Multiscale statistical modeling of large-scale structure:

from baryon acoustic oscillations to galaxy formation histories

by

Jayashree Behera

Integrated M.S., National Institute of Technology Rourkela, India

2016

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Physics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Abstract

This dissertation presents a comprehensive study of the large-scale structure of the Universe through two interrelated avenues: the modeling of Baryon Acoustic Oscillations (BAOs) in higher-order statistics, and the prediction of galaxy formation histories using machine learning with semi-stochastic corrections along with the validation of such models using dark matter-only simulations. The overarching goal is to improve the extraction of cosmological information and the realism of galaxy property predictions by leveraging both simulation-driven statistics and data-driven inference.

The first part (Chapters 1 and 2) focuses on modeling the BAO feature in the bispectrum—the Fourier-space analogue of the three-point correlation function. Unlike the power spectrum, the bispectrum captures non-Gaussianity and mode coupling from nonlinear gravitational evolution and galaxy bias. We develop a "wiggle-only" framework that isolates the oscillatory BAO component from the broadband bispectrum shape. Using `GLAM` N-body simulations (with and without BAO), we analyze signal evolution across redshifts and triangle configurations. Our aligned-template method enables percent-level recovery of the BAO dilation parameter α, accounting for template systematics. Robustness and precision are demonstrated using 1000 $h^{-3}Gpc^3$ of realizations, showing the bispectrum's potential to complement two-point statistics in surveys DESI and Euclid.

The second part of the thesis (Chapters 3, 4 and 5) addresses the limitations of machine learning in modeling the baryonic assembly history of galaxies within dark matter halos. Traditional models tend to smooth over short-timescale variability in star formation and chemical enrichment histories, due to their architectural bias toward minimizing global loss functions. We propose a novel correction scheme that decomposes galaxy histories in Fourier space, identifies missing high-frequency power, and re-injects statistically consistent fluctuations into the predicted histories. This framework is applied to galaxies from the

IllustrisTNG simulation. The modified histories restore variability, improving accuracy on observables such as the stellar-halo mass and mass-metallicity relation, spectral energy distributions (SEDs), and photometric color distributions. The corrections particularly enhance the bimodality of galaxy colors, the scatter in metallicity at fixed mass, and the recovery of quenched populations in satellite systems. This framework is further extended to dark matter-only simulations to evaluate the generalizability of neural network-based predictions in the absence of baryonic training data. We assess the role of mass accretion history, halo concentration, and cosmic environment in enabling accurate galaxy property inference, and demonstrate how semi-stochastic corrections improve the fidelity of the generated histories.

These components address distinct but complementary aspects of large-scale structure modeling. The bispectrum-based analysis enhances the precision and robustness of cosmological distance measurements by leveraging higher-order statistics, unlocking additional information from the nonlinear regime of structure formation. In contrast, the semi-stochastic modeling of galaxy formation histories improves the realism of mock galaxy catalogs, correcting biases in data-driven predictions and enriching their utility in forward-modeling frameworks. Additionally, the thesis includes several complementary studies (Chapter 6): an angular multipole analysis of the bispectrum for probing anisotropic clustering, the mitigation of imaging systematics in galaxy survey data, and spectroscopic classification of AGN and QSO targets within DESI. Together, these efforts support the broader goal of connecting theory, simulation, and observation — advancing both the accuracy of cosmological inference and the fidelity of galaxy population modeling for current and future surveys.

Note on Chapter-End Quotations: In keeping with my roots in Odisha, India, each chapter concludes with a short quotation in Odia, my native language. These reflective excerpts are drawn from Odia literature, classical philosophy, and folk tradition, and are intended to symbolically parallel the cosmological and astrophysical ideas explored in this work. Each quote is followed by an English translation to maintain accessibility while honoring the worldview that shaped my earliest engagement with science.

Multiscale statistical modeling of large-scale structure: from baryon acoustic oscillations to galaxy formation histories

by

Jayashree Behera

Integrated M.S., National Institute of Technology Rourkela, India

2016

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Physics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Approved by:

Major Professor
Lado Samushia

Copyright

Abstract

This dissertation presents a comprehensive study of the large-scale structure of the Universe through two interrelated avenues: the modeling of Baryon Acoustic Oscillations (BAOs) in higher-order statistics, and the prediction of galaxy formation histories using machine learning with semi-stochastic corrections along with the validation of such models using dark matter-only simulations. The overarching goal is to improve the extraction of cosmological information and the realism of galaxy property predictions by leveraging both simulation-driven statistics and data-driven inference.

The first part (Chapters 1 and 2) focuses on modeling the BAO feature in the bispectrum—the Fourier-space analogue of the three-point correlation function. Unlike the power spectrum, the bispectrum captures non-Gaussianity and mode coupling from nonlinear gravitational evolution and galaxy bias. We develop a "wiggle-only" framework that isolates the oscillatory BAO component from the broadband bispectrum shape. Using `GLAM` N-body simulations (with and without BAO), we analyze signal evolution across redshifts and triangle configurations. Our aligned-template method enables percent-level recovery of the BAO dilation parameter α, accounting for template systematics. Robustness and precision are demonstrated using 1000 $h^{-3}Gpc^3$ of realizations, showing the bispectrum's potential to complement two-point statistics in surveys DESI and Euclid.

The second part of the thesis (Chapters 3, 4 and 5) addresses the limitations of machine learning in modeling the baryonic assembly history of galaxies within dark matter halos. Traditional models tend to smooth over short-timescale variability in star formation and chemical enrichment histories, due to their architectural bias toward minimizing global loss functions. We propose a novel correction scheme that decomposes galaxy histories in Fourier space, identifies missing high-frequency power, and re-injects statistically consistent fluctuations into the predicted histories. This framework is applied to galaxies from the

IllustrisTNG simulation. The modified histories restore variability, improving accuracy on observables such as the stellar-halo mass and mass-metallicity relation, spectral energy distributions (SEDs), and photometric color distributions. The corrections particularly enhance the bimodality of galaxy colors, the scatter in metallicity at fixed mass, and the recovery of quenched populations in satellite systems. This framework is further extended to dark matter-only simulations to evaluate the generalizability of neural network-based predictions in the absence of baryonic training data. We assess the role of mass accretion history, halo concentration, and cosmic environment in enabling accurate galaxy property inference, and demonstrate how semi-stochastic corrections improve the fidelity of the generated histories.

These components address distinct but complementary aspects of large-scale structure modeling. The bispectrum-based analysis enhances the precision and robustness of cosmological distance measurements by leveraging higher-order statistics, unlocking additional information from the nonlinear regime of structure formation. In contrast, the semi-stochastic modeling of galaxy formation histories improves the realism of mock galaxy catalogs, correcting biases in data-driven predictions and enriching their utility in forward-modeling frameworks. Additionally, the thesis includes several complementary studies (Chapter 6): an angular multipole analysis of the bispectrum for probing anisotropic clustering, the mitigation of imaging systematics in galaxy survey data, and spectroscopic classification of AGN and QSO targets within DESI. Together, these efforts support the broader goal of connecting theory, simulation, and observation — advancing both the accuracy of cosmological inference and the fidelity of galaxy population modeling for current and future surveys.

Note on Chapter-End Quotations: In keeping with my roots in Odisha, India, each chapter concludes with a short quotation in Odia, my native language. These reflective excerpts are drawn from Odia literature, classical philosophy, and folk tradition, and are intended to symbolically parallel the cosmological and astrophysical ideas explored in this work. Each quote is followed by an English translation to maintain accessibility while honoring the worldview that shaped my earliest engagement with science.

Contents

List of Figures

List of Tables

Acknowledgments

As this chapter of my academic journey comes to a close, I'm deeply aware that this thesis reflects not only my efforts but the insight, encouragement, and support of many individuals. It has been shaped by a community of mentors, colleagues, and loved ones who made this path richer and more meaningful.

First and foremost, I express my deepest appreciation to my advisor, Dr. Lado Samushia, whose mentorship has played a key role in shaping my development as a researcher. His ability to balance structured guidance with the freedom to explore independently allowed me to find my voice in science. His support for conference travel opened doors to the broader scientific community, fostering new perspectives and professional connections that have been invaluable in gaining confidence and clarity as an academic.

I am grateful to Dr. Rita Tojeiro and Dr. Harry Chittenden, whose collaboration was central to the work presented in Chapters 4 and 5. Rita's input guided both the scientific direction and the clarity of the results. I would like to thank Ashley Ross and Stephanie Juneau for expanding my exposure to new areas within astrophysics. Discussions with Ashley deepened my understanding of large-scale structure catalogs, while working with Stephanie opened doors to black hole studies. Each of these collaborations offered new perspectives, technical skills, and areas of excitement that significantly enriched my doctoral experience.

I acknowldge the Physics Department, my peers and colleagues at K-State and in the DESI collaboration for creating an environment of support, curiosity, and shared growth that made this journey truly rewarding.

A heartfelt thank you to my committee members, Dr. Bharat Ratra, Dr. Glenn Horton-Smith, and Dr. David Auckly for their time and thoughtful feedback. A special appreciation to Dr. Ratra, whose efforts to organize seminars with prominent researchers provided timely inspiration and and insight into the forefront of the field.

