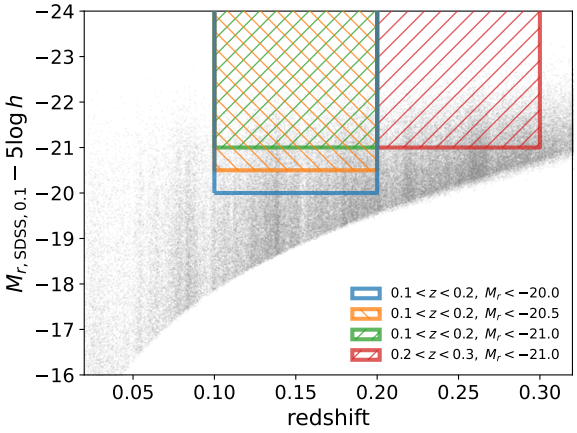

Figure 19: Distribution of r-band absolute magnitude M_r vs. redshift. The full DESI BGS SV3 sample is shown by the grey points. Our four volume-limited, absolute magnitude-thresholded samples are constructed through the cuts represented by the corresponding colored boundaries.

SV3 samples. This is sufficiently large so that cosmic-variance-like uncertainties from the data dominate over the sample variance of this simulation volume. However, future studies will need to use larger volume simulations to compensate for DESI's volume, which will be 100 times that of SV3.

Table 5: DESI subsamples used for our analyses. The full sample size is given by N_{tot}, while N_{cyl} is the number of centers of the cylinders that meet our spatial completeness criteria.

M_r threshold	Redshift range	N_{tot}	N_{cyl}
-20.0	$0.1 < z < 0.2$	20,241	15,936
-20.5	$0.1 < z < 0.2$	11,036	8,686
-21.0	$0.1 < z < 0.2$	5,096	4,031
-21.0	$0.2 < z < 0.3$	14,874	12,543

3.3 Observable Summary Statistics

To extract clustering information from the galaxy samples, we use three summary statistics: number density $n_{\rm gal}$, the projected two-point correlation function $w_{\rm p}(r_{\rm p})$, and the CiC distribution $P(N_{\rm CiC})$. We compare the observations[2] with our best-fit models for these three summary statistics for each sample in Figure 26.

The number density is calculated via the sum of the inverse individual probability (IIP; see Section 3.4.2) weights of the galaxies in the sample divided by the comoving volume they were sampled from. For the HOD number density predictions, the comoving volume of SMDPL is $400^3 h^{-3}{\rm Mpc}^3$, while the volumes of the DESI samples depend on the redshift cuts and the survey area. The DESI SV3 BGS survey area is 173.3 sq deg, which corresponds to comoving volumes of $2.83 \times 10^6 h^{-3}{\rm Mpc}^3$ and $6.95 \times 10^6 h^{-3}{\rm Mpc}^3$ for samples with redshift ranges of $0.1 < z < 0.2$ and $0.2 < z < 0.3$, respectively.

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

$$w_{\rm p}(r_{\rm p}) = 2 \int_0^{\pi_{\rm max}} \xi(r_{\rm p}, \pi) d\pi \tag{11}$$

where ξ is the two-point correlation function, π is line-of-sight separation distance, and $r_{\rm p}$ is perpendicular separation distance. For consistency with [112], we choose $\pi_{\rm max} = 40 h^{-1}{\rm Mpc}$ and use twelve logarithmically spaced bins between $r_{\rm p}$ of $0.158 h^{-1}{\rm Mpc}$ and $39.81 h^{-1}{\rm Mpc}$. We concatenate all 18 random files from the SV3 LSS catalogs but draw a random 20% subsample to reduce excessive computational time. We utilize the `pycorr`[3] package to apply the [65] estimator, line-of-sight integration, and fiber assignment weights. The performance-critical pair searching is powered by `Corrfunc` [102].

Counts-in-cylinders (CiC) is a type of counts-in-cells statistic (i.e., it quantifies the local density of neighbors in a cell around each object; the development of such metrics has a

[2] i.e., Bezanson Points

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

long history; e.g., [52, 130, 119, 3]) that defines neighbors using a cylindrical cell along the line-of-sight direction. As in [112], we use relatively small-scale cylinders by choosing the radius to be $R_{\rm CiC} = 2h^{-1}$Mpc and the half-length to be $L_{\rm CiC} = 10h^{-1}$Mpc. Cylinders of this scale primarily probe the number of intra-halo galaxies and are therefore sensitive to satellite occupation. Conveniently, using a small cylindrical volume is also a computationally favorable choice. The CiC distribution $P(N_{\rm CiC})$ can be evaluated in bins of $N_{\rm CiC}$ – for which we use ten linearly spaced bins between -0.5 and 9.5 plus twenty logarithmically spaced bins between 9.5 and 149.5; alternatively, the majority of available information can be captured by computing the first three to five moments of the $N_{\rm CiC}$ distribution. We describe our methods used to compute counts-in-cylinders in detail in Section 3.4.

We test the ability of each summary statistic to provide information about the HOD by sampling uniformly from HOD parameters around their 1σ confidence interval from the [112] $M_r < -20.5$ sample. We predict each of our summary statistics plus noise according to a random draw from the covariance matrix calculated in Section 3.3.1, including CiC up to the tenth moment. We then train a random forest [20] to predict the HOD parameters from these summary statistics and provide a visualization of the resulting SHapley Additive exPlanations (SHAP [72]) feature importance in Figure 20. To briefly summarize, number density is highly important for predicting $\log M_{\rm min}$, the two-point correlation function is broadly informative across all parameters, and the first few CiC moments are particularly important for constraining satellite HOD parameters.

3.3.1 Covariance of Summary Statistics

To constrain our HOD model, we compare the following summary statistics as measured in our data to model predictions: number density; the two-point correlation function (computed in 12 bins in $r_{\rm p}$); and CiC (for 28 bins in $N_{\rm CiC}$). We calculate the covariance matrix of these summary statistics by jackknife resampling using the 20 regions displayed in Figure 18.

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

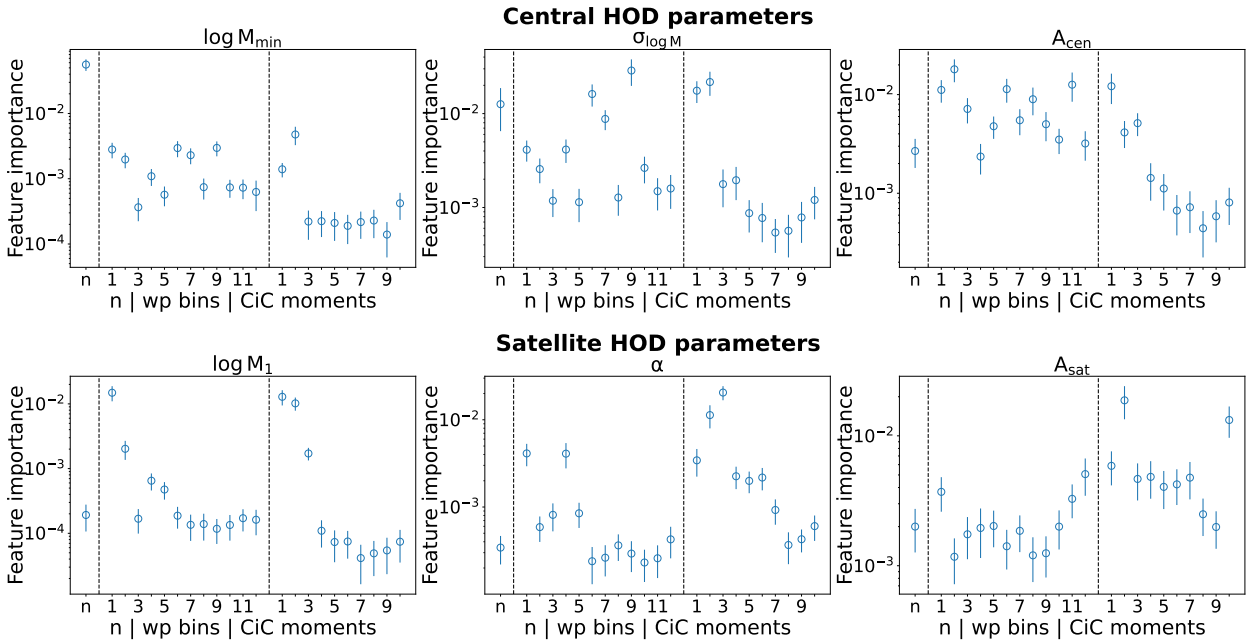

Figure 20: SHAP feature importances for each of our summary statistics for inferring HOD parameters. Each panel plots the importance of each feature (i.e., each quantity that is used to predict the HOD parameters via a machine learning model), calculated by the mean absolute SHAP value for the given HOD parameter. Summary statistics with high feature importance are more useful for predicting the parameter. For the satellite HOD parameters (bottom row), the first few CiC moments provide the majority of the constraining information. See Figure 28 for beeswarm plots of the full distribution of SHAP values of the six most important features for each parameter.

matrix of our summary statistics can then be estimated by

$$\Sigma_{ij} = \frac{N_\mathrm{J} - 1}{N_\mathrm{J}} \sum_{k=0}^{N_\mathrm{J}} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j) \tag{12}$$

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

3.4 Counts-in-Cylinders

Counts-in-cylinders (CiC; derived in [87] and previously used by [93, 112]) is sensitive to higher-order n-point functions, which makes it complementary to two-point statistics commonly used in the literature. Despite its utility, CiC is not widely adopted in galaxy-halo connection studies, due to difficulties in correcting for systematics, excessive computational time, and significantly increased dimensionality of the full covariance matrix. In this section, we present our methodology to mitigate all of these problems and implement each of these methods in an open source Python package `galtab`[4].

After a brief explanation of our observational cylinder geometry in Section 3.4.1, we present our weighting method in Section 3.4.2 based on individual inverse probabilities and inverse conditional probabilities (IIP×ICP), which corrects CiC calculations to account for clustering bias in surveys with fiber collisions. This approach is analogous to and inspired by pair inverse probabilities (PIP) weighting [14], which we used to correct our $w_{\rm p}(r_{\rm p})$ measurement. To minimize the increase in dimensionality, we suggest using only the first three to five moments of the CiC distribution, defined in Section 3.4.3, which retain most of the constraining information. Our analysis uses information from the entire CiC distribution, but our results are not significantly affected by using only the first five CiC moments instead. Additionally, we present a galaxy placeholder pretabulation method in Section 3.4.4 to speed up our Markov-chain Monte Carlo (MCMC) procedure. This makes our CiC prediction runtime comparable to traditional Monte Carlo $w_{\rm p}(r_{\rm p})$ prediction methods but with the significant advantage of producing precise, deterministic values, which yield much higher MCMC sampling efficiency than stochastic Monte Carlo predictions.

3.4.1 Observational Cylinder Geometry

While a cylinder perfectly aligns with the velocity distortion in an idealized simulation, for observations, we must slightly distort its round face into a truncated cone so that it is always perpendicular to the line-of-sight direction. We also allow a slight curve to this

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

truncated cone's top and bottom faces, to keep them normal to the line of sight. Then there are only two search criteria: angular distance and line-of-sight separation. The line-of-sight separation cut is L_{CiC} and we define the angular radius cut to be

$$\theta_{\mathrm{CiC}} = \arccos\left(1 - \frac{3R_{\mathrm{CiC}}^2 L_{\mathrm{CiC}}}{(d + L_{\mathrm{CiC}})^3 - (d - L_{\mathrm{CiC}})^3}\right) \tag{13}$$

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

3.4.2 IIP×ICP Weighting

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

$$P(i) = \frac{\texttt{sum(bitweights[i])} + 1}{129}, \tag{14}$$

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

$$P(i \;\&\; j) = \frac{\texttt{sum(bitweights[i] \& bitweights[j])} + 1}{129} \tag{15}$$

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

In order to measure the CiC distribution, we must calculate the expectation value of the number of galaxies we expect to find in the cylinder around every galaxy individually, $N_{\mathrm{CiC},i}$. For this task, we sum the inverse conditional probabilities (ICPs) of each neighboring galaxy's fiber assignment (conditional on the fiber assignment of the cylinder's central galaxy). Using the definitions from Equations 14 and 15,

$$\mathrm{ICP}_{j|i} = \frac{P(i)}{P(i \ \& \ j)} \tag{16}$$

$$N_{\mathrm{CiC},i} = \frac{1}{f_{\mathrm{rand}}} \sum_{j \in C_i} \mathrm{ICP}_{j|i} \tag{17}$$

where C_i is the set of indices of galaxies contained by the cylinder surrounding the ith galaxy, and f_{rand} is the fraction of randoms enclosed in the cylinder compared to the expected number occupying a circle of angular radius θ_{CiC} in order to account for incompleteness in spatial coverage. Note that we do not include cylinders with $f_{\mathrm{rand}} < 0.9$. This cut excludes approximately 21% of the cylinders at $z \sim 0.15$ and 16% of the cylinders at $z \sim 0.25$, as listed in Table 5.