To Kim Coy, thank you for being the heart of the department. Your care and attention to every student's well-being – remembering birthdays, offering warmth to newcomers, and providing clarity in critical things – made the department feel like a second home.

The late-night board game sessions with Anbu and Thiba during the final year provided the much-needed laughter and relaxation. I thank Sudeep, my long-time friend, for his steady stream of messages and memes. I would like to acknowledge Raga, Hanyu, Vaisakh and Navya, colleagues from the DESI collaboration, for their encouragement and guidance during critical phases of my final year, particularly in DESI-involvement and postdoc applications. Special thanks go to Hanyu, Abhishek, Mehdi and Giorgi for their everyday camaraderie, shared frustrations, and simple kindnesses that made academic life feel less solitary.

To my sisters, Ms. Swarnalata Behera (Didi) and Dr. Bhagyalaxmi Behera (Kunia) – your influence shaped me in more ways than you know. Many of my creative interests trace back to Didi's talents. Kunia's dedication, discipline, and work ethic set a standard I always admired, even if I could never quite match it. Their accomplishments raised the bar and became a quiet source of motivation early on. To my little cheerleaders — my nieces and nephew, Bulbul, Chulbul, and Gugul — thank you for the innocent joy and proud little cheers. My late brother-in-law, Deepak Kumar Behera, was the first to believe in my decision to pursue physics, and his early encouragement gave me the confidence to begin. I also acknowledge Puspendra Sahu for his kind presence and continued support along the way.

Finally, I am deeply grateful to my parents, Madhusudan Behera and Bijayalaxmi Behera, whose unwavering love, sacrifice, and quiet strength are the foundation of every step I've taken and all that I've achieved. And I am equally thankful to my husband, Abhishek Jana, for being a constant source of motivation and encouragement throughout this journey. Their patience, steady support and reassurance, especially during moments of doubt and exhaustion, has made this journey not only possible but also meaningful.

This thesis may carry my name, but it holds the fingerprints of many. Thank you all for being part of this journey - academic, personal, and everything in between.

Dedication

To my beloved parents —

For everything they gave me without ever expecting anything in return. For the strength in my father's silence and the wisdom in my mother's actions. For showing me that love need not be spoken aloud, and that true strength often arrives without noise.

They gave more than they had, and became more than they were asked to be — learning, adapting, and enduring - not for recognition, but for their daughters. In unnoticed hours and unspoken moments, her quiet resolve and his steady effort built the foundation of my achievements.

Though they never named their sacrifices or spoke their lessons aloud, I've tried to carry their lessons here — in thought, in values, and in the quiet spaces of this thesis.

This work is, in every sense, for them.

Chapter 1

Introduction

1.1 Standard Rulers in Cosmology

Understanding the large-scale geometry and expansion history of the Universe requires precise distance measurements across cosmic time. Two essential tools in this effort are *standard candles* and *standard rulers*, which provide complementary approaches to inferring cosmological distances. Standard candles rely on known intrinsic luminosity, while standard rulers depend on known physical size. Together, they allow us to map the expansion history of the Universe and constrain the nature of dark energy.

1.1.1 Distance Measures in Cosmology

In an expanding universe, redshift serves as a proxy for lookback time. However, converting redshift into a physical distance requires knowledge of the expansion rate and the underlying cosmological model. Observables such as flux, angular size, and redshift must be interpreted using either calibrated luminosities (standard candles) or calibrated physical scales (standard rulers). These methods serve as indirect distance estimators crucial for building the cosmological distance-redshift relation.

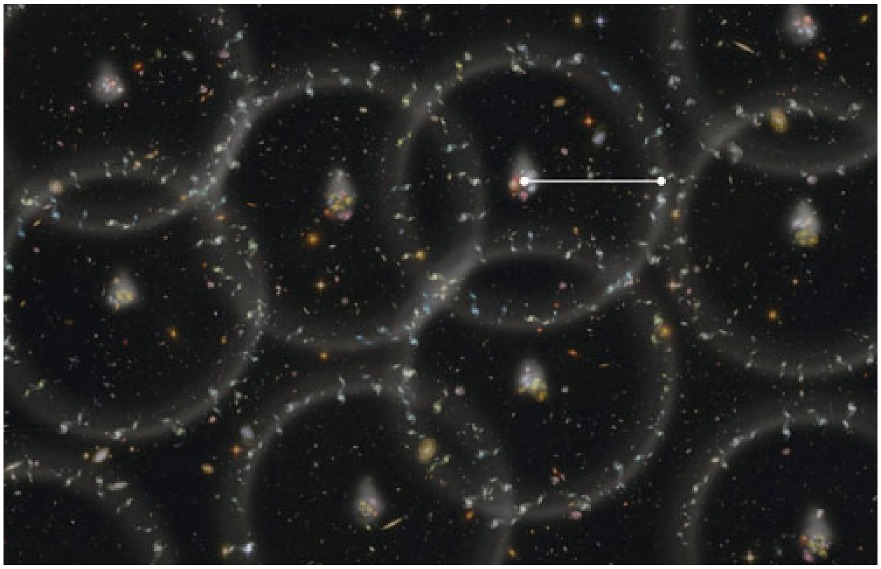

Figure 1.1: Schematic illustration of BAO shells in the large-scale galaxy distribution. Each galaxy tends to have neighbors located at a preferred comoving separation of about 150 Mpc, corresponding to the sound horizon at the drag epoch. Source : BAO Cosmology at UCLA

1.1.2 Standard Candles: Supernovae Type Ia

A *standard candle* is an astrophysical object whose intrinsic luminosity is either known or can be calibrated. By comparing the intrinsic and observed brightness, one can infer its distance via the inverse-square law. The most widely used standard candles in cosmology are *Type Ia supernovae* due to their relatively uniform peak luminosity. These events are believed to arise from thermonuclear explosions of white dwarfs in binary systems. After correcting for empirical correlations with light curve shape and color, SNe Ia can provide precise measurements of luminosity distance. Observations of Type Ia supernovae at different redshifts led to the groundbreaking discovery of the accelerated expansion of the universe, a result now interpreted as evidence for dark energy.

SNe Ia are typically used to probe distances in the redshift range $0 < z \lesssim 2$. They offer high precision at low redshift but are sensitive to astrophysical systematics such as dust extinction, host galaxy effects, and possible evolution of progenitor populations.

1.1.3 Standard Rulers: Baryon Acoustic Oscillations

A standard ruler is a feature of known physical size whose apparent angular or redshift extent can be used to infer cosmological distances. While the most precisely measured standard ruler comes from the acoustic scale imprinted in the Cosmic Microwave Background, Baryon Acoustic Oscillations (BAOs) provide an independent and robust standard ruler at late times, observable in the clustering of galaxies. The BAO scale originates from sound waves that propagated through the photon–baryon plasma in the early universe. These oscillations froze at the time of recombination, leaving an overdensity “shell” at a characteristic comoving scale of ~ 150 Mpc. This scale is preserved in the large-scale clustering of galaxies and can be used to infer cosmic distances.

This scale manifests as a slight excess probability of finding pairs of galaxies separated by that characteristic distance, leaving a detectable imprint in observational statistics.

Unlike supernovae, which require calibration of luminosity, the BAO scale can be computed from first principles using linear perturbation theory and precise measurements of the early Universe (e.g., cosmic microwave background observations). This makes it a robust and absolute standard ruler.

BAO measurements have been used in galaxy surveys such as SDSS, BOSS, eBOSS, and DESI, and have also been detected in the Lyman-α forest at higher redshifts ($z > 2.3$).

1.1.4 Comparison and Complementarity

While both standard rulers and candles serve the same end, measuring cosmological distances, they do so via fundamentally different means and are subject to different systematics. Type Ia supernovae provide high-resolution distance measurements at low redshift and are especially useful for constraining the Hubble constant and late-time cosmic acceleration. However, they require careful calibration and modeling of astrophysical systematics.

BAOs, on the other hand, provide absolute distance measures based on early-universe physics, and are less affected by nonlinear astrophysical processes at late times. Their strength lies in their theoretical robustness and ability to probe a wide range of redshifts with minimal modeling uncertainties.

The two techniques are highly complementary. Supernovae dominate low-redshift cosmology, anchoring the Hubble diagram, while BAOs excel at mapping distances over a wide redshift range and constraining the expansion rate of the universe with sub-percent precision. Together, these techniques form the backbone of modern observational cosmology. Their complementarity allows for cross-validation and joint analyses that significantly tighten constraints on the expansion history and properties of dark energy.

1.2 Cosmological Structure Formation and BAOs

The cosmic structures observed today such as galaxies, clusters, and the vast filamentary network of the cosmic web, have evolved from small initial fluctuations in the early universe. These fluctuations, seeded during inflation, grew over time under the influence of gravity. Embedded within this process is a key feature imprinted by the interaction between baryons and photons in the early universe: the Baryon Acoustic Oscillation (BAO) scale. This section provides an overview of the physics behind structure formation and the origin of the BAO signal.