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

$$\mathrm{IIP}_i = \frac{1}{P(i)}. \tag{18}$$

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

$$\frac{\mathrm{IIP}_{i+}\lceil N_{\mathrm{CiC},i} \rceil + \mathrm{IIP}_{i-}\lfloor N_{\mathrm{CiC},i} \rfloor}{\mathrm{IIP}_i} = N_{\mathrm{CiC},i} \tag{19}$$

3.4.3 Calculating the CiC Moments

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

$$\mu_1 = \sum_{i=1}^{N} w_i N_{\mathrm{CiC},i} \tag{20}$$

$$\mu_2 = \sqrt{\sum_{i=1}^{N} w_i (N_{\mathrm{CiC},i} - \mu_1)^2} \tag{21}$$

$$\mu_{k>2} = \frac{1}{\mu_2^k} \sum_{i=1}^{N} w_i (N_{\mathrm{CiC},i} - \mu_1)^k. \tag{22}$$

where $N_{\mathrm{CiC},i}$ is the number of neighbors inside the cylinder surrounding the i^{th} galaxy in the sample and w_i is the i^{th} IIP weight, but normalized to $\sum w_i = 1$ (see Section 3.4.2 for details on $N_{\mathrm{CiC},i}$ and IIP weights). Note that μ_1 is the mean, μ_2 is the standard deviation, and for $k > 2$, μ_k are standardized central moments (skewness, kurtosis, etc.), uncorrected for degree-of-freedom bias, which is a negligible source of systematics for large sample sizes compared to other uncertainties. In figures, we refer to μ_k as CiC_k to be explicit that they are moments of CiC.

3.4.4 Pretabulation with Placeholder Galaxies

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

Our procedure is illustrated in Figure 21. Our method requires a fiducial HOD model to compute the expected occupation, $\langle N_{\mathrm{cen}} \rangle$ and $\langle N_{\mathrm{sat}} \rangle$, for each halo. For our fiducial model, we choose the best fit of [112] that corresponds to the magnitude threshold of each of our samples. We populate each halo with $N_{\mathrm{cen,ph}}$ central placeholders and $N_{\mathrm{sat,ph}}$ satellite placeholders. We determine the number of satellite placeholders for each halo with the hyperparameter W_{max} according to the equation

$$N_{\mathrm{sat,ph}} = \left\lceil \frac{\langle N_{\mathrm{sat}} \rangle}{W_{\mathrm{max}}} \right\rceil \tag{23}$$

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

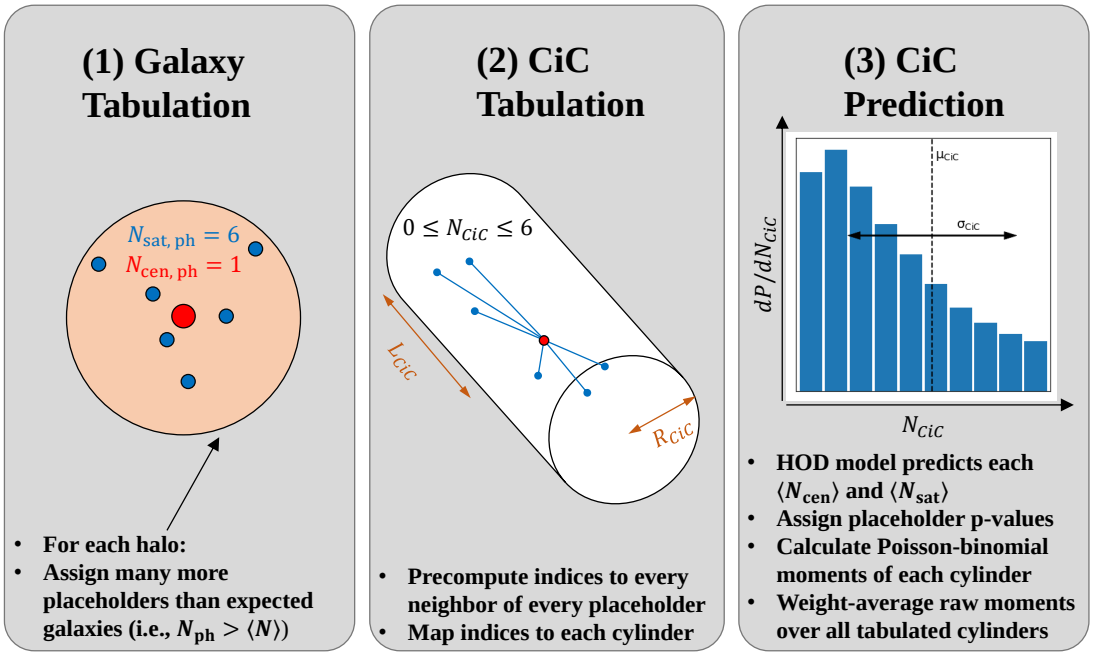

Figure 21: Demonstration of our placeholder algorithm used to pretabulate counts-in-cylinders pair indices. Given a fiducial model, we populate placeholder centrals for most halos with a non-zero probability of hosting a halo. We populate many more placeholder satellites than expected in the fiducial model so that the resulting binomial satellite occupation distribution sufficiently resembles the assumed Poisson distribution. We then tabulate the placeholder indices in each halo for rapid CiC prediction using one of the two modes described in Sections 3.4.5 and 3.4.6.

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

$$Q_{\rm min} = \frac{\int_{\langle N_{\rm cen} \rangle_{\rm min}}^{1} \Phi(\langle N_{\rm cen} \rangle) \langle N_{\rm cen} \rangle d\langle N_{\rm cen} \rangle}{\int_{0}^{1} \Phi(\langle N_{\rm cen} \rangle) \langle N_{\rm cen} \rangle d\langle N_{\rm cen} \rangle} \tag{24}$$

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

between $\langle N_{\rm cen}\rangle$ and $\langle N_{\rm cen}\rangle + d\langle N_{\rm cen}\rangle$.

To balance accuracy and runtime (see Figure 22), we set $W_{\rm max} = 0.05$ and $Q_{\rm min} = 10^{-4}$. In `galtab`, these hyperparameters can be tuned via the `max_weight` and `min_quant` keyword arguments, respectively.

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

As is usually done in Monte Carlo HOD realizations, these galaxy probability values are assumed to be independent. Therefore, the halo occupation of centrals follows a Bernoulli distribution, the same as typical Monte Carlo frameworks. However, the halo occupation of satellites follows a binomial distribution in our framework, which only converges to the desired Poisson distribution in the low $P_i \lesssim 0.05$ limit, hence our choice of $W_{\rm max} = 0.05$.

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

3.4.5 Pretabulated CiC Prediction: Monte Carlo Mode

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

$$P(N_{\rm CiC}) = \frac{\sum_{i=1}^{N} P_i P(N_{{\rm CiC},i})}{\sum_{i=1}^{N} P_i}. \tag{25}$$

In general, each $P(N_{{\rm CiC},i})$ is a Poisson binomial distribution, whose calculation scales exponentially with the number of neighbors in the ith cylinder, which is infeasible. Therefore, the full distribution can only be calculated using our Monte Carlo mode prediction. In this mode, we also pretabulate $n_{\rm MC}$ random seeds over $[0, 1)$ for each galaxy, which we use as Bernoulli quantiles after assigning the P_i of each placeholder. This allows us to effectively

create $n_{\rm MC}$ independent realizations that can still produce quasi-deterministic and almost continuous (but non-differentiable) predictions. We find that using $n_{\rm MC} = 10$ random seeds produces reasonably stable results without excessive runtime. We will show in Section 3.4.6 that predictions of the CiC moments can be made without invoking random seeds, allowing them to be perfectly continuous and differentiable.

3.4.6 Pretabulated CiC Prediction: Analytic Mode

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

$$\langle N_{\rm CiC} \rangle = \frac{\sum_{i=1}^{N} P_i \langle N_{{\rm CiC},i} \rangle}{\sum_{i=1}^{N} P_i} \tag{26}$$

where

$$\langle N_{{\rm CiC},i} \rangle = \sum_{j \in C_i} P_j \tag{27}$$

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

It is possible to calculate a similar relation for the standard deviation and the higher standardized moments we have defined in Equations 21 and 22. However, these relations are much more complicated. Note that the mean is a special case because it is the first raw moment (which allows Equation 26) as well as the first cumulant (which allows Equation 27).

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

$$N_{{\rm CiC},i} = \sum_{j \in C_i} X_j \tag{28}$$

where X_j is the occupation of the jth placeholder, which follows a Bernoulli distribution (0 or 1) with mean P_j. Therefore, the first cumulant of this Bernoulli distribution is $\kappa_1(X_j) = P_j$,

and the subsequent Bernoulli cumulants can be derived from the recursion relation

$$\kappa_{k+1}(X_j) = P_j(1 - P_j)\frac{d\kappa_k(X_j)}{dP_j}. \tag{29}$$

Given the first $k_{\max}$ Bernoulli cumulants of each placeholder, we can calculate the first $k_{\max}$ Poisson binomial cumulants of the ith cylinder. We can take the kth cumulant of each random variable on both sides of Equation 28:

$$\kappa_k(N_{\mathrm{CiC},i}) = \sum_{j \in C_i} \kappa_k(X_j). \tag{30}$$

From the moments of each $N_{\mathrm{CiC},i}$, we would like the moments of the combined CiC distribution, which is a weighted superposition of each individual cylinder's distribution, as expressed in Equation 25. For this step, the most convenient set of moments to use are raw moments. The kth raw moment of $N_{\mathrm{CiC},i}$ can be obtained from its first k cumulants according to

$$\langle N^k_{\mathrm{CiC},i} \rangle = \kappa_k(N_{\mathrm{CiC},i}) + \sum_{j=1}^{k-1} \kappa_j(N_{\mathrm{CiC},i}) \langle N^{k-j}_{\mathrm{CiC},i} \rangle. \tag{31}$$

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

$$\langle N^k_{\mathrm{CiC}} \rangle = \frac{\sum_{i=1}^{N} P_i \langle N^k_{\mathrm{CiC},i} \rangle}{\sum_{i=1}^{N} P_i}. \tag{32}$$

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

$$\langle (N_{\mathrm{CiC}} - \langle N_{\mathrm{CiC}} \rangle)^k \rangle = \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j} \langle N^j_{\mathrm{CiC}} \rangle \langle N_{\mathrm{CiC}} \rangle^{k-j} \tag{33}$$

from which we can calculate the standard moments given in Equations 20 through 22 using

$$\mu_1 = \langle N_{\mathrm{CiC}} \rangle, \tag{34}$$

$$\mu_2 = \sqrt{\langle (N_{\mathrm{CiC}} - \langle N_{\mathrm{CiC}} \rangle)^2 \rangle}, \tag{35}$$

and

$$\mu_{k>2} = \frac{1}{\mu_2^k} \langle (N_{\mathrm{CiC}} - \langle N_{\mathrm{CiC}} \rangle)^k \rangle. \tag{36}$$

3.4.7 Computational Performance

In Section 3.4.4 and Figure 22, we have described our hyperparameter tuning of $W_{\max}$ and $Q_{\min}$ to balance runtime and accuracy. These parameters control the number of placeholders, N, as well as the average number of placeholders per cylinder, C. To store all pretabulated indices, the memory usage of `galtab` scales with $\mathcal{O}(NC)$.

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

By far, the most computationally expensive step of our procedure is the summation of occupations (or cumulants, for the analytic mode; see Equation 30) of placeholders per cylinder. To fully optimize this calculation, we employ just-in-time (JIT) compilation via the JAX library [19]. This also automatically ports the computation to the GPU, if available, which can speed up the predictions by at least an order of magnitude faster than the times reported in Figure 22.

3.5 Constraining the HOD

3.5.1 HOD Model

We employ a decorated HOD model based on the formulation of [127]. In this framework, the halo occupations of central and satellite galaxies over a given magnitude threshold are

described by Bernoulli and Poisson distributions, respectively. Their means are functions of halo mass M_h, described by

$$\langle N_{\rm cen}\rangle(M_h) = \frac{1}{2}\left(1 + {\rm erf}\left(\frac{\log(M_h/M_{\rm min})}{\sigma_{\log M}}\right)\right) \tag{37}$$

and

$$\langle N_{\rm sat}\rangle(M_h) = \left(\frac{M_h - M_0}{M_1}\right)^{\alpha} \tag{38}$$

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

$$\log M_0 = \log M_{\rm min} + Q_0(\log M_1 - \log M_{\rm min}) \tag{39}$$

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

Adding further flexibility into our model, we include assembly bias parameters $A_{\rm cen}$ and $A_{\rm sat}$ to introduce a halo occupation dependence on the NFW concentration. These parameters both range from [-1, 1], and allow for redistribution of the central and satellite occupation, respectively, from low to high concentration halos for positive A, or vice versa. See [50] for further details on the decorated HOD parameterization.

3.5.2 MCMC Fits

We use Markov-chain Monte Carlo (MCMC) to constrain the HOD model using each galaxy sample. We make use of the `emcee` [42] implementation, in which several walkers simultaneously sample a likelihood function throughout parameter space, and occasionally trade locations to construct MCMC chains. Ignoring the normalization constant, the log-likelihood is given by

$$\ln\mathcal{L} = -\frac{1}{2}(\vec{x}_{\rm model} - \vec{x}_{\rm data})^{\intercal}\Sigma^{\dagger}(\vec{x}_{\rm model} - \vec{x}_{\rm data}) \tag{40}$$

where Σ is the covariance matrix from Equation 12 and $\Sigma^{\dagger}$ is its Moore-Penrose pseudo-inverse [88], which prevents the reduced dimensionality of our likelihood from affecting the likelihood numerically. Loss in dimensionality occurs when we use the full CiC distribution (but not when we reduce this information into CiC moments) due to some eigenvalues in the covariance matrix equaling zero when there are at least 20 summary statistics, which is our number of jackknife realizations. We use the implementation available in the `logpdf` method of the `multivariate_normal` class from SciPy [111].

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

We initialize our MCMC chains around the best-fit parameters of the corresponding magnitude threshold sample from [112], with very slight variation between the MCMC walkers. We let these chains run for 60,000 trial points (3,000 iterations $\times$ 20 walkers), and conservatively remove a burn-in of 2,000 trial points to calculate our posteriors displayed in Figures 23, 24, and 25, as well as the maximum-likelihood points and confidence regions reported in Tables 6 and 7, respectively. Our relatively small number of trial points is acceptable thanks to our deterministic likelihood evaluations and our prior on $\log M_0$ that confines the MCMC to a stable region of parameter space. The autocorrelation lengths of our chains ended up ranging from 100-300. This is about a factor of two shorter than the autocorrelation lengths we obtain using Monte Carlo CiC evaluations, and possibly orders of magnitudes shorter than the result from Monte Carlo $w_{\rm p}(r_{\rm p})$ evaluations.

To quantify how well our maximum-likelihood models agree with the data, we calculate χ^2 along with the probability of measuring data with at least this value of χ^2 by chance using the chi-squared cumulative distribution function[5]. In Table 6, we report this probability and translate it into the z-score of a Gaussian to quantify the "number of sigmas" of tension that exists between our model and data.

[5] i.e., the Newman Score

3.6 Results and Discussion

The measurements from the DESI One-Percent Survey already produce reasonably tight constraints on the HOD. For each of the four threshold samples defined in Table 5, the corresponding best-fit HOD parameters are given in Table 6, and 1σ confidence intervals are given in Table 7. We have also summarized these constraints as a function of M_r threshold and redshift into easier-to-digest plots in Figure 27. In this figure, we show that as luminosity increases from M_r of -20.0 to -21.0, the characteristic halo mass for central galaxies gradually increases from roughly $10^{12.0}$ to $10^{12.4}$ $M_\odot$. We find a similar increasing trend for the characteristic halo masses containing one (and two) satellite galaxies for each sample; the inferred slope α of the $\langle N_{\rm sat}\rangle(M_{\rm halo})$ relation does not evolve significantly compared to the shown error bars. Finally, we show the parameters which trace assembly bias; these are very significantly greater than zero for centrals in the lower two magnitude threshold samples, while satellite assembly bias is consistent with zero throughout. With only one $z = 0.25$ sample, we find no significant signals of redshift evolution.

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

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

The constraints we find on assembly bias are consistent with the findings from studies based on SDSS data. Despite the smaller sample size currently available from DESI, our

$w_{\rm p}(r_{\rm p})$ + CiC analysis produces much stronger constraints than characterizing SDSS clustering with $w_{\rm p}(r_{\rm p})$ alone (e.g., [124, 110]). In fact, we achieve very similar constraining power to [112], even though we use the same set of summary statistics. This may imply that the assembly bias signal is stronger at the higher redshifts probed by BGS. Additionally, the purity of the DESI samples may be higher due to the high targeting completeness in the 1% survey, which allows us to avoid having to assign redshifts to untargeted galaxies based upon the nearest neighbors in the sky.

While the HOD model can consistently produce good fits to $w_{\rm p}$ and n simultaneously (possibly to the point of overfitting), incorporating CiC measurements results in mismatches between the model and data in some cases. Although introducing assembly bias parameters has slightly reduced this tension, the $M_r < -21.0$ sample at $z \sim 0.15$ still exhibits a tension of nearly 2σ between our models and the data. This tension is reported in Table 6 and is readily apparent in Figure 26 (though one must use caution when assessing the mismatch by eye since the summary statistics can be strongly covariant).

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

As one example, the HOD we have used in this work assumes that the stellar-to-halo-mass relation has a log-normal scatter, but the UniverseMachine simulations [7] exhibit a slight skew to this scatter in several tested samples. In principle, it is simple to test the addition of one more parameter to allow a skew-log-normal scatter.

Another modification that may be justified is to relax the assumed isotropic NFW distribution of satellite galaxies. This is a common assumption, yet it has long been known that the distribution of subhalos is anisotropic, due to the preferential accretion of mergers

along filaments [126]. Additionally, recent studies have found a significant difference in the radial profile of the halo mass associated with subhalos from NFW [40, 75]. Such modifications would be more complex but will be particularly important as small-scale clustering measurements improve since they are sensitive to the spatial distribution of satellites.

Additionally, we have only tested for assembly bias tied to halo concentration, and have ignored other occupation correlations that may be based upon halo spin or age [27, 98]. Another possibility is that the occupation of satellites is correlated with the occupation of the central in the same halo due to galactic conformity [11, 58]. Both of these possibilities would likely produce similar statistical imprints. However, a primary question to investigate is whether these alternate assumptions lead to a biased inference of HOD parameters such as characteristic halo masses and assembly bias. If so, all of our results could be overly confident[6].

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

Depending on the computing resources available and the dimensionality of the analysis, `galtab` may provide even more drastic speedups. Due to the implementation in JAX, the expensive steps are automatically executed on a GPU when available. Additionally, our framework allows the predictions to be differentiable with respect to HOD parameters (assuming the occupation model is compatible with JAX arrays, for which those available in `halotools` require slight modifications). In principle, this allows for the use of alternative MCMC methods with improved scalability to high-dimensional or strongly covariant posterior estimation, such as Hamiltonian Monte Carlo [82].

Our development of the `galtab` package provides a useful tool for further analyses of the galaxy-halo connection that may require differentiable predictions. By combining these new tools with upcoming enlarged samples from DESI, we anticipate that coming studies

[6] i.e., Zentner Points™

will soon shift focus from mere detections of assembly bias to studying its implications for galaxy formation in much finer detail.

3.7 Galtab Appendix - SHAP Feature Importance Calculations

As briefly discussed in Section 3.3 and plotted in Figure 20, we have roughly quantified the importance of each summary statistic in inferring the HOD model parameters by testing how influential each quantity for machine learning-based predictions. We performed this test using an artificial dataset based upon uniformly sampling 1000 sets of HOD parameters via Latin Hypercube Sampling over the projected one-dimensional 1σ confidence interval of the fiducial fits for the $M_r < -20.5$ threshold sample of [112].

For each of the 1000 sets of HOD parameters, we predicted the values of all of the summary statistics via the methods described in Section 3.4.6. We then trained a scikit-learn [86] random forest regression model to perform the inverse mapping (i.e., predicting HOD parameters from the values of the summary statistics).

We then calculate SHAP feature importance values for each feature (i.e., each quantity used as an input to the random forest). SHAP values are explained in detail in [72]. In brief, they attempt to quantify the amount of "impact" each feature has on model predictions. To be explicit, a large positive SHAP value corresponds to a feature for which increases in the feature value cause large increases in model predictions, and vice versa. This allows us to analyze and distinguish the effects of positive or negative changes in each feature on the model predictions that result.

We show the full beeswarm distribution of SHAP values for each HOD parameter in Figure 28. We assign importance values shown in Figure 20 by taking the mean absolute values of these distributions. Features that have large SHAP importances will correspond to those quantities which are most useful for predicting a given HOD parameter.

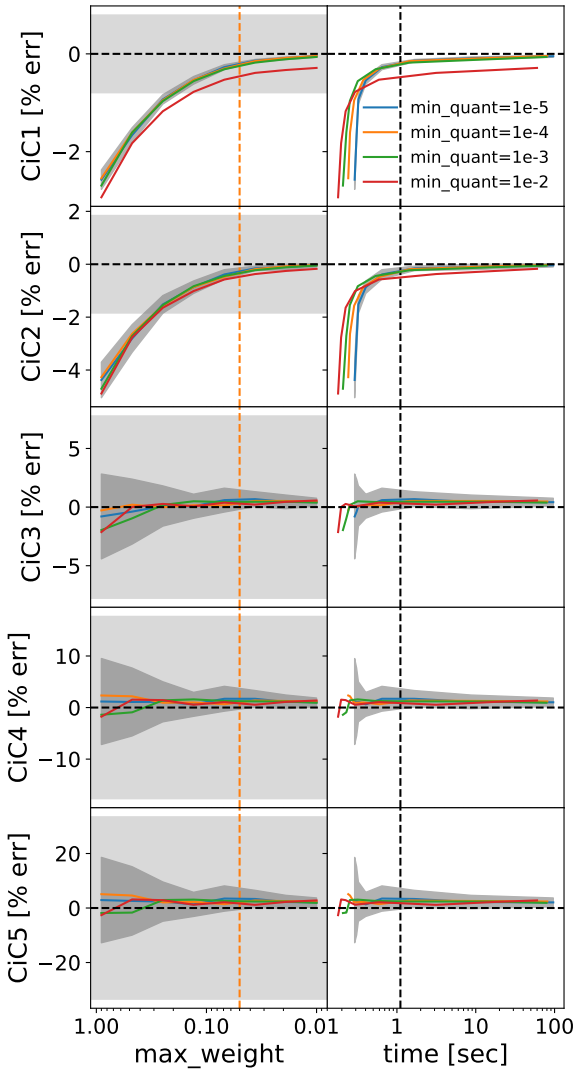

Figure 22: Hyperparameter tuning of `galtab` to achieve sufficient accuracy of CiC moments. The left panels show the tuning of the $W_{\rm max}$ parameter, which is translated to a CPU runtime in the panels on the right side of the figure, with lower values of $W_{\rm max}$ requiring longer times, but achieving higher accuracy. Line colors correspond to the denoted value of $Q_{\rm min}$, the dark grey bands correspond to a standard deviation due to tabulation stochasticity, horizontal dashed lines correspond to truth values from `halotools`, and the light grey band corresponds to a `halotools` standard deviation. The vertical dashed line in the left panels corresponds to our chosen value of $W_{\rm max} = 0.05$, which intentionally yields a similar runtime as `halotools`: approximately one CPU-second, as specified by the vertical dashed line in the right panels.