1.2.1 Gravitational Instability and Linear Perturbations

The initial density field of the universe was nearly homogeneous, with fluctuations at the level of one part in 10^5, as revealed by observations of the Cosmic Microwave Background (CMB). These fluctuations are well described by a Gaussian random field in the linear regime. As the universe expanded and cooled, regions with slightly higher density grew via gravitational instability, attracting more matter and forming the seeds of future structures.

In the linear regime, the evolution of these perturbations can be described analytically. The Fourier modes of the density contrast $\delta(\vec{k})$ grow proportionally to the scale factor during matter domination. The growth rate and shape of the evolving density field are influenced by the cosmological parameters and the relative contributions of matter, radiation, and dark energy.

A key ingredient in this process is the *transfer function*, which describes how initial perturbations at different scales evolve differently due to the interplay of radiation pressure, baryon inertia, and dark matter. At small scales, the growth of perturbations is suppressed by radiation pressure in the early universe, while larger scales are largely unaffected. This differential evolution shapes the matter power spectrum and contributes to the emergence of features like the BAO.

1.2.2 Formation of the BAO Scale

Before recombination, baryons and photons were tightly coupled due to Thomson scattering. In this photon-baryon fluid, overdense regions created pressure gradients, which in turn generated spherical sound waves that propagated outward from the initial perturbation. These acoustic oscillations traveled at the sound speed in the fluid, which is itself a function of the baryon and photon densities.

As the universe cooled and expanded, it reached the epoch of recombination, around $z \sim 1100$, when electrons and protons combined to form neutral hydrogen. At this point, photons decoupled from baryons and began to free-stream, forming the CMB, while the baryons effectively "froze" in place. The radius these sound waves had traveled before decoupling sets the characteristic scale of the BAO, known as the *sound horizon at the drag epoch*, ~ 150 Mpc in comoving coordinates.

This scale becomes imprinted in the matter distribution as a slight preference for pairs of galaxies to be separated by this distance. It appears as a peak in the two-point correlation function and as oscillations in the power spectrum. Because the physical processes determining this scale are well understood and rooted in early-universe physics, the BAO provides a robust standard ruler for measuring cosmological distances.

1.2.3 Observational Tracers of Large-Scale Structure

To observe the BAO feature, we rely on large-scale maps of the matter distribution. However, we cannot directly observe dark matter, which makes up the majority of the mass in the universe. Instead, we use galaxies and other luminous tracers as biased proxies for the underlying matter field.

Galaxies form within dark matter halos, and their spatial distribution reflects, though not perfectly, the distribution of the halos themselves. The degree to which galaxies trace the matter field is quantified by the galaxy bias, which can be scale-dependent and evolve with redshift and galaxy type.

In addition to galaxies, other tracers like the *Lyman-α* forest and quasars are used to probe the matter field, especially at higher redshifts ($z > 2$). The Lyman-α forest consists of absorption features in the spectra of distant quasars, caused by intervening neutral hydrogen in the intergalactic medium. These features trace the large-scale density field and allow for BAO measurements at epochs inaccessible to traditional galaxy surveys.

Together, these tracers enable the detection of the BAO scale across a wide range of redshifts, providing critical constraints on the expansion history of the universe and the nature of dark energy.

1.3 Observational Statistics in Large-Scale Structure

The universe on large scales is statistically homogeneous and isotropic, and the distribution of matter is best described as a random field. Since we cannot predict the precise location of each galaxy, the large-scale distribution of matter in the universe is described statistically, rather than deterministically, due to the stochastic nature of initial conditions and the nonlinear evolution of density fluctuations. Observational statistics provide a formalism for quantifying this structure using measurable quantities such as galaxy positions and redshifts. This section introduces the core statistical tools used in cosmology: the two-point and three-point correlation functions and their Fourier-space analogues, the power spectrum and bispectrum.

1.3.1 The Need for Statistical Measures

Galaxy surveys measure the spatial positions of millions of galaxies. These positions are affected by initial conditions, gravitational evolution, and biasing mechanisms. Rather than modeling each object individually, cosmologists employ statistical measures that describe the ensemble properties of the matter field. The underlying assumption is that the universe is a realization of a statistically defined field governed by a small number of physical parameters.

Statistical homogeneity and isotropy imply that all information about the matter distribution is encoded in correlation functions of increasing order. These functions measure the likelihood of finding galaxies at particular separations or configurations, and they are the primary observables for extracting cosmological information from large-scale structure.

1.3.2 Two-Point Correlation Function and Power Spectrum

The two-point correlation function (2PCF), $\xi(r)$, measures the excess probability of finding a pair of galaxies separated by a distance r, relative to a random (unclustered) distribution. It is formally defined as:

$$dP = \bar{n}^2[1 + \xi(r)]dV_1 dV_2, \tag{1.1}$$

where $\bar{n}$ is the mean number density of galaxies and dV_1, dV_2 are volume elements. A positive $\xi(r)$ indicates clustering at scale r, while $\xi(r) = 0$ corresponds to a random distribution.

In a statistically isotropic universe, the correlation function depends only on the separation $r = |\vec{r}|$, so we write $\xi(r)$. This function encodes the strength of clustering at various scales. On large scales ($r \gtrsim 10$ Mpc), it probes the linear regime of structure formation, where theoretical models are most reliable. The BAO feature appears as a broad peak in $\xi(r)$ at around 150 Mpc, corresponding to the sound horizon.

In practice, the 2PCF is estimated from galaxy catalogs using pair counts. Several estimators have been proposed to compute the 2PCF, aimed at minimizing biases and maximizing the signal-to-noise ratio. The most commonly used estimators include the Peebles and Hauser[8], Davis and Peebles[9], Hamilton[10], and Landy and Szalay[11] estimators. A widely used estimator that minimizes variance and edge effects is the Landy-Szalay estimator.

$$\xi(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)}, \tag{1.2}$$

where:

- $DD(r)$: number of galaxy–galaxy (data–data) pairs at separation r.
- $RR(r)$: number of random–random pairs.
- $DR(r)$: number of data–random pairs.

The estimator requires a synthetic random catalog that matches the survey geometry and selection function.

Multipole Expansion and Redshift-Space Distortions

In redshift space, peculiar velocities distort the observed positions of galaxies along the line of sight, introducing anisotropies in the clustering signal. To capture this, the 2PCF is decomposed using a multipole expansion:

$$\xi(s, \mu) = \sum_{\ell} \xi_\ell(s) P_\ell(\mu), \tag{1.3}$$

where:

- s: redshift-space separation
- μ: cosine of the angle between the separation vector and the line of sight
- $P_\ell(\mu)$: Legendre polynomial of order ℓ

The multipole moments $\xi_\ell(s)$ are computed via:

$$\xi_\ell(s) = \frac{2\ell + 1}{2} \int_{-1}^{1} \xi(s, \mu) P_\ell(\mu)\, d\mu. \tag{1.4}$$

The monopole ($\ell = 0$) captures angle-averaged clustering, while the quadrupole ($\ell = 2$) and hexadecapole ($\ell = 4$) quantify redshift-space distortions such as the Kaiser effect[12] and small-scale velocity dispersions[13].

Projected Correlation Function

The redshift-space 2PCF can also be expressed in terms of separations perpendicular (r_p) and parallel (π) to the line of sight, i.e., $\xi(r_p, \pi)$. The projected correlation function $w_p(r_p)$ is then defined by integrating $\xi(r_p, \pi)$ along π:

$$w_p(r_p) = 2 \int_0^{\pi_{\max}} \xi(r_p, \pi)\, d\pi. \tag{1.5}$$

This projection smooths over redshift-space distortions and yields a measurement of the real-space clustering signal[9].

Power Spectrum

The power spectrum $P(\vec{k})$ is the Fourier transform of the two-point correlation function:

$$P(\vec{k}) = \int \xi(\vec{r}) e^{i\vec{k}\cdot\vec{r}} d^3 r. \tag{1.6}$$

Likewise, the inverse relation is:

$$\xi(\vec{r}) = \frac{1}{(2\pi)^3} \int P(\vec{k}) e^{-i\vec{k}\cdot\vec{r}} d^3 k. \tag{1.7}$$

For isotropic fields, both $\xi(r)$ and $P(k)$ depend only on the magnitude of their arguments. The power spectrum describes how variance in the density field is distributed across different spatial scales. Peaks and wiggles in $P(k)$ correspond to preferred scales in the matter distribution, most notably, the BAO feature, which appears as a series of damped oscillations in

$P(k)$.

The power spectrum is particularly advantageous because it connects naturally to theoretical models, such as those derived from linear perturbation theory and numerical simulations.

1.3.3 Three-Point Correlation Function and the Bispectrum

The three-point correlation function (3PCF), $\zeta(r_{12}, r_{23}, r_{31})$, generalizes the 2PCF to triplets of galaxies. It quantifies the excess probability over random of finding three galaxies forming a triangle with sides r_{12}, r_{23}, and r_{31}:

$$\zeta = \frac{DDD}{RRR} - 1, \tag{1.8}$$

where DDD and RRR refer to triplet counts in the data and random catalogs, respectively. The 3PCF contains additional information about the shape of galaxy clustering and is sensitive to gravitational nonlinearities, bias, and primordial non-Gaussianity.