Table 6: Maximum-likelihood HOD parameters for each sample. For each set of best-fit parameters, the goodness of fit is given by the Akaike Information Criterion (AIC), the chi-squared (χ^2), the degrees of freedom (DoF), the p value corresponding to the probability of measuring $\geq \chi^2$ by chance, and the corresponding z score measure of tension. The fits without CiC, and without assembly bias are included for comparison.

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

Table 7: Confidence intervals of the HOD parameters from the 16th, 50th, and 84th percentiles of the marginalized posteriors. The confidence intervals without CiC constraints, and without assembly bias, are included for comparison.

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

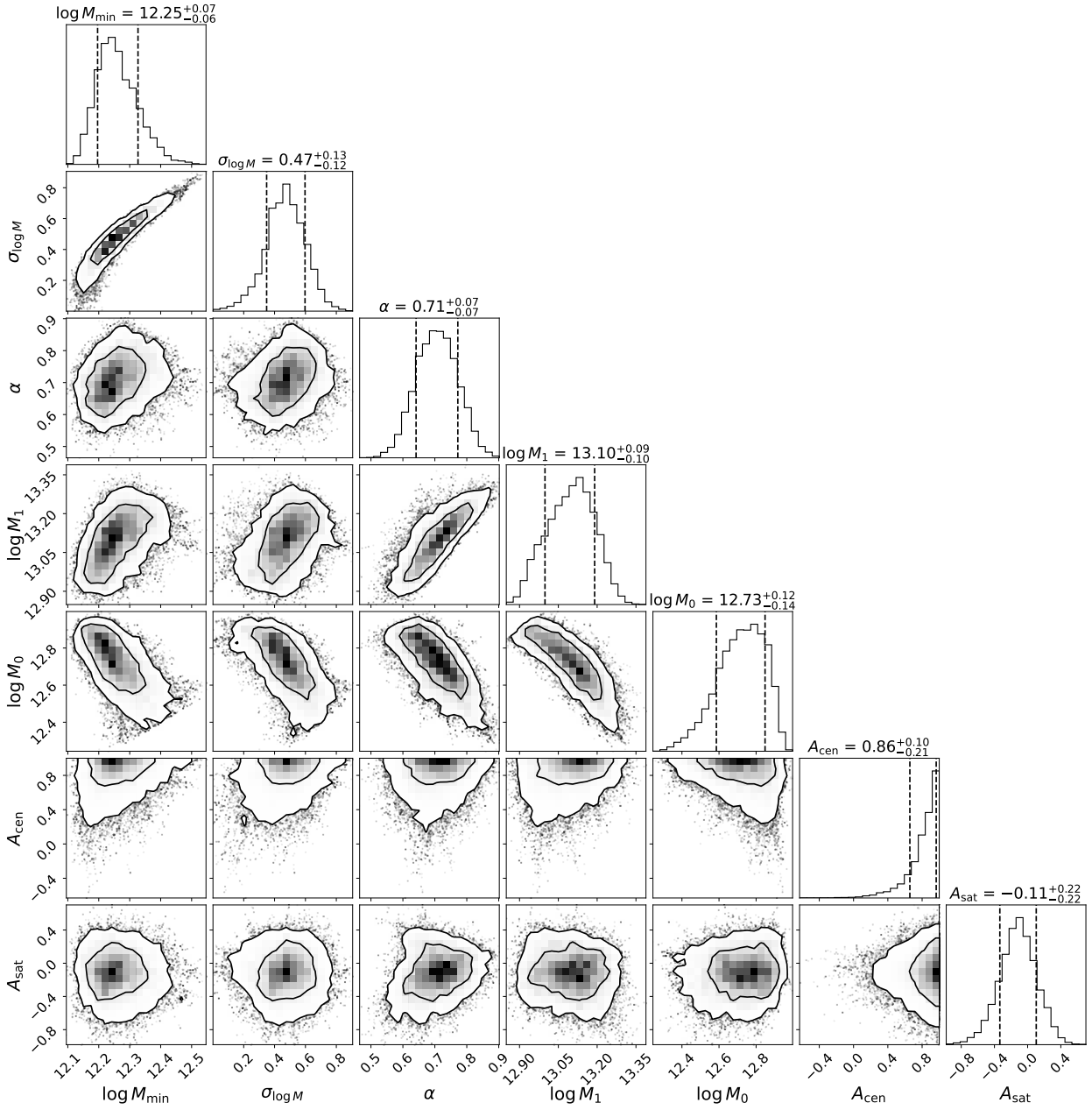

Figure 23: Posterior distribution of the HOD parameters of the -20.5 threshold sample from MCMC sampling. The 68% and 95% confidence regions are displayed by contour lines for each two-dimensional projection, and the 68% confidence intervals are marked with dashed vertical lines for each one-dimensional projection.

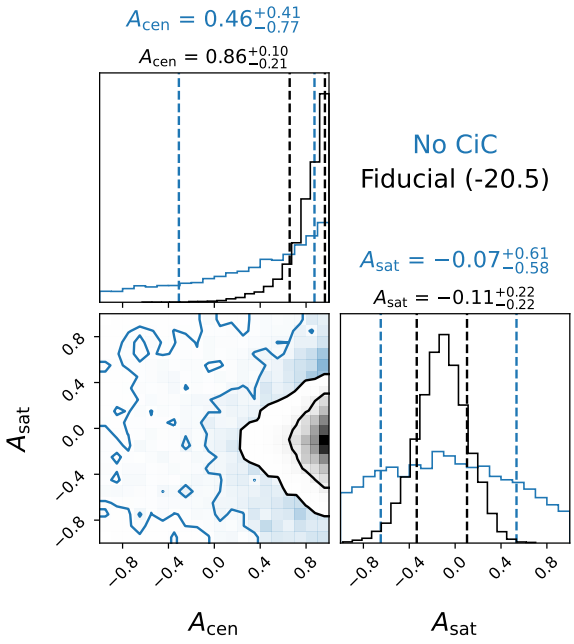

Figure 24: Posterior distribution of the assembly bias parameters of the -20.5 threshold sample from MCMC sampling. Overplotted in blue is the result we obtain without including any constraints from CiC, yielding very little information about assembly bias.

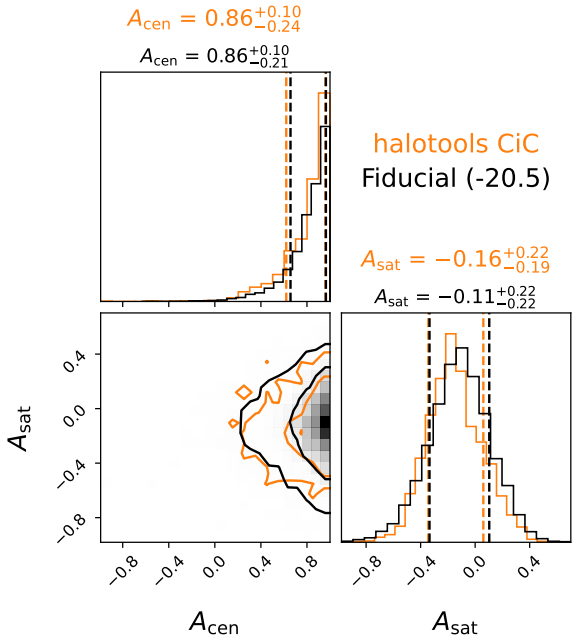

Figure 25: Posterior distribution of the assembly bias parameters of the -20.5 threshold sample from MCMC sampling. Overplotted in orange is the result we obtain from calculating CiC from `halotools` instead of `galtab` for the same number of MCMC iterations. Due to the stochasticity of the `halotools` predictions, its acceptance rate was three times lower in this case, causing much slower posterior convergence.

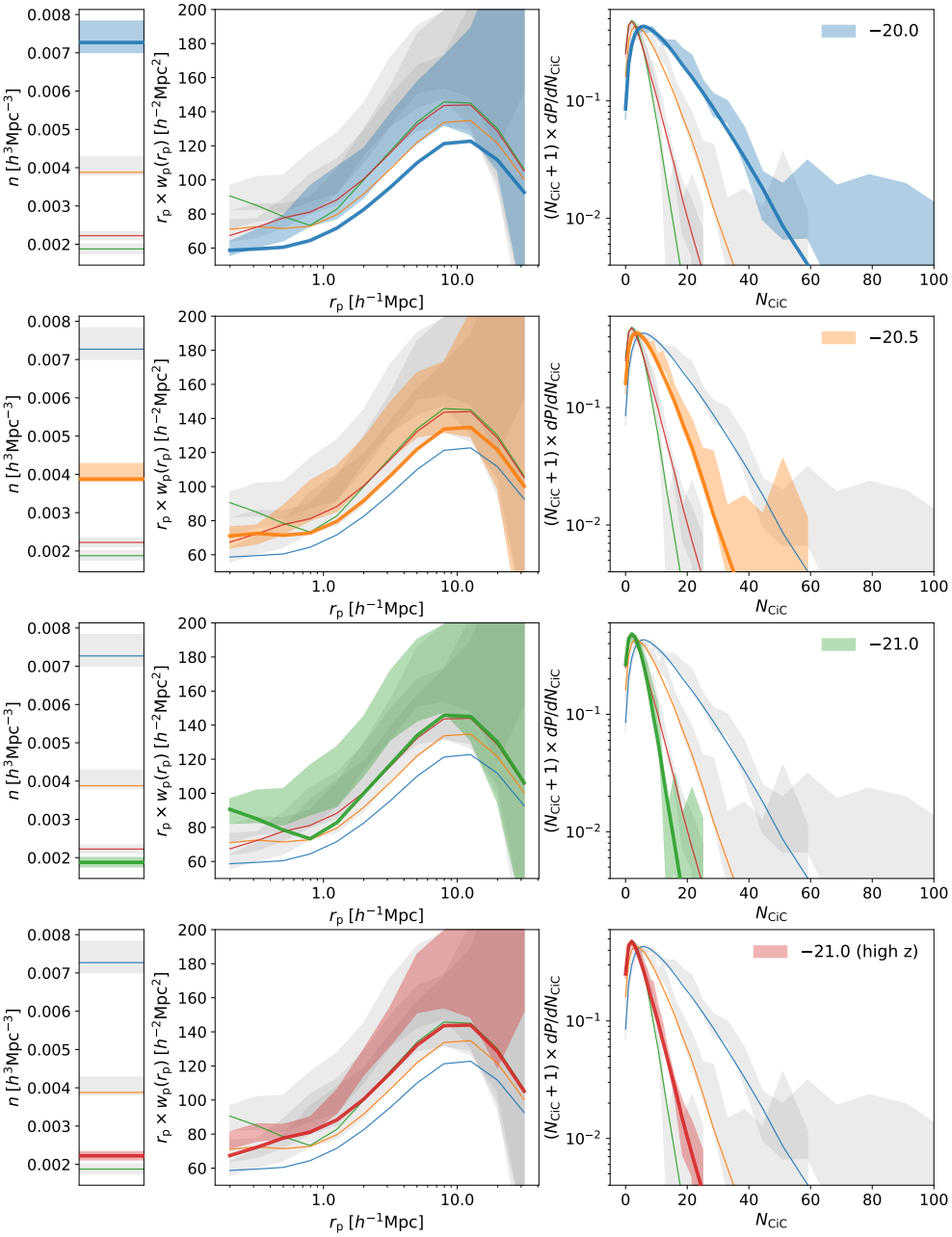

Figure 26: Measurements and our maximum-likelihood predictions of number density (left panels), the projected correlation function (center panels), and the CiC distribution (right panels). The 1σ confidence intervals from the measurements of a given quantity are represented by shaded regions of the color corresponding to the sample, while the maximum-likelihood predictions are represented by solid lines following the same color scheme. The parameters of the best-fit models and their tensions versus the data are reported in Table 6.