The Fourier counterpart of the 3PCF is the bispectrum $B(\vec{k}_1, \vec{k}_2)$, which is defined through the expectation value of the Fourier modes of the density field:

$$\langle \delta(\vec{k}_1)\delta(\vec{k}_2)\delta(\vec{k}_3) \rangle = (2\pi)^3 \delta_D(\vec{k}_1 + \vec{k}_2 + \vec{k}_3) B(\vec{k}_1, \vec{k}_2), \tag{1.9}$$

where δ_D is the Dirac delta function enforcing momentum conservation ($\vec{k}_1 + \vec{k}_2 + \vec{k}_3 = 0$). The three wavevectors $\vec{k}_1, \vec{k}_2, \vec{k}_3$ form a closed triangle in Fourier space, and the bispectrum depends on both the scale and the shape of this triangle.

The bispectrum is sensitive to the shape of structures and the non-Gaussian features of the density field. It provides information not present in the power spectrum, such as nonlinear gravitational evolution, scale-dependent bias, and primordial non-Gaussianity[14;15]. While more complex to measure and model, it offers a rich avenue for improving cosmological constraints. More information on bispectrum is in Section 1.5.

1.3.4 Alternative and Complementary Statistics

While the two-point and three-point correlation functions are the standard tools for characterizing large-scale structure, several alternative statistics have been developed to probe specific aspects of galaxy formation and the galaxy–halo connection. These observables offer complementary constraints to traditional correlation functions and are often employed in Halo Occupation Distribution (HOD) modeling and galaxy–halo relation studies.

1.3.5 Physical Interpretation and Complementarity

Each statistical measure probes different aspects of structure formation. The power spectrum captures the variance of fluctuations, while the bispectrum and higher-order statistics measure phase correlations, essential for understanding the detailed shape and evolution of structures. The 2PCF and power spectrum are sensitive to the amplitude and scale of fluctuations, while higher-order statistics such as the bispectrum and 3PCF reveal mode coupling and non-Gaussian structure.

Importantly, these statistics are complementary. For example, combining the power spectrum and bispectrum can break degeneracies between parameters like bias, growth rate, and the amplitude of fluctuations. They also respond differently to systematic effects, making cross-validation possible.

1.3.6 Challenges in Measuring Clustering Statistics

Despite their power, these statistics come with practical challenges. Survey geometry, selection effects, and redshift-space distortions complicate the interpretation of measurements. Shot noise from discrete tracers adds additional variance, particularly on small scales.

In Fourier space, the finite volume of the survey limits the smallest measurable wavenumber, while binning in k-space and triangle configuration space for the bispectrum adds complexity to analysis pipelines. Estimating accurate covariance matrices for these statistics is also non-trivial and often relies on mock catalogs or analytical approximations.

Nevertheless, two- and three-point statistics remain foundational tools for extracting cosmological information from the large-scale structure of the universe.

1.3.7 Alternative Cosmological Clustering Statistics

Beyond the traditional correlation functions and bispectrum, several alternative statistical approaches have been developed to probe cosmological information encoded in the large-scale structure. These statistics are particularly useful for testing models of gravity, dark energy, and non-Gaussian initial conditions, and are complementary to standard analyses.

Void Statistics

Cosmic voids are underdense regions in the matter distribution and offer a unique window into cosmological structure formation. Because voids are less affected by nonlinear clustering and galaxy bias, they serve as clean probes of cosmological parameters such as the matter density Ω_m, dark energy equation of state, and modified gravity parameters.

Void statistics include the void size function, radial density profiles, and clustering of void centers. These observables have been used to place competitive constraints on dark energy and deviations from General Relativity[16;17].

Marked Correlation Functions

Marked statistics apply weights ("marks") to galaxies based on local density, luminosity, or other physical properties, and modify traditional correlation functions accordingly. The marked two-point function enhances sensitivity to environmental dependencies in clustering and is especially powerful for constraining galaxy bias and the growth rate of structure[18;19].

Counts-in-Cells and Density PDFs

Instead of focusing on pairwise correlations, counts-in-cells measure the probability distribution of galaxy counts within spherical or cubic volumes of fixed size. This approach captures the full distribution of density fluctuations, going beyond the mean and variance to probe non-Gaussian features of the field. It is sensitive to the amplitude of matter fluctuations, halo bias, and primordial non-Gaussianity[20].

Topological Measures

Topological statistics such as Minkowski Functionals and Betti numbers quantify the shape, connectivity, and void-filling structure of the cosmic web. These methods are sensitive to the geometry of the density field and are robust to systematic effects like shot noise or selection function variations. They have been used to probe non-Gaussianity, dark energy, and cosmic isotropy[21–23].

1.4 BAO Features in Two-Point Statistics

The Baryon Acoustic Oscillation (BAO) scale is one of the most prominent features in the large-scale distribution of matter, and its signature is most clearly observed in two-point statistics. This section reviews how the BAO manifests in both configuration and Fourier space, the methodology for isolating the signal, and the reconstruction techniques used to sharpen it. These two-point statistics form the basis for most current cosmological constraints using BAO.

1.4.1 BAO in the Power Spectrum and Correlation Function

The BAO feature arises from the sound horizon imprinted in the matter distribution, as discussed in Section 1.2.2. In configuration space, this appears as a small but broad peak in the two-point correlation function $\xi(r)$ at a separation of ~ 150 Mpc, corresponding to the comoving scale of the sound horizonas in Figure 1.2.

In Fourier space, the same physical process leads to a series of oscillations superimposed on the matter power spectrum $P(k)$ as in Figure 1.2. These oscillations are often referred to as "wiggles" and are damped at small scales due to nonlinear effects. The power spectrum is typically decomposed into a "smooth" broadband component and a "wiggly" oscillatory

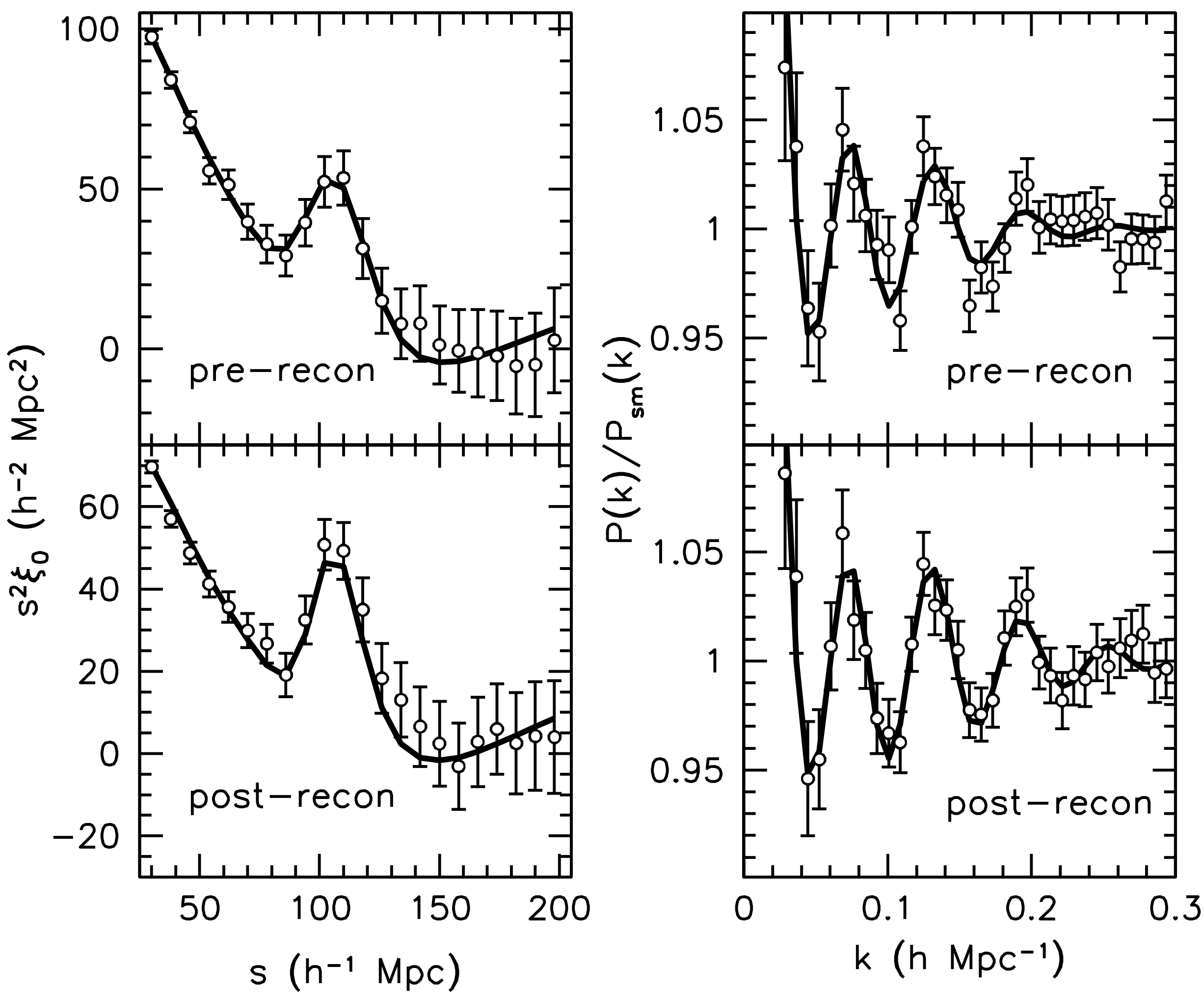

Figure 1.2: Measurements of two-point clustering statistics ($\xi(s)$ in the left panel and $P(k)$ in the right panel) from the CMASS sample from BOSS Data Release 11. In order to highlight the signals in each case, it is common to multiply the correlation function by s^2 and divide the power spectrum by a smooth power spectrum without oscillations ($P_{\mathrm{sm}}(k)$). Top row shows the measurements before reconstruction while the bottom row shows the results after the reconstruction. It is noticeable that the reconstruction sharpens the BAO signal. *Source:* Anderson et al.[1]

component:

$$P(k) = P_{\mathrm{smooth}}(k) \times P_{\mathrm{wiggle}}(k). \tag{1.10}$$

This decomposition is not unique, but common methods use spline fits or theoretical models from Boltzmann solvers like CAMB[24] or CLASS[25] to define the smooth part. The oscillatory part $P_{\mathrm{wiggle}}(k)$ contains the key cosmological information related to distances.