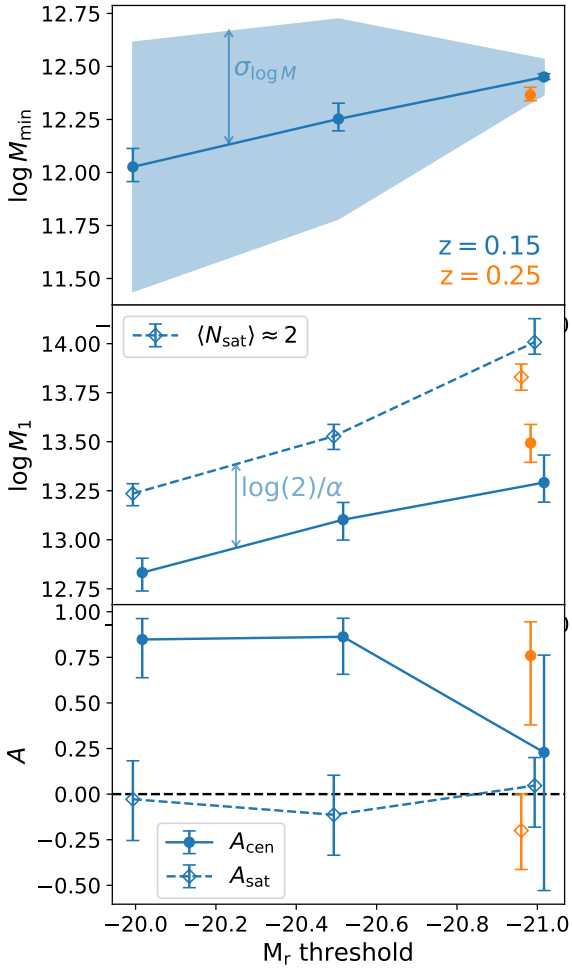

Figure 27: Variation of HOD parameters with luminosity and redshift. Median values of the one-dimensional marginalized posteriors for the characteristic masses, $\log M_{\rm min}$ (top panel) and $\log M_1$ (middle panel) are plotted, as well as the assembly bias parameters $A_{\rm cen}$ and $A_{\rm sat}$ (bottom panel). The capped error bars on these points span the 16th to the 84th percentile of the posterior for a given parameter. Median values derived from our posteriors of other HOD parameters $\sigma_{\log M}$ (top panel) and α (middle panel) are labeled; $\sigma_{\log M}$ characterizes the spread in the M_r-$M_{\rm halo}$ relation, and $\log(2)/\alpha$ characterizes the log-difference between the halo masses corresponding to $\langle N_{\rm sat}\rangle \approx 1$ and $\langle N_{\rm sat}\rangle \approx 2$. We apply small x-offsets to easily distinguish the points, but all M_r thresholds are exactly -20.0, -20.5, or -21.0.

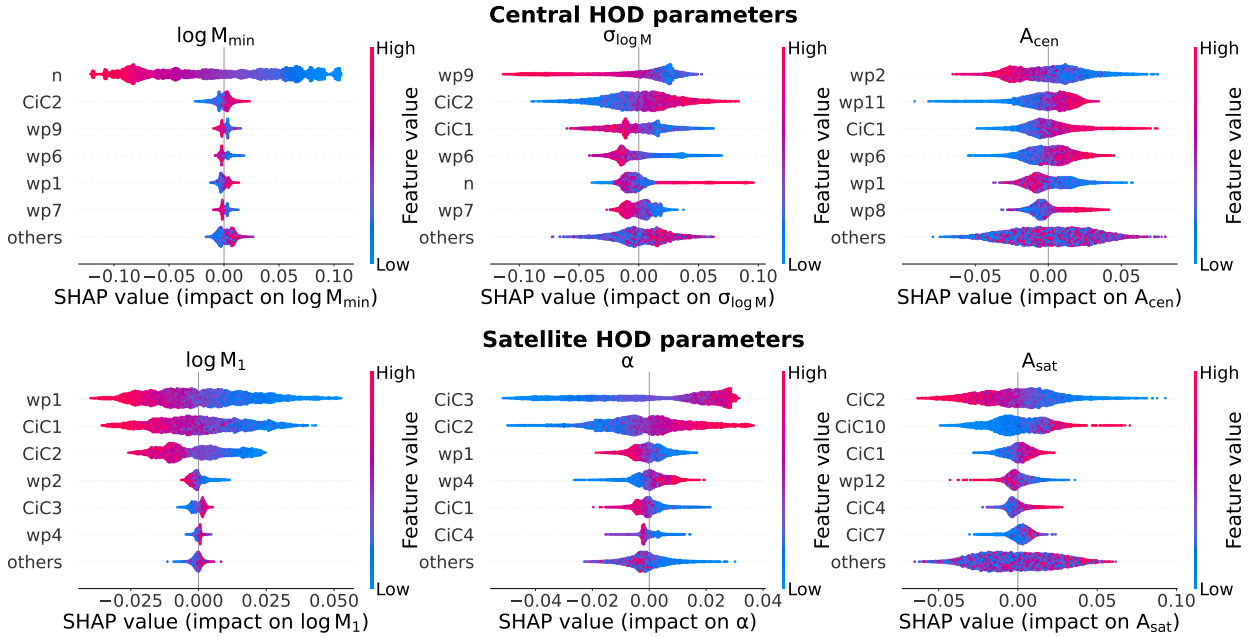

Figure 28: The impact of each of our summary statistics on HOD inference, based upon SHAP feature importances. The upper panel of each sub-figure shows beeswarms of the SHAP values for each feature's impact on predicting the given HOD parameter. Each panel shows the six most important quantities in order of importance, and the panels are organized in the same way those in Figure 20. See Figure 20 for a more condensed version of this information which focuses on the mean absolute SHAP value as an importance metric.

4.0 Conclusions

In this thesis, I have deepened our understanding of the galaxy-halo connection (GHC) by integrating a new dataset, DESI, and enabled the integration of future datasets by developing new methodologies. I have shed light on the role of assembly bias in the statistical relationship between galaxies and their host dark matter halos.

In Chapter 2, I focused on the development of the CLIMBER tool, which allows for the generation of mock galaxy catalogs based on the UniverseMachine, a cutting-edge GHC model. Utilizing CLIMBER, I explored the potential of upcoming surveys, PFS, WAVES, and MOONS, to further deepen our understanding of galaxy clustering and GHC constraints at intermediate redshifts. These mock catalogs will help pave the wave for future observational program proposals calculate their expected scientific outputs. It will also aid in interpreting observational results accurately be providing a tool for science validation through recovering known model parameters. Finally, I showed that the GHC constraints from PFS and MOONS surveys will be limited by cosmic variance, and future extensions should focus on increasing their respective sky areas. See Section 2.6 for a more detailed discussion.

In Chapter 3, I introduced the `galtab` pretabulation code, a powerful tool that drastically improves the efficiency of HOD inference using counts-in-cells statistics — primarily implemented for counts-in-cylinders. Applying `galtab` to early DESI data, I have found a 3σ detection of assembly bias, revealing a connection between galaxy luminosity and the assembly history of their host halo. This finding represents a significant step forward in our understanding of the physical processes governing galaxy formation and highlights the potential of `galtab` in future HOD analyses. See Section 3.6 for a more detailed discussion.

Through these contributions, I have advanced our understanding of the GHC and its implications for galaxy formation and evolution. However, I have also unraveled some new questions that require future investigations. In Chapter 3, I have identified a modest data-model mismatch in the HOD analysis. I have also suggested some areas of the HOD which could be extended to improve its realism with a relatively small degree-of-freedom increase. For example, such extensions could allow for skew to the assumed log-normal scatter in the

stellar-to-halo-mass relation or non-isotropic variations to the assumed NFW distribution of satellites.

Given the unprecedented amount of spectroscopic data coming in the coming decade, future investigations of the GHC will be significantly aided with faster and fully differentiable predictions of galaxy spectra from star-formation histories according to assembly correlation GHC prescriptions. Therefore, I plan to begin contributing to the ongoing development process of the `diffmah` [48], `diffstar` [4], and `dsps` [47] frameworks. This will be my primary responsibility starting this fall when I begin my first post-doctoral position at Argonne National Laboratory. I am looking forward to embarking on this new stage of my research career while continuing to push forward many of the goals of my thesis.

Bibliography

[1] Kevork N. Abazajian, Jennifer K. Adelman-McCarthy, Marcel A. Agüeros, Sahar S. Allam, Carlos Allende Prieto, Deokkeun An, Kurt S. J. Anderson, Scott F. Anderson, James Annis, Neta A. Bahcall, and et al. The Seventh Data Release of the Sloan Digital Sky Survey. ApJS, 182(2):543–558, June 2009.

[2] T. M. C. Abbott, F. B. Abdalla, A. Alarcon, J. Aleksić, S. Allam, S. Allen, A. Amara, J. Annis, J. Asorey, S. Avila, and et al. Dark Energy Survey year 1 results: Cosmological constraints from galaxy clustering and weak lensing. Phys. Rev. D, 98(4):043526, August 2018.

[3] Kurt L. Adelberger, Charles C. Steidel, Mauro Giavalisco, Mark Dickinson, Max Pettini, and Melinda Kellogg. A Counts-in-Cells Analysis Of Lyman-break Galaxies At Redshift Z ~3. ApJ, 505(1):18–24, September 1998.

[4] Alex Alarcon, Andrew P. Hearin, Matthew R. Becker, and Jonás Chaves-Montero. Diffstar: a fully parametric physical model for galaxy assembly history. MNRAS, 518(1):562–584, January 2023.

[5] Lauren Anderson, Eric Aubourg, Stephen Bailey, Dmitry Bizyaev, Michael Blanton, Adam S. Bolton, J. Brinkmann, Joel R. Brownstein, Angela Burden, Antonio J. Cuesta, Luiz A. N. da Costa, Kyle S. Dawson, Roland de Putter, Daniel J. Eisenstein, James E. Gunn, Hong Guo, Jean-Christophe Hamilton, Paul Harding, Shirley Ho, Klaus Honscheid, Eyal Kazin, David Kirkby, Jean-Paul Kneib, Antoine Labatie, Craig Loomis, Robert H. Lupton, Elena Malanushenko, Viktor Malanushenko, Rachel Mandelbaum, Marc Manera, Claudia Maraston, Cameron K. McBride, Kushal T. Mehta, Olga Mena, Francesco Montesano, Demetri Muna, Robert C. Nichol, Sebastián E. Nuza, Matthew D. Olmstead, Daniel Oravetz, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Kaike Pan, John Parejko, Isabelle Pâris, Will J. Percival, Patrick Petitjean, Francisco Prada, Beth Reid, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Lado Samushia, Ariel G. Sánchez, David J. Schlegel, Donald P. Schneider, Claudia G. Scóccola, Hee-Jong Seo, Erin S. Sheldon, Audrey Simmons, Ramin A. Skibba, Michael A. Strauss, Molly E. C. Swanson, Daniel Thomas, Jeremy L. Tinker, Rita Tojeiro, Mariana Vargas Magaña, Licia Verde, Christian Wagner, David A. Wake, Benjamin A. Weaver, David H. Weinberg, Martin White, Xiaoying Xu, Christophe Yèche, Idit Zehavi, and Gong-Bo Zhao. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Release 9 spectroscopic galaxy sample. MNRAS, 427(4):3435–3467, December 2012.

[6] Matthew R. Becker. Connecting Galaxies with Halos Across Cosmic Time: Stellar mass assembly distribution modeling of galaxy statistics. *arXiv e-prints*, page arXiv:1507.03605, July 2015.

[7] Peter Behroozi, Risa H. Wechsler, Andrew P. Hearin, and Charlie Conroy. UNIVERSEMACHINE: The correlation between galaxy growth and dark matter halo assembly from z = 0-10. MNRAS, 488(3):3143–3194, September 2019.

[8] Peter S. Behroozi, Risa H. Wechsler, and Hao-Yi Wu. The ROCKSTAR Phase-space Temporal Halo Finder and the Velocity Offsets of Cluster Cores. ApJ, 762(2):109, January 2013.

[9] Eric F. Bell and Roelof S. de Jong. Stellar Mass-to-Light Ratios and the Tully-Fisher Relation. ApJ, 550(1):212–229, March 2001.

[10] Andreas A. Berlind and David H. Weinberg. The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass. ApJ, 575(2):587–616, August 2002.