The exact shape and phase of the wiggles are sensitive to the sound horizon at the drag epoch, and hence to cosmological parameters like the matter density (Ω_m), the Hubble parameter (h), and the baryon fraction (Ω_b/Ω_m). BAO analyses thus focus on measuring these oscillations precisely to constrain the expansion history.

1.4.2 Extracting the BAO Signal

To isolate the BAO feature, most analyses compare the observed power spectrum or correlation function to a theoretical template that models the BAO oscillations. The observed signal is rescaled relative to the template using a dilation parameter α, which encodes deviations in the measured BAO scale from that predicted by a fiducial cosmology.

$$P_{\mathrm{obs}}(k) \approx P_{\mathrm{model}}(\alpha k), \tag{1.11}$$

In the isotropic approximation, where angular and radial distances are assumed to scale uniformly, the BAO scale is constrained through a spherically averaged combination of angular diameter distance $D_A(z)$ and Hubble parameter $H(z)$. This is encapsulated in the volume-averaged distance $D_V(z)$, defined as:

$$D_V(z) = \left[(1+z)^2 D_A^2(z) \frac{cz}{H(z)}\right]^{1/3}. \tag{1.12}$$

The isotropic dilation parameter is then:

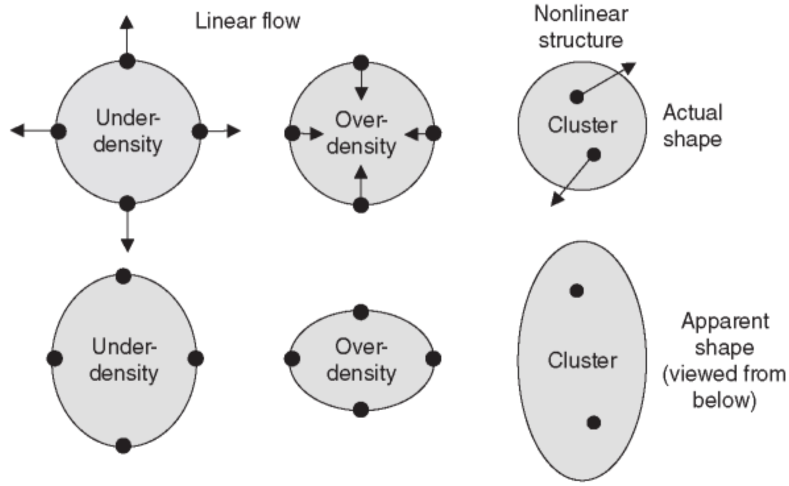

Figure 1.3: Explanatory diagram showing how real-space structures (top row) look in redshift-space (bottom row).

$$\alpha = \frac{D_V(z)\, r_{d,\mathrm{fid}}}{D_V^{\mathrm{fid}}(z)\, r_d}, \tag{1.13}$$

where r_d is the comoving sound horizon at the drag epoch, and the superscript "fid" denotes the value assumed in the fiducial cosmology. A value of $\alpha \neq 1$ indicates that the observed acoustic scale differs from the fiducial expectation, providing a constraint on cosmological parameters that affect distances and expansion history[26].

In practice, redshift-space distortions and line-of-sight effects break isotropy[2;12;27;28] (Figure. 1.3, 1.4). To account for this, modern analyses model the anisotropic BAO signal using separate dilation parameters in the transverse and radial directions:

$$\alpha_\perp = \frac{D_A(z)\, r_{d,\mathrm{fid}}}{D_A^{\mathrm{fid}}(z)\, r_d}, \quad \alpha_\parallel = \frac{H^{\mathrm{fid}}(z)\, r_{d,\mathrm{fid}}}{H(z)\, r_d}. \tag{1.14}$$

From these, one can define the average isotropic dilation and the anisotropy parameter:

$$\alpha = \alpha_\perp^{2/3} \alpha_\parallel^{1/3}, \quad \epsilon = \left(\frac{\alpha_\parallel}{\alpha_\perp}\right)^{1/3} - 1. \tag{1.15}$$

These parameters are constrained by jointly fitting the monopole and quadrupole moments of the two-point correlation function or power spectrum. This allows cosmological

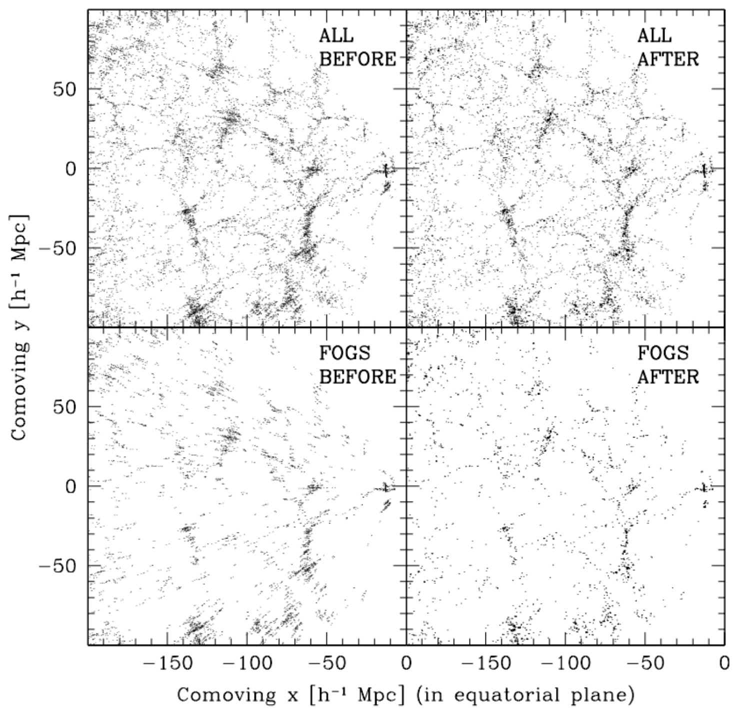

Figure 1.4: An illustration of the "Fingers of God" (FoG) effect — the elongation of virialized structures along the line of sight, adapted from Tegmark et al. [2]. The figure displays galaxies from a slice of the SDSS dataset, projected through the declination axis in two-dimensional comoving space. The top row shows all 67,626 galaxies in the slice, while the bottom row highlights those identified as exhibiting the FoG effect. The right column presents the same galaxies after modeling and removing the FoG distortions. The observer is located at the origin (x,y)=(0,0), and the radial smearing caused by line-of-sight velocities is clearly visible in the lower left panel as elongated structures pointing toward the observer.

constraints on the angular diameter distance and Hubble expansion rate at a given redshift.

To account for nonlinear effects, the BAO template is also damped by a Gaussian smoothing factor:

$$D(k) = \exp(-k^2\Sigma^2), \tag{1.16}$$

where Σ is a nuisance parameter that quantifies the scale of nonlinear BAO suppression. This damping is crucial for accurately modeling the reduced amplitude of BAO oscillations on small scales (Figure 1.2).

These methods are now standard in BAO analyses across modern galaxy surveys such as the Baryon Oscillation Spectroscopic Survey (BOSS), the extended BOSS (eBOSS)[29;30], both of which were part of the Sloan Digital Sky Survey (SDSS)[31], and the Dark Energy Spectroscopic Instrument DESI[32].

Fitting procedures are often performed using both the monopole and quadrupole moments of the correlation function or power spectrum, allowing for anisotropic dilation parameters that separately probe transverse and radial distances.

1.4.3 BAO Reconstruction Techniques

Over time, nonlinear gravitational evolution distorts and damps the BAO feature, reducing its contrast in observed statistics as in the top row of Figure 1.2. To mitigate this effect, a technique known as *reconstruction* was developed by Eisenstein et al.[33]. The method attempts to reverse the bulk flows that blur the BAO signal by estimating the displacement field from the observed galaxy distribution and moving galaxies back toward their initial positions.

Reconstruction has become standard in BAO analyses, as it restores the sharpness of the peak in $\xi(r)$ and enhances the amplitude of oscillations in $P(k)$, significantly improving the precision of distance measurements. The effect of reconstruction is particularly pronounced at low redshift, where nonlinearities are stronger.

While powerful, reconstruction assumes that the displacement field can be reliably estimated from the observed galaxy field. This assumption is valid in the linear or quasi-linear regime but may break down on small scales or in highly nonlinear environments. Furthermore, reconstruction partially removes information from higher-order statistics and transfers it to the two-point function, which has implications for analyses using the bispectrum, as discussed in later sections.

Reconstruction methods continue to evolve, with ongoing work aiming to improve the modeling of bias, redshift-space distortions, and shot noise during the reconstruction process[34–36].

1.5 The Bispectrum

While the power spectrum captures the variance in the density field and is sufficient to describe Gaussian initial conditions, gravitational evolution introduces non-Gaussian features that cannot be fully captured by two-point statistics alone. The bispectrum, the Fourier-space analogue of the three-point correlation function, is the lowest-order statistic sensitive to phase correlations and mode coupling. It provides access to new cosmological information, including details about structure formation, bias, and even early-universe physics.

1.5.1 Definition and Physical Significance

The bispectrum $B(\vec{k}_1, \vec{k}_2)$ is defined as the expectation value of the product of three Fourier modes of the overdensity field $\delta(\vec{k})$ as in Equation 1.9.

The bispectrum is nonzero only if the underlying field is non-Gaussian. Even if the initial conditions are Gaussian, non-linear gravitational evolution and galaxy bias generate non-zero bispectrum signals at late times. Importantly, the bispectrum is sensitive to the relative phases of Fourier modes, capturing how large- and small-scale modes interact, information inaccessible to the power spectrum alone.

Because of its shape dependence, the bispectrum can distinguish between different physical effects, such as local vs. equilateral primordial non-Gaussianity[37;38] or scale-dependent galaxy bias[39;40]. It also enhances constraints on parameters like the growth rate of structure and σ_8, particularly when combined with the power spectrum.

1.5.2 Measurement Challenges

Despite its rich information content, the bispectrum is significantly more difficult to measure and model than the power spectrum. One key complication is the high dimensionality of triangle configurations, while the power spectrum depends only on the scalar k, the bispectrum depends on the shape and size of triangle triplets.

To reduce computational complexity, bispectrum measurements are often binned using triangle types (e.g., equilateral, isosceles, squeezed) or compressed into summary statistics such as multipoles. Fast algorithms for bispectrum estimation from galaxy surveys have been developed, including methods based on FFTs and grid-based binning[41;42].

Survey geometry and selection functions introduce mode coupling and leakage between configurations, which must be accounted for when comparing to theory. Redshift-space distortions further complicate interpretation by introducing angular dependencies that break isotropy.

Finally, accurate estimation of the covariance matrix for the bispectrum is a major challenge. Unlike the power spectrum, which is nearly Gaussian on large scales, the bispectrum is inherently non-Gaussian and requires a large number of mock catalogs or analytical models for its full covariance.

1.5.3 Status of Bispectrum Modeling

Theoretical modeling of the bispectrum has traditionally relied on leading-order perturbation theory, often referred to as the "tree-level" approximation. This provides an analytic expression for the bispectrum in the mildly nonlinear regime, assuming a bias expansion for galaxies:

$$B_g(\vec{k}_1, \vec{k}_2, \vec{k}_3) = b_1^3 B_m(\vec{k}_1, \vec{k}_2, \vec{k}_3) + \text{bias and RSD terms,} \tag{1.17}$$

where B_m is the matter bispectrum, and b_1 is the linear bias parameter. This expression can be extended using perturbation theory kernels F_2 and G_2, which encode second-order gravitational effects and velocity couplings[43;44].

However, tree-level predictions are only accurate on very large scales ($k \lesssim 0.1\ h/\mathrm{Mpc}$), and fail to capture the full bispectrum shape at quasi-linear and nonlinear scales. To address this, more sophisticated approaches have been developed:

- **Effective Field Theory (EFT)** models include counterterms that absorb small-scale physics, improving predictions up to $k \sim 0.3\ h/\mathrm{Mpc}$[45;46].
- **Simulation-based emulators** interpolate between high-fidelity N-body simulation results over a grid of cosmological parameters[47].
- **Halo model extensions** incorporate halo biasing and profile contributions to describe small-scale bispectrum contributions.
- **Measurement frameworks**, such as the `Triumvirate` package[48], provide efficient and scalable tools for computing bispectrum multipoles and three-point correlation functions using FFT-based estimators. These tools are essential for validating models against simulations and mocks in modern large-volume surveys.

These models are continually being refined, but achieving percent-level accuracy across all triangle shapes and scales remains challenging, especially for cosmological inference where model error must be controlled. As a result, most bispectrum analyses currently use conservative scale cuts and calibrate their models against large ensembles of mock catalogs.

1.6 BAO Features in the Bispectrum

The bispectrum, as a higher-order statistic, contains additional information about the distribution of matter beyond what is captured in the power spectrum. In particular, the Baryon Acoustic Oscillation (BAO) feature also manifests in the bispectrum, where it provides complementary constraints on cosmic distances and the growth of structure. This section outlines the motivation for BAO detection in the bispectrum, summarizes previous work in this area, and highlights the differences between full-shape and wiggle-only modeling approaches.

1.6.1 Motivation for BAO Detection in the Bispectrum

The BAO signal in the bispectrum arises due to interference between different modes in the density field, many of which are affected by the same underlying acoustic oscillations. Because the bispectrum is sensitive to phase correlations between modes, it enhances the detectability of subtle features like the BAO, especially in triangle configurations where the acoustic signal is constructively amplified[15;41;42].

In comparison to the power spectrum, the bispectrum provides sensitivity to different combinations of the cosmological distance scale. It can probe anisotropic effects in both transverse and radial directions, improving constraints on the Hubble parameter $H(z)$ and angular diameter distance $D_A(z)$. Additionally, the bispectrum helps break degeneracies between bias, growth rate, and amplitude parameters like σ_8 when analyzed jointly with the power spectrum[39;49].

Importantly, the bispectrum retains valuable cosmological information even when reconstruction has been applied to the density field. Since reconstruction suppresses nonlinearities and sharpens the BAO peak in the power spectrum, it may also redistribute information into higher-order correlations. Some of this displaced information can be recovered by jointly analyzing the post-reconstruction bispectrum and power spectrum[50].

1.6.2 Previous Work

Initial studies of the BAO feature in the bispectrum used perturbation theory to model its shape and position. For example, Sefusatti et al.[41] computed the tree-level matter bispectrum and showed that BAO appears as a series of small-amplitude oscillations, most prominent in certain squeezed and equilateral triangle configurations.

Subsequent work extended this framework to biased tracers and redshift space, incorporating galaxy bias and velocity distortions. The BOSS and eBOSS surveys have published bispectrum measurements alongside power spectrum results, using multipole decomposition to isolate angular dependence. For instance, Gil-Marín et al.[39] performed a joint analysis of the bispectrum monopole and power spectrum multipoles to constrain cosmological parameters from BOSS DR12.

More recently, models based on Effective Field Theory (EFT) have enabled fits to the bispectrum on quasi-linear scales up to $k \sim 0.25\ h/\mathrm{Mpc}$. These efforts show that the bispectrum improves constraints on $f\sigma_8$, $\alpha_\perp$, and $\alpha_\parallel$, though the modeling remains sensitive to systematics such as bias parametrization and scale cuts[45;46;49].

While these full-shape fits incorporate the entire bispectrum signal, they are computationally expensive and prone to model misspecification. The BAO oscillations are a subdominant contribution to the full bispectrum amplitude, making it challenging to isolate them without bias from the broadband shape.

1.6.3 Wiggle-Only vs. Full-Shape Approaches

One strategy to mitigate modeling uncertainties is to focus exclusively on the oscillatory component of the bispectrum, the so-called "wiggle-only" approach. This method isolates the BAO signature by constructing a bispectrum model where the smooth broadband shape has been factored out or subtracted. This can be done, for example, by taking the ratio of bispectra from simulations with and without the BAO feature in the initial power spectrum. Wiggle-only models offer several advantages:

- They reduce dependence on nonlinear bias models and EFT counterterms.
- They are less sensitive to scale-dependent systematics and broadband modeling errors.
- They allow for more aggressive inclusion of higher k-modes, extending the usable range of the bispectrum for BAO extraction.

However, the trade-off is that they discard non-oscillatory information that could help constrain other cosmological parameters. Nevertheless, for the specific task of measuring the

BAO scale, especially in the context of combining multiple observables, wiggle-only modeling provides a robust and efficient approach.

This methodology forms the core of the work presented in Chapter 2, where we construct a simple flexible template for the BAO signal in the bispectrum and validate its accuracy using simulations with and without acoustic oscillations.