[11] Angela M. Berti, Alison L. Coil, Peter S. Behroozi, Daniel J. Eisenstein, Aaron D. Bray, Richard J. Cool, and John Moustakas. PRIMUS: One- and Two-halo Galactic Conformity at 0.2 ¡ z ¡ 1. ApJ, 834(1):87, January 2017.

[12] Angela M. Berti, Alison L. Coil, Andrew P. Hearin, and Peter S. Behroozi. Main-sequence Scatter is Real: The Joint Dependence of Galaxy Clustering on Star Formation and Stellar Mass. AJ, 161(1):49, January 2021.

[13] Florian Beutler, Chris Blake, Matthew Colless, D. Heath Jones, Lister Staveley-Smith, Lachlan Campbell, Quentin Parker, Will Saunders, and Fred Watson. The 6dF Galaxy Survey: baryon acoustic oscillations and the local Hubble constant. MNRAS, 416(4):3017–3032, October 2011.

[14] Davide Bianchi and Will J. Percival. Unbiased clustering estimation in the presence of missing observations. MNRAS, 472(1):1106–1118, November 2017.

[15] Marvin Blank, Andrea V. Macciò, Aaron A. Dutton, and Aura Obreja. NIHAO - XXII. Introducing black hole formation, accretion, and feedback into the NIHAO simulation suite. MNRAS, 487(4):5476–5489, August 2019.

[16] Michael R. Blanton, Matthew A. Bershady, Bela Abolfathi, Franco D. Albareti, Carlos Allende Prieto, Andres Almeida, Javier Alonso-García, Friedrich Anders, Scott F. Anderson, Brett Andrews, and et al. Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe. AJ, 154(1):28, July 2017.

[17] G. R. Blumenthal, S. M. Faber, J. R. Primack, and M. J. Rees. Formation of galaxies and large-scale structure with cold dark matter. Nature, 311:517–525, October 1984.

[18] J. R. Bond, S. Cole, G. Efstathiou, and N. Kaiser. Excursion Set Mass Functions for Hierarchical Gaussian Fluctuations. ApJ, 379:440, October 1991.

[19] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

[20] Leo Breiman. Random Forests. *Machine Learning*, 45:5–32, January 2001.

[21] G. Bruzual and S. Charlot. Stellar population synthesis at the resolution of 2003. MNRAS, 344(4):1000–1028, October 2003.

[22] A. C. Carnall, R. J. McLure, J. S. Dunlop, and R. Davé. Inferring the star formation histories of massive quiescent galaxies with BAGPIPES: evidence for multiple quenching mechanisms. MNRAS, 480(4):4379–4401, November 2018.

[23] Gilles Chabrier. Galactic Stellar and Substellar Initial Mass Function. PASP, 115(809):763–795, July 2003.

[24] Alison L. Coil, Michael R. Blanton, Scott M. Burles, Richard J. Cool, Daniel J. Eisenstein, John Moustakas, Kenneth C. Wong, Guangtun Zhu, James Aird, Rebecca A. Bernstein, Adam S. Bolton, and David W. Hogg. The PRIsm MUlti-object Survey (PRIMUS). I. Survey Overview and Characteristics. ApJ, 741(1):8, November 2011.

[25] Matthew Colless, Gavin Dalton, Steve Maddox, Will Sutherland, Peder Norberg, Shaun Cole, Joss Bland-Hawthorn, Terry Bridges, Russell Cannon, Chris Collins, Warrick Couch, Nicholas Cross, Kathryn Deeley, Roberto De Propris, Simon P. Driver, George Efstathiou, Richard S. Ellis, Carlos S. Frenk, Karl Glazebrook, Carole Jackson, Ofer Lahav, Ian Lewis, Stuart Lumsden, Darren Madgwick, John A. Peacock, Bruce A. Peterson, Ian Price, Mark Seaborne, and Keith Taylor. The 2dF Galaxy Redshift Survey: spectra and redshifts. MNRAS, 328(4):1039–1063, December 2001.

[26] Charlie Conroy and James E. Gunn. The Propagation of Uncertainties in Stellar Population Synthesis Modeling. III. Model Calibration, Comparison, and Evaluation. ApJ, 712(2):833–857, April 2010.

[27] S. Contreras, J. Chaves-Montero, M. Zennaro, and R. E. Angulo. The cosmological dependence of halo and galaxy assembly bias. MNRAS, 507(3):3412–3422, November 2021.

[28] Asantha Cooray and Ravi Sheth. Halo models of large scale structure. Phys. Rep., 372(1):1–129, December 2002.

[29] Robert A. Crain, Joop Schaye, Richard G. Bower, Michelle Furlong, Matthieu Schaller, Tom Theuns, Claudio Dalla Vecchia, Carlos S. Frenk, Ian G. McCarthy, John C. Helly, Adrian Jenkins, Yetli M. Rosas-Guevara, Simon D. M. White, and James W. Trayford. The EAGLE simulations of galaxy formation: calibration of subgrid physics and model variations. MNRAS, 450(2):1937–1961, June 2015.

[30] Elisabete da Cunha, Stéphane Charlot, and David Elbaz. A simple model to interpret the ultraviolet, optical and infrared emission from galaxies. MNRAS, 388(4):1595–1617, August 2008.

[31] Romeel Davé, Daniel Anglés-Alcázar, Desika Narayanan, Qi Li, Mika H. Rafieferantsoa, and Sarah Appleby. SIMBA: Cosmological simulations with black hole growth and feedback. MNRAS, 486(2):2827–2849, June 2019.

[32] M. Davis and P. J. E. Peebles. A survey of galaxy redshifts. V. The two-point position and velocity correlations. ApJ, 267:465–482, April 1983.

[33] Natalí S. M. de Santi, Helen Shao, Francisco Villaescusa-Navarro, L. Raul Abramo, Romain Teyssier, Pablo Villanueva-Domingo, Yueying Ni, Daniel Anglés-Alcázar, Shy Genel, Elena Hernandez-Martinez, Ulrich P. Steinwandel, Christopher C. Lovell, Klaus Dolag, Tiago Castro, and Mark Vogelsberger. Robust field-level likelihood-free inference with galaxies. *arXiv e-prints*, page arXiv:2302.14101, February 2023.

[34] DESI Collaboration, B. Abareshi, J. Aguilar, S. Ahlen, Shadab Alam, David M. Alexander, R. Alfarsy, L. Allen, C. Allende Prieto, O. Alves, and et al. Overview of the Instrumentation for the Dark Energy Spectroscopic Instrument. AJ, 164(5):207, November 2022.

[35] DESI Collaboration, A. G. Adame, J. Aguilar, S. Ahlen, S. Alam, G. Aldering, D. M. Alexander, R. Alfarsy, C. Allende Prieto, M. Alvarez, and et al. The Early Data Release of the Dark Energy Spectroscopic Instrument. *arXiv e-prints*, page arXiv:2306.06308, June 2023.

[36] S. P. Driver, L. J. Davies, M. Meyer, C. Power, A. S. G. Robotham, I. K. Baldry, J. Liske, and P. Norberg. The Wide Area VISTA Extra-Galactic Survey (WAVES). In Nicola R. Napolitano, Giuseppe Longo, Marcella Marconi, Maurizio Paolillo, and Enrichetta Iodice, editors, *The Universe of Digital Sky Surveys*, volume 42, page 205, January 2016.

[37] A. Einstein. Die Grundlage der allgemeinen Relativitätstheorie. *Annalen der Physik*, 354(7):769–822, January 1916.

[38] A. C. Fabian. Observational Evidence of Active Galactic Nuclei Feedback. ARA&A, 50:455–489, September 2012.

[39] G. G. Fazio, J. L. Hora, L. E. Allen, M. L. N. Ashby, P. Barmby, L. K. Deutsch, J. S. Huang, S. Kleiner, M. Marengo, S. T. Megeath, G. J. Melnick, M. A. Pahre, B. M. Patten, J. Polizotti, H. A. Smith, R. S. Taylor, Z. Wang, S. P. Willner, W. F. Hoffmann, J. L. Pipher, W. J. Forrest, C. W. McMurty, C. R. McCreight, M. E. McKelvey, R. E. McMurray, D. G. Koch, S. H. Moseley, R. G. Arendt, J. E. Mentzell, C. T. Marx, P. Losch, P. Mayman, W. Eichhorn, D. Krebs, M. Jhabvala, D. Y. Gezari, D. J. Fixsen, J. Flores, K. Shakoorzadeh, R. Jungo, C. Hakun, L. Workman, G. Karpati, R. Kichak, R. Whitley, S. Mann, E. V. Tollestrup, P. Eisenhardt, D. Stern, V. Gorjian, B. Bhattacharya, S. Carey, B. O. Nelson, W. J. Glaccum, M. Lacy, P. J. Lowrance, S. Laine, W. T. Reach, J. A. Stauffer, J. A. Surace, G. Wilson, E. L. Wright, A. Hoffman, G. Domingo, and M. Cohen. The Infrared Array Camera (IRAC) for the Spitzer Space Telescope. ApJS, 154(1):10–17, September 2004.

[40] Catherine E. Fielder, Yao-Yuan Mao, Andrew R. Zentner, Jeffrey A. Newman, Hao-Yi Wu, and Risa H. Wechsler. Illuminating dark matter halo density profiles without subhaloes. MNRAS, 499(2):2426–2444, December 2020.

[41] Katharina M. Fierlinger, Andreas Burkert, Evangelia Ntormousi, Peter Fierlinger, Marc Schartmann, Alessandro Ballone, Martin G. H. Krause, and Roland Diehl. Stellar feedback efficiencies: supernovae versus stellar winds. MNRAS, 456(1):710–730, February 2016.

[42] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. emcee: The MCMC Hammer. PASP, 125(925):306, March 2013.

[43] Liang Gao, Volker Springel, and Simon D. M. White. The age dependence of halo clustering. MNRAS, 363(1):L66–L70, October 2005.

[44] E. J. Groth and P. J. E. Peebles. Statistical analysis of catalogs of extragalactic objects. VII. Two- and three-point correlation functions for the high-resolution Shane-Wirtanen catalog of galaxies. ApJ, 217:385–405, October 1977.

[45] Qi Guo, Simon White, Michael Boylan-Kolchin, Gabriella De Lucia, Guinevere Kauffmann, Gerard Lemson, Cheng Li, Volker Springel, and Simone Weinmann. From dwarf spheroidals to cD galaxies: simulating the galaxy population in a ΛCDM cosmology. MNRAS, 413(1):101–131, May 2011.

[46] F. K. Hansen, A. J. Banday, and K. M. Górski. Testing the cosmological principle of isotropy: local power-spectrum estimates of the WMAP data. MNRAS, 354(3):641–665, November 2004.

[47] Andrew P. Hearin, Jonás Chaves-Montero, Alex Alarcon, Matthew R. Becker, and Andrew Benson. DSPS: Differentiable stellar population synthesis. MNRAS, 521(2):1741–1756, May 2023.

[48] Andrew P. Hearin, Jonás Chaves-Montero, Mathew R. Becker, and Alex Alarcon. A Differentiable Model of the Assembly of Individual and Populations of Dark Matter Halos. *The Open Journal of Astrophysics*, 4(1):7, July 2021.

[49] Andrew P. Hearin and Douglas F. Watson. The dark side of galaxy colour. MNRAS, 435(2):1313–1324, October 2013.

[50] Andrew P. Hearin, Andrew R. Zentner, Frank C. van den Bosch, Duncan Campbell, and Erik Tollerud. Introducing decorated HODs: modelling assembly bias in the galaxy-halo connection. MNRAS, 460(3):2552–2570, August 2016.

[51] Philip F. Hopkins, Andrew Wetzel, Dušan Kereš, Claude-André Faucher-Giguère, Eliot Quataert, Michael Boylan-Kolchin, Norman Murray, Christopher C. Hayward, Shea Garrison-Kimmel, Cameron Hummels, Robert Feldmann, Paul Torrey, Xiangcheng Ma, Daniel Anglés-Alcázar, Kung-Yi Su, Matthew Orr, Denise Schmitz, Ivanna Escala, Robyn Sanderson, Michael Y. Grudić, Zachary Hafen, Ji-Hoon Kim, Alex Fitts, James S. Bullock, Coral Wheeler, T. K. Chan, Oliver D. Elbert, and Desika Narayanan. FIRE-2 simulations: physics versus numerics in galaxy formation. MNRAS, 480(1):800–863, October 2018.