1.7 Simulation-Based Approaches for BAO Studies

Analytical models of large-scale structure are inherently limited by nonlinear effects, redshift-space distortions, galaxy bias, and other complexities. To accurately model and validate BAO signals, particularly in the bispectrum, cosmologists rely on large ensembles of numerical simulations. These simulations provide a way to generate realistic matter and galaxy distributions, test theoretical templates, and estimate covariance matrices for statistical inference.

1.7.1 Role of N-body Simulations

N-body simulations numerically evolve a set of particles under gravity, capturing the nonlinear clustering of dark matter from initial conditions to late times. These simulations serve as the primary testbed for understanding the evolution of the BAO feature across different statistical measures. Simulations can be used to:

- Validate analytical or perturbation theory models for power spectrum and bispectrum.
- Quantify the impact of nonlinear effects, redshift-space distortions, and scale-dependent bias.
- Estimate covariance matrices by generating large ensembles of mock realizations.

In particular, simulations allow precise measurement of the “true” BAO scale by comparing outputs with and without acoustic features in the initial power spectrum. This technique is central to wiggle-only modeling and other BAO isolation strategies.

1.7.2 GLAM and PMILL Simulations

Two key simulation suites relevant for BAO studies are the GLAM (Gadget-based Lagrangian Approximation Method) and PMILL (P-Millennium) simulations.

GLAM simulations provide high-volume, fast approximate realizations of large-scale structure, with sufficient resolution to capture BAO features in the power spectrum and bispectrum. Because GLAM is computationally efficient, it can generate hundreds of mock catalogs for covariance estimation and template testing[51].

PMILL simulations, on the other hand, are full N-body simulations with higher mass and spatial resolution. They accurately model small-scale clustering and provide a reference benchmark for nonlinear effects. PMILL is especially useful for cross-validating approximate methods like GLAM or for calibrating bias parameters in halo models[52].

In many applications, the initial power spectrum is modified to create two versions of the simulation: one with BAO "wiggles" ($P_{\text{wiggle}}(k)$) and one where the oscillations have been smoothed out ($P_{\text{no-wiggle}}(k)$). The resulting fields can be directly compared to isolate the acoustic contribution in any statistic.

1.7.3 Validating Theoretical Models Against Simulations

The BAO templates used in data analysis are typically validated by fitting to measurements from simulations. This ensures that the theoretical model accurately recovers the BAO scale under realistic conditions, including nonlinear evolution and shot noise.

A common validation approach involves:

- Measuring the power spectrum or bispectrum from simulations with and without BAO
- Fitting a template to recover the scale dilation parameter α
- Comparing recovered α to the known input scale

Agreement at the sub-percent level is generally required for the template to be considered robust. Discrepancies can highlight issues such as inadequate damping models, template mismatch, or breakdown of the perturbative regime.

The accuracy of validation depends critically on simulation volume. For instance, a single simulation may suffer from sample variance, while a large ensemble, such as $N = 100–1000$ realizations, can average over cosmic variance and produce stable estimates of the mean and uncertainty.

Modern analyses increasingly rely on simulation-based emulators and hybrid models that interpolate between exact N-body results and theoretical expectations. These tools are essential for achieving the precision required by next-generation surveys like DESI, Euclid, and Roman.

1.8 Summary and Motivation for This Work

This chapter reviews the theoretical foundations, statistical tools, and modeling techniques relevant to the study of Baryon Acoustic Oscillations (BAOs) in large-scale structure. The discussion begins with an overview of standard candles and rulers as cosmological distance probes, highlighting the robustness and broad applicability of the BAO scale. The statistical measures used to extract this signal are then outlined, ranging from two-point functions such as the power spectrum to higher-order statistics like the bispectrum.

The bispectrum, in particular, offers access to additional cosmological information due to its sensitivity to phase correlations and nonlinear evolution. However, modeling the full bispectrum remains challenging, especially on quasi-linear and nonlinear scales. While recent work has incorporated the bispectrum into joint analyses using Effective Field Theory and halo models, these methods are computationally demanding and subject to modeling uncertainties.

To address this, simulation-based approaches have emerged as a promising avenue. In particular, the use of simulations with and without acoustic oscillations in the initial conditions allows for clean separation of the BAO feature from the broadband bispectrum shape. This forms the basis of *wiggle-only* modeling strategies, which aim to isolate and characterize the acoustic signal directly, bypassing the need for full-shape modeling and extensive bias parametrizations.

The work presented in the following chapter builds on these developments by introducing a simple and flexible template for the BAO feature in the bispectrum, derived from large-volume GLAM simulations with and without BAO. This template is tested across a wide variety of triangle configurations and redshifts, demonstrating that the BAO scale can be recovered with sub-percent accuracy, even in the presence of model mismatch. The methodology focuses exclusively on the oscillatory component, offering a robust path forward for future analyses of the bispectrum in large galaxy surveys.

Chapter 2 provides a detailed presentation of this approach, including the simulation setup, the bispectrum measurement pipeline, and validation results across different scales, redshifts, and configurations.

“ସୃଷ୍ଟି ଆରମ୍ଭ ହେଉଛି ଏକ ଧ୍ୱନି ରୂପରେ — ଯାହା ନୀତି ଓ ନିୟମକୁ ଉନ୍ମୋଚନ କରେ।”

— Inspired by the Brihadaranyaka Upanishad

“Creation begins as a sound — revealing law and order.”

Chapter 2

Modeling the BAO feature in Bispectrum.

This chapter originally appeared in the literature as :

Jayashree Behera, Mehdi Rezaie, Lado Samushia and Julia Ereza, Modeling the BAO feature in Bispectrum, arXiv:2312.05942[53].

Abstract

We investigate how well a simple leading order perturbation theory model of the bispectrum can fit the BAO feature in the measured bispectrum monopole of galaxies. Previous works showed that perturbative models of the galaxy bispectrum start failing at the wavenumbers of $k \sim 0.1h\ \mathrm{Mpc}^{-1}$. We show that when the BAO feature in the bispectrum is separated it can be successfully modeled up to much higher wavenumbers. We validate our modeling on `GLAM` simulations that were run with and without the BAO feature in the initial conditions. We also quantify the amount of systematic error due to BAO template being offset from the true cosmology. We find that the systematic errors do not exceed 0.3 per cent for reasonable deviations of up to 3 per cent from the true value of the sound horizon.

2.1 Introduction

The Baryon Acoustic Oscillation (BAO) signal is a statistical feature imprinted into the spatial distribution of galaxies. Small overdensities in the initial distribution of dark matter serve as seeds for acoustic wave propagation in the early baryon-photon fluid. The propagation of these waves freezes at later times and results in a slightly overdense spherical shell of a radius of about $r_{\rm d} \sim 100h^{-1}$ Mpc around each primordial overdensity. From the observational point of view, this means that every galaxy created around the initial overdensity has a slightly higher chance of having a neighboring galaxy on this spherical shell. Later this feature manifests itself as a local maximum (peak) at $r_{\rm d}$ in the two-point correlation function of galaxies and as a decaying oscillatory feature of frequency $2\pi/r_{\rm d}$ in the power spectrum. The BAO scale has been measured from the two-point correlation function with a sub-percent precision. Theoretical modeling and interpretation of the measured BAO scale in the two-point correlation function are relatively straightforward. The measurements in conjunction with the theory can be used to determine the distance-redshift relationship and therefore to constrain cosmological models. The same feature must also be present in other measures of galaxy clustering but it may not be as easy to model. The main goal of our paper is to explore the possibility of modeling the BAO feature in the Fourier space measurements of the galaxy three-point correlation function.

Many methods for measuring the BAO peak position (or frequency) have been proposed in recent years. They all follow the same schematic idea proposed in[33].

- Theoretically computed two-point correlation function (or the power spectrum) is split into "smooth" and "wiggly" components, $\xi = \xi_{\rm s}\xi_{\rm w}$ or $P = P_{\rm s}P_{\rm w}$. This can be done in many different ways but the final result is not sensitive to details of the decomposition. With proper decomposition, $P_{\rm w}$ ends up being a decaying oscillation around unity[54–56].
- The wiggly component is separated from the measurements in a similar way.
- The leading order effect of non-linear evolution on the wiggly part of the power spectrum is scale-dependent damping of the oscillations, $P_{\rm w} \rightarrow (P_{\rm w} - 1)D(k) + 1$. The

simplest model $D(k) = \exp(-\Sigma^2 k^2)$, where Σ is a nuisance parameter, works reasonably well[7;57–65].

- The damped model is supplemented by nuisance parameters to account for small inaccuracies in separating the smooth part. These nuisance parameters are usually polynomials in k and they are smooth in the sense that they have little effect on the frequency of the oscillatory signal[35;66–68].
- The model is then dilated $P_{\rm w} \to AP_{\rm w}(\alpha k)$ to align its oscillations with the ones in the measurements.