[52] E. P. Hubble. *Realm of the Nebulae*. 1936.

[53] Tomoaki Ishiyama, Francisco Prada, Anatoly A. Klypin, Manodeep Sinha, R. Benton Metcalf, Eric Jullo, Bruno Altieri, Sofía A. Cora, Darren Croton, Sylvain de la Torre, David E. Millán-Calero, Taira Oogi, José Ruedas, and Cristian A. Vega-Martínez. The Uchuu simulations: Data Release 1 and dark matter halo concentrations. MNRAS, 506(3):4210–4231, September 2021.

[54] Benjamin D. Johnson, Joel Leja, Charlie Conroy, and Joshua S. Speagle. Stellar Population Inference with Prospector. ApJS, 254(2):22, June 2021.

[55] D. Heath Jones, Mike A. Read, Will Saunders, Matthew Colless, Tom Jarrett, Quentin A. Parker, Anthony P. Fairall, Thomas Mauch, Elaine M. Sadler, Fred G. Watson, Donna Burton, Lachlan A. Campbell, Paul Cass, Scott M. Croom, John Dawe, Kristin Fiegert, Leela Frankcombe, Malcolm Hartley, John Huchra, Dionne James, Emma Kirby, Ofer Lahav, John Lucey, Gary A. Mamon, Lesa Moore, Bruce A. Peterson, Sayuri Prior, Dominique Proust, Ken Russell, Vicky Safouris, Ken-Ichi Wakamatsu, Eduard Westra, and Mary Williams. The 6dF Galaxy Survey: final redshift release (DR3) and southern large-scale structures. MNRAS, 399(2):683–698, October 2009.

[56] N. Kaiser. On the spatial correlations of Abell clusters. ApJ, 284:L9–L12, September 1984.

[57] M. Kajisawa, T. Ichikawa, T. Yamada, Y. K. Uchimoto, T. Yoshikawa, M. Akiyama, and M. Onodera. MOIRCS Deep Survey. VIII. Evolution of Star Formation Activity as a Function of Stellar Mass in Galaxies Since z ~3. ApJ, 723(1):129–145, November 2010.

[58] Guinevere Kauffmann, Cheng Li, Wei Zhang, and Simone Weinmann. A re-examination of galactic conformity and a comparison with semi-analytic models of galaxy formation. MNRAS, 430(2):1447–1456, April 2013.

[59] Daniel D. Kelson, Rik J. Williams, Alan Dressler, Patrick J. McCarthy, Stephen A. Shectman, John S. Mulchaey, Edward V. Villanueva, Jeffrey D. Crane, and Ryan F. Quadri. The Carnegie-Spitzer-IMACS Redshift Survey of Galaxy Evolution since z = 1.5. I. Description and Methodology. ApJ, 783(2):110, March 2014.

[60] Anatoly Klypin, Gustavo Yepes, Stefan Gottlöber, Francisco Prada, and Steffen Heß. MultiDark simulations: the story of dark matter halo concentrations and density profiles. MNRAS, 457(4):4340–4359, April 2016.

[61] Anatoly A. Klypin, Sebastian Trujillo-Gomez, and Joel Primack. Dark Matter Halos in the Standard Cosmological Model: Results from the Bolshoi Simulation. ApJ, 740(2):102, October 2011.

[62] Alexander Knebe and Volkmar Wießner. Triaxial versus Spherical Dark Matter Halo Profiles. PASA, 23(3):125–128, November 2006.

[63] Andrey V. Kravtsov, Andreas A. Berlind, Risa H. Wechsler, Anatoly A. Klypin, Stefan Gottlöber, Brandon Allgood, and Joel R. Primack. The Dark Side of the Halo Occupation Distribution. ApJ, 609(1):35–49, July 2004.

[64] Mariska Kriek, Pieter G. van Dokkum, Ivo Labbé, Marijn Franx, Garth D. Illingworth, Danilo Marchesini, and Ryan F. Quadri. An Ultra-Deep Near-Infrared Spectrum of a Compact Quiescent Galaxy at z = 2.2. ApJ, 700(1):221–231, July 2009.

[65] Stephen D. Landy and Alexander S. Szalay. Bias and Variance of Angular Correlation Functions. ApJ, 412:64, July 1993.

[66] Alexie Leauthaud, Jeremy Tinker, Kevin Bundy, Peter S. Behroozi, Richard Massey, Jason Rhodes, Matthew R. George, Jean-Paul Kneib, Andrew Benson, Risa H. Wechsler, Michael T. Busha, Peter Capak, Marina Cortês, Olivier Ilbert, Anton M. Koekemoer, Oliver Le Fèvre, Simon Lilly, Henry J. McCracken, Mara Salvato, Tim Schrabback, Nick Scoville, Tristan Smith, and James E. Taylor. 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. ApJ, 744(2):159, January 2012.

[67] Benjamin V. Lehmann, Yao-Yuan Mao, Matthew R. Becker, Samuel W. Skillman, and Risa H. Wechsler. The Concentration Dependence of the Galaxy-Halo Connection: Modeling Assembly Bias with Abundance Matching. ApJ, 834(1):37, January 2017.

[68] Joel Leja, Benjamin D. Johnson, Charlie Conroy, Pieter van Dokkum, Joshua S. Speagle, Gabriel Brammer, Ivelina Momcheva, Rosalind Skelton, Katherine E. Whitaker, Marijn Franx, and Erica J. Nelson. An Older, More Quiescent Universe from Panchromatic SED Fitting of the 3D-HST Survey. ApJ, 877(2):140, June 2019.

[69] Sarah K. Leslie, Eva Schinnerer, Daizhong Liu, Benjamin Magnelli, Hiddo Algera, Alexander Karim, Iary Davidzon, Ghassem Gozaliasl, Eric F. Jiménez-Andrade, Philipp Lang, Mark T. Sargent, Mladen Novak, Brent Groves, Vernesa Smolčić, Giovanni Zamorani, Mattia Vaccari, Andrew Battisti, Eleni Vardoulaki, Yingjie Peng,

and Jeyhan Kartaltepe. The VLA-COSMOS 3 GHz Large Project: Evolution of Specific Star Formation Rates out to z $\sim$ 5. ApJ, 899(1):58, August 2020.

[70] Shihong Liao, Liang Gao, Carlos S. Frenk, Robert J. J. Grand, Qi Guo, Facundo A. Gómez, Federico Marinacci, Rüdiger Pakmor, Shi Shao, and Volker Springel. Ultra-diffuse galaxies in the Auriga simulations. MNRAS, 490(4):5182–5195, December 2019.

[71] Tobias J. Looser, Simon J. Lilly, Larry P. T. Sin, Bruno M. B. Henriques, Roberto Maiolino, and Michele Cirasuolo. Optimizing high-redshift galaxy surveys for environmental information. MNRAS, 504(2):3029–3057, June 2021.

[72] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *arXiv e-prints*, page arXiv:1705.07874, May 2017.

[73] R. Maiolino, M. Cirasuolo, J. Afonso, F. E. Bauer, R. Bowler, O. Cucciati, E. Daddi, G. De Lucia, C. Evans, H. Flores, A. Gargiulo, B. Garilli, P. Jablonka, M. Jarvis, J. P. Kneib, S. Lilly, T. Looser, M. Magliocchetti, Z. Man, F. Mannucci, S. Maurogordato, R. J. McLure, P. Norberg, P. Oesch, E. Oliva, S. Paltani, C. Pappalardo, Y. Peng, L. Pentericci, L. Pozzetti, A. Renzini, M. Rodrigues, F. Royer, S. Serjeant, L. Vanzi, V. Wild, and G. Zamorani. MOONRISE: The Main MOONS GTO Extragalactic Survey. *The Messenger*, 180:24–29, June 2020.

[74] D. Christopher Martin, James Fanson, David Schiminovich, Patrick Morrissey, Peter G. Friedman, Tom A. Barlow, Tim Conrow, Robert Grange, Patrick N. Jelinsky, Bruno Milliard, Oswald H. W. Siegmund, Luciana Bianchi, Yong-Ik Byun, Jose Donas, Karl Forster, Timothy M. Heckman, Young-Wook Lee, Barry F. Madore, Roger F. Malina, Susan G. Neff, R. Michael Rich, Todd Small, Frank Surber, Alex S. Szalay, Barry Welsh, and Ted K. Wyder. The Galaxy Evolution Explorer: A Space Ultraviolet Survey Mission. ApJ, 619(1):L1–L6, January 2005.

[75] Lorena Mezini, Catherine E. Fielder, Andrew R. Zentner, Yao-Yuan Mao, Kuan Wang, and Hao-Yi Wu. The Influence of Subhaloes on Host Halo Properties. *arXiv e-prints*, page arXiv:2304.13809, April 2023.

[76] H. J. Mo and S. D. M. White. An analytic model for the spatial clustering of dark matter haloes. MNRAS, 282(2):347–361, September 1996.

[77] Surhud More, Frank C. van den Bosch, Marcello Cacciato, Ramin Skibba, H. J. Mo, and Xiaohu Yang. Satellite kinematics - III. Halo masses of central galaxies in SDSS. MNRAS, 410(1):210–226, January 2011.

[78] John Moustakas, Alison L. Coil, James Aird, Michael R. Blanton, Richard J. Cool, Daniel J. Eisenstein, Alexander J. Mendez, Kenneth C. Wong, Guangtun Zhu, and Stéphane Arnouts. PRIMUS: Constraints on Star Formation Quenching and Galaxy Merging, and the Evolution of the Stellar Mass Function from z = 0-1. ApJ, 767(1):50, April 2013.

[79] Adam Muzzin, Danilo Marchesini, Mauro Stefanon, Marijn Franx, Bo Milvang-Jensen, James S. Dunlop, J. P. U. Fynbo, Gabriel Brammer, Ivo Labbé, and Pieter van Dokkum. A Public K_s -selected Catalog in the COSMOS/ULTRAVISTA Field: Photometry, Photometric Redshifts, and Stellar Population Parameters. ApJS, 206(1):8, May 2013.

[80] Adam Muzzin, Danilo Marchesini, Pieter G. van Dokkum, Ivo Labbé, Mariska Kriek, and Marijn Franx. A Near-Infrared Spectroscopic Survey of K-Selected Galaxies at z~2.3: Comparison of Stellar Population Synthesis Codes and Constraints from the Rest-Frame NIR. ApJ, 701(2):1839–1864, August 2009.

[81] Julio F. Navarro, Carlos S. Frenk, and Simon D. M. White. The Structure of Cold Dark Matter Halos. ApJ, 462:563, May 1996.

[82] Radford Neal. MCMC Using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. 2011.

[83] Dylan Nelson, Volker Springel, Annalisa Pillepich, Vicente Rodriguez-Gomez, Paul Torrey, Shy Genel, Mark Vogelsberger, Ruediger Pakmor, Federico Marinacci, Rainer Weinberger, Luke Kelley, Mark Lovell, Benedikt Diemer, and Lars Hernquist. The IllustrisTNG simulations: public data release. *Computational Astrophysics and Cosmology*, 6(1):2, May 2019.

[84] Christine O'Donnell, Peter Behroozi, and Surhud More. Observing correlations between dark matter accretion and galaxy growth - I. Recent star formation activity in isolated Milky Way-mass galaxies. MNRAS, 501(1):1253–1272, February 2021.

[85] Alan N. Pearl, Rachel Bezanson, Andrew R. Zentner, Jeffrey A. Newman, Andy D. Goulding, Katherine E. Whitaker, Sean D. Johnson, and Jenny E. Greene. CLIMBER: Galaxy-Halo Connection Constraints from Next-generation Surveys. ApJ, 925(2):180, February 2022.

[86] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles

Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, October 2011.

[87] P. J. E. Peebles. *The large-scale structure of the universe*. 1980.

[88] R. Penrose. A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, July 1955.

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

[90] Francisco Prada, Anatoly A. Klypin, Antonio J. Cuesta, Juan E. Betancort-Rijo, and Joel Primack. Halo concentrations in the standard Λ cold dark matter cosmology. MNRAS, 423(4):3018–3030, July 2012.

[91] William H. Press and Paul Schechter. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. ApJ, 187:425–438, February 1974.

[92] Rachel M. Reddick, Risa H. Wechsler, Jeremy L. Tinker, and Peter S. Behroozi. The Connection between Galaxies and Dark Matter Structures in the Local Universe. ApJ, 771(1):30, July 2013.

[93] Beth A. Reid and David N. Spergel. Constraining the Luminous Red Galaxy Halo Occupation Distribution Using Counts-In-Cylinders. ApJ, 698(1):143–154, June 2009.

[94] Adam G. Riess, Alexei V. Filippenko, Peter Challis, Alejandro Clocchiatti, Alan Diercks, Peter M. Garnavich, Ron L. Gilliland, Craig J. Hogan, Saurabh Jha, Robert P. Kirshner, B. Leibundgut, M. M. Phillips, David Reiss, Brian P. Schmidt, Robert A. Schommer, R. Chris Smith, J. Spyromilio, Christopher Stubbs, Nicholas B. Suntzeff, and John Tonry. Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. AJ, 116(3):1009–1038, September 1998.

[95] Aldo Rodríguez-Puebla, Vladimir Avila-Reese, Xiaohu Yang, Sebastien Foucaud, Niv Drory, and Y. P. Jing. The Stellar-to-Halo Mass Relation of Local Galaxies Segregates by Color. ApJ, 799(2):130, February 2015.

[96] Aldo Rodríguez-Puebla, Joel R. Primack, Peter Behroozi, and S. M. Faber. Is main-sequence galaxy star formation controlled by halo mass accretion? MNRAS, 455(3):2592–2606, January 2016.

[97] Edwin E. Salpeter. The Luminosity Function and Stellar Evolution. ApJ, 121:161, January 1955.

[98] Gabriela Sato-Polito, Antonio D. Montero-Dorta, L. Raul Abramo, Francisco Prada, and Anatoly Klypin. The dependence of halo bias on age, concentration, and spin. MNRAS, 487(2):1570–1579, August 2019.

[99] Joop Schaye, Robert A. Crain, Richard G. Bower, Michelle Furlong, Matthieu Schaller, Tom Theuns, Claudio Dalla Vecchia, Carlos S. Frenk, I. G. McCarthy, John C. Helly, Adrian Jenkins, Y. M. Rosas-Guevara, Simon D. M. White, Maarten Baes, C. M. Booth, Peter Camps, Julio F. Navarro, Yan Qu, Alireza Rahmati, Till

Sawala, Peter A. Thomas, and James Trayford. The EAGLE project: simulating the evolution and assembly of galaxies and their environments. MNRAS, 446(1):521–554, January 2015.

[100] Evan E. Schneider and Brant E. Robertson. Introducing CGOLS: The Cholla Galactic Outflow Simulation Suite. ApJ, 860(2):135, June 2018.

[101] N. Scoville, H. Aussel, M. Brusa, P. Capak, C. M. Carollo, M. Elvis, M. Giavalisco, L. Guzzo, G. Hasinger, C. Impey, J. P. Kneib, O. LeFevre, S. J. Lilly, B. Mobasher, A. Renzini, R. M. Rich, D. B. Sanders, E. Schinnerer, D. Schminovich, P. Shopbell, Y. Taniguchi, and N. D. Tyson. The Cosmic Evolution Survey (COSMOS): Overview. ApJS, 172(1):1–8, September 2007.

[102] Manodeep Sinha and Lehman H. Garrison. CORRFUNC - a suite of blazing fast correlation functions on the CPU. MNRAS, 491(2):3022–3041, January 2020.

[103] Zachary Slepian, Daniel J. Eisenstein, Florian Beutler, Chia-Hsun Chuang, Antonio J. Cuesta, Jian Ge, Héctor Gil-Marín, Shirley Ho, Francisco-Shu Kitaura, Cameron K. McBride, Robert C. Nichol, Will J. Percival, Sergio Rodríguez-Torres, Ashley J. Ross, Román Scoccimarro, Hee-Jong Seo, Jeremy Tinker, Rita Tojeiro, and Mariana Vargas-Magaña. The large-scale three-point correlation function of the SDSS BOSS DR12 CMASS galaxies. MNRAS, 468(1):1070–1083, June 2017.

[104] Rachel S. Somerville and Romeel Davé. Physical Models of Galaxy Formation in a Cosmological Framework. ARA&A, 53:51–113, August 2015.

[105] Volker Springel. The cosmological simulation code GADGET-2. MNRAS, 364(4):1105–1134, December 2005.

[106] Kate Storey-Fisher, Jeremy Tinker, Zhongxu Zhai, Joseph DeRose, Risa H. Wechsler, and Arka Banerjee. The Aemulus Project VI: Emulation of beyond-standard galaxy clustering statistics to improve cosmological constraints. *arXiv e-prints*, page arXiv:2210.03203, October 2022.

[107] Masahiro Takada, Richard S. Ellis, Masashi Chiba, Jenny E. Greene, Hiroaki Aihara, Nobuo Arimoto, Kevin Bundy, Judith Cohen, Olivier Doré, Genevieve Graves, James E. Gunn, Timothy Heckman, Christopher M. Hirata, Paul Ho, Jean-Paul Kneib, Olivier Le Fèvre, Lihwai Lin, Surhud More, Hitoshi Murayama, Tohru Nagao, Masami Ouchi, Michael Seiffert, John D. Silverman, Laerte Sodré, David N.

Spergel, Michael A. Strauss, Hajime Sugai, Yasushi Suto, Hideki Takami, and Rosemary Wyse. Extragalactic science, cosmology, and Galactic archaeology with the Subaru Prime Focus Spectrograph. PASJ, 66(1):R1, February 2014.

[108] Jeremy L. Tinker, ChangHoon Hahn, Yao-Yuan Mao, Andrew R. Wetzel, and Charlie Conroy. Halo histories versus galaxy properties at z = 0 II: large-scale galactic conformity. MNRAS, 477(1):935–945, June 2018.

[109] Adam R. Tomczak, Ryan F. Quadri, Kim-Vy H. Tran, Ivo Labbé, Caroline M. S. Straatman, Casey Papovich, Karl Glazebrook, Rebecca Allen, Gabreil B. Brammer, Michael Cowley, Mark Dickinson, David Elbaz, Hanae Inami, Glenn G. Kacprzak, Glenn E. Morrison, Themiya Nanayakkara, S. Eric Persson, Glen A. Rees, Brett Salmon, Corentin Schreiber, Lee R. Spitler, and Katherine E. Whitaker. The SFR-M* Relation and Empirical Star-Formation Histories from ZFOURGE* at 0.5 ¡ z ¡ 4. ApJ, 817(2):118, February 2016.

[110] Mohammadjavad Vakili and ChangHoon Hahn. How Are Galaxies Assigned to Halos? Searching for Assembly Bias in the SDSS Galaxy Clustering. ApJ, 872(1):115, February 2019.

[111] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, February 2020.

[112] Kuan Wang, Yao-Yuan Mao, Andrew R. Zentner, Hong Guo, Johannes U. Lange, Frank C. van den Bosch, and Lorena Mezini. Evidence of galaxy assembly bias in SDSS DR7 galaxy samples from count statistics. MNRAS, 516(3):4003–4024, November 2022.

[113] Kuan Wang, Yao-Yuan Mao, Andrew R. Zentner, Frank C. van den Bosch, Johannes U. Lange, Chad M. Schafer, Antonio S. Villarreal, Andrew P. Hearin, and Duncan Campbell. How to optimally constrain galaxy assembly bias: supplement projected correlation functions with count-in-cells statistics. MNRAS, 488(3):3541–3567, September 2019.

[114] Risa H. Wechsler and Jeremy L. Tinker. The Connection Between Galaxies and Their Dark Matter Halos. ARA&A, 56:435–487, September 2018.

[115] M. W. Werner, T. L. Roellig, F. J. Low, G. H. Rieke, M. Rieke, W. F. Hoffmann, E. Young, J. R. Houck, B. Brandl, G. G. Fazio, J. L. Hora, R. D. Gehrz, G. Helou, B. T. Soifer, J. Stauffer, J. Keene, P. Eisenhardt, D. Gallagher, T. N. Gautier, W. Irace, C. R. Lawrence, L. Simmons, J. E. Van Cleve, M. Jura, E. L. Wright, and D. P. Cruikshank. The Spitzer Space Telescope Mission. ApJS, 154(1):1–9, September 2004.

[116] Andrew R. Wetzel and Daisuke Nagai. The Physical Nature of the Cosmic Accretion of Baryons and Dark Matter into Halos and Their Galaxies. ApJ, 808(1):40, July 2015.

[117] Katherine E. Whitaker, Marijn Franx, Joel Leja, Pieter G. van Dokkum, Alaina Henry, Rosalind E. Skelton, Mattia Fumagalli, Ivelina G. Momcheva, Gabriel B. Brammer, Ivo Labbé, Erica J. Nelson, and Jane R. Rigby. Constraining the Low-mass Slope of the Star Formation Sequence at 0.5 ¡ z ¡ 2.5. ApJ, 795(2):104, November 2014.

[118] Katherine E. Whitaker, Alexandra Pope, Ryan Cybulski, Caitlin M. Casey, Gergö Popping, and Min S. Yun. The Constant Average Relationship between Dust-obscured Star Formation and Stellar Mass from z = 0 to z = 2.5. ApJ, 850(2):208, December 2017.

[119] S. D. M. White. The hierarchy of correlation functions and its relation to other measures of galaxy clustering. MNRAS, 186:145–154, January 1979.

[120] S. D. M. White and M. J. Rees. Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering. MNRAS, 183:341–358, May 1978.

[121] Simon D. M. White and Carlos S. Frenk. Galaxy Formation through Hierarchical Clustering. ApJ, 379:52, September 1991.

[122] Idit Zehavi, Zheng Zheng, David H. Weinberg, Joshua A. Frieman, Andreas A. Berlind, Michael R. Blanton, Roman Scoccimarro, Ravi K. Sheth, Michael A. Strauss, Issha Kayo, Yasushi Suto, Masataka Fukugita, Osamu Nakamura, Neta A. Bahcall, Jon Brinkmann, James E. Gunn, Greg S. Hennessy, Željko Ivezić, Gillian R. Knapp, Jon Loveday, Avery Meiksin, David J. Schlegel, Donald P. Schneider, Istvan Szapudi, Max Tegmark, Michael S. Vogeley, Donald G. York, and SDSS Collaboration. The Luminosity and Color Dependence of the Galaxy Correlation Function. ApJ, 630(1):1–27, September 2005.

[123] Andrew R. Zentner. The Excursion Set Theory of Halo Mass Functions, Halo Clustering, and Halo Growth. *International Journal of Modern Physics D*, 16(5):763–815, January 2007.

[124] Andrew R. Zentner, Andrew Hearin, Frank C. van den Bosch, Johannes U. Lange, and Antonio Villarreal. Constraints on assembly bias from galaxy clustering. MNRAS, 485(1):1196–1209, May 2019.

[125] Andrew R. Zentner, Andrew P. Hearin, and Frank C. van den Bosch. Galaxy assembly bias: a significant source of systematic error in the galaxy-halo relationship. MNRAS, 443(4):3044–3067, October 2014.

[126] Andrew R. Zentner, Andrey V. Kravtsov, Oleg Y. Gnedin, and Anatoly A. Klypin. The Anisotropic Distribution of Galactic Satellites. ApJ, 629(1):219–232, August 2005.

[127] Zheng Zheng, Alison L. Coil, and Idit Zehavi. Galaxy Evolution from Halo Occupation Distribution Modeling of DEEP2 and SDSS Galaxy Clustering. ApJ, 667(2):760–779, October 2007.

[128] Ying Zu and Rachel Mandelbaum. Mapping stellar content to dark matter haloes using galaxy clustering and galaxy-galaxy lensing in the SDSS DR7. MNRAS, 454(2):1161–1191, December 2015.

[129] Ying Zu and Rachel Mandelbaum. Mapping stellar content to dark matter haloes - III. Environmental dependence and conformity of galaxy colours. MNRAS, 476(2):1637–1653, May 2018.

[130] Fritz Zwicky. *Morphological astronomy*. 1957.