In the simplest version of this analysis (when the fit is performed on the angularly averaged power spectrum), the dilation parameter α depends on the distance redshift relationship,

$$\alpha \equiv \alpha_{\rm V} = \frac{D_{\rm v}(z) r_{\rm d}^{\rm tmp}}{D_{\rm v}^{\rm fid}(z) r_{\rm d}}, \tag{2.1}$$

where $D_{\rm v}$ is the distance to the galaxies,

$$D_{\rm v}(z) = \left[\frac{czD_{\rm A}(z)^2}{H(z)}\right]^{\frac{1}{3}}, \tag{2.2}$$

$D_{\rm A}$ is the angular diameter distance, and H is the Hubble parameter. The superscript fid denotes values in "fiducial" cosmology assumed when converting redshifts of galaxies to distances and the superscript tmp is used to denote the BAO scale in the BAO template which may be (but does not have to be) identical to the fiducial cosmology. Quantities without superscripts are the real values corresponding to the data. The fiducial values assumed in the theoretical model and the BAO peak position in the template are known exactly. Thus, measuring α allows us to infer the ratio of the distance at a given redshift to the BAO scale, a quantity that is extremely sensitive to the properties of dark energy. The exact procedure used in high-precision data analysis is more involved and requires great care in every small detail. The fit is performed to both the angular monopole and the quadrupole of the correlation function, which enables us to also measure the anisotropic dilation parameter $\epsilon = H^{\rm fid} r_{\rm d}^{\rm tmp}/(H r_{\rm d})$[67;69–75], but the general idea remains the same. These scalings have been shown to be robust for a wide variety of cosmological models[76–79].

This procedure for extracting the BAO signal from the power spectrum has been successfully used on many galaxy samples[68;80–83]. The resulting distance measurements are leading sources of dark energy parameter constraints[84–96].

The physics governing the production of Baryon Acoustic Oscillations (BAOs) in the matter power spectrum is well understood[97–102]. In the very early universe, the distribution of matter was nearly uniform. Tiny overdensities began to grow under the influence of gravity, initiating the formation of dark matter halos and, eventually, galaxies. As baryonic matter was pulled into these overdensities, its temperature increased, generating outward radiation pressure that resisted further compression. This competition between gravity and pressure set up oscillations — effectively spherical sound waves — that propagated through the photon–baryon fluid.

Around 380,000 years after the Big Bang, the universe cooled enough for electrons and protons to combine into neutral atoms. This recombination era allowed photons to decouple from baryons, leading to a dramatic drop in pressure support. As a result, gravitational collapse became more efficient, accelerating the growth of large-scale structure. The remnants of these early sound waves — the imprints of baryon–photon interactions — are preserved today as BAOs: subtle modulations in the spatial distribution of baryonic matter.

The characteristic scale of these oscillations, determined by the physics of the early universe and its constituents (baryonic matter, dark matter, and dark energy), provides a powerful standard ruler. This imprint appears as a peak in the two-point correlation function at a comoving separation of $r_{\rm d} \sim 100$ Mpc. Because these oscillatory features exist on large, nearly linear scales, they are also expected to be observable in the galaxy distribution[33;103–106].

The same BAO feature is also present in other clustering measures e.g. the three-point correlation function, or its Fourier image, the bispectrum. In general, the bispectrum is more difficult to model compared to the power spectrum. Recent studies showed that the perturbation theory based models fail to accurately predict bispectrum measurements at surprisingly large scales[40;46;107–112]. Several recent works managed to measure the distance scale from the BAO feature in the bispectrum despite these difficulties[42;49]. A number of recent works[42;113;114] measured the BAO peak also in the galaxy three-point statistics. More works have studied different aspects of modeling the full bispectrum[7;39;49;50;115–121] and by comparing theoretical models of increasing difficulty with simulations[122–124].

When it comes to the BAO peak position, studies tend to focus on two-point statistics[1;26;35;125–127]. This is justified by the fact that the reconstruction of the galaxy field[33;128] moves some of the information from the three-point statistics – e.g. the three-point correlation function or the bispectrum – into the two-point statistics. However, measuring the BAO peak position from the bispectrum directly – or the combination of the non-reconstructed power spectrum and bispectrum – is still useful as the effect of the reconstruction on the information content of the higher-order statistics is, at the moment, not completely clear[129].

In this work, we explore the possibility of isolating the BAO feature in the galaxy bispectrum and modeling it with a simple model supplemented by enough nuisance parameters to account for inaccuracies in modeling its overall shape. The BAO feature in the Fourier space extends to high wavenumbers, but the quantity of interest is not the exact profile of the feature as a function of a wavenumber but its frequency, which is set by large-scale physics. The hope is that, if the BAO feature in the bispectrum is isolated, it will be possible to model it effectively even if the modeling of the full bispectrum is difficult. Our results are encouraging and suggest that surprisingly simple models of the BAO feature in the bispectrum are able to reproduce the BAO frequency with minimal bias. This is not surprising as the BAO signal is imprinted by linear physics in the early Universe. While the exact amount of power at high wavenumbers is difficult to theoretically compute, the quantity of interest in the BAO analysis is the beat frequency of the feature, and this quantity turns out to be more robust.

We use a simple, linear-theory-based model to describe the BAO feature in the bispectrum, in which the effect of nonlinear evolution is modeled by a wavenumber-dependent Gaussian damping term. (see section 2.3). We use `GLAM` simulations (see section 2.2) to val-

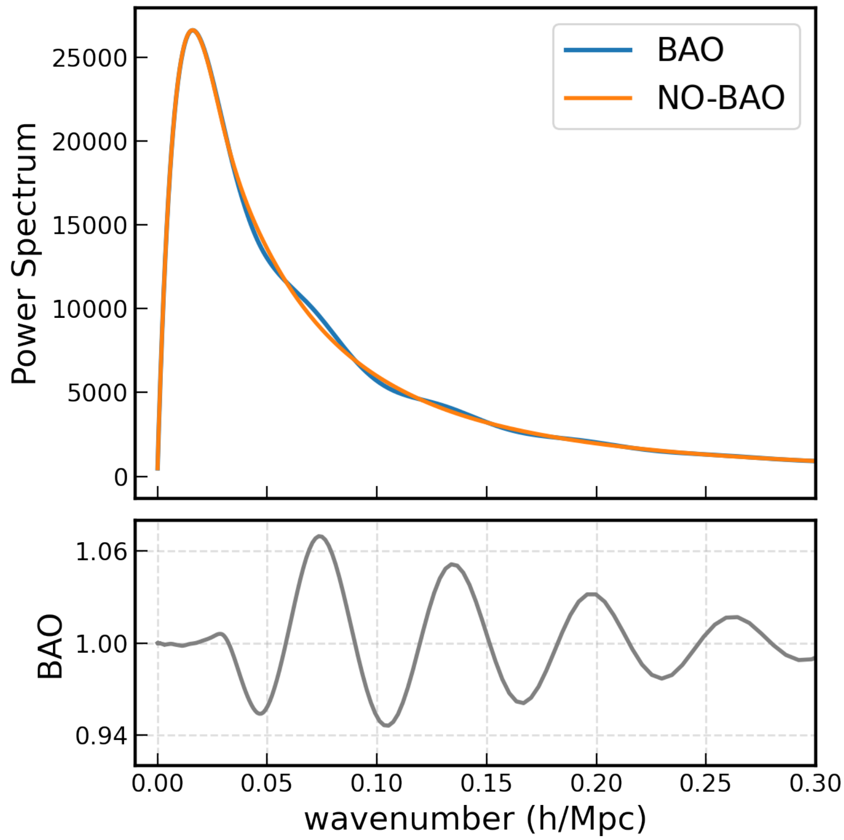

Figure 2.1: Initial power spectrum of `GLAM` *withBAO* and *noBAO* simulations and it's BAO signature.

idate this model. These simulations were run in pairs with and without the BAO feature in the initial conditions but otherwise identical. These simulations provide us with the "empirical" data for the smooth and wiggly bispectra. When fitting our model to simulations, we find that the extracted BAO scale is unbiased even when the fiducial template for the BAO is constructed in a slightly offset cosmological model (see section 2.4). The recovered values of α are within 0.3 per cent of the true values even when the BAO feature in the fitting template deviates from the true value by as much as 3 per cent.

2.2 BAO Signal in Simulations

2.2.1 GLAM simulations

We use GLAM-PMILL simulations[130] to test methods for BAO feature separation and theoretical modeling. The GLAM-PMILL catalogs used in this work are publicly available in the Skies & Universes website[i].

The GLAM simulations[51] are N-body cosmological simulations that, in the case of the PMILL series, follow the evolution of $2,000^3$ dark matter particles. Each particle has a mass of $1.065 \times 10^{10}\,h^{-1}\mathrm{M}_\odot$ in a comoving periodic box of size $L = 1\,h^{-1}\mathrm{Gpc}$, with $N_\mathrm{s} = 136$ time-steps and a mesh of $N_\mathrm{g} = 4,000$ cells per side. This results in a spatial resolution of $0.250\,h^{-1}\mathrm{Mpc}$.

The simulations were produced in a flat ΛCDM cosmology, adopting the cosmological parameters assumed in the Planck Millennium simulation (PMILL; Baugh et al.[52]). PMILL uses the best-fitting parameters from the first Planck 2013 data release[131]. Each instance of the simulation was run with two initial conditions, starting at $z_\mathrm{ini} = 100$: one with the BAO signal imprinted in the initial conditions, and one without. The pairs of simulations are otherwise identical, i.e. they had matching random phases in the initial conditions. The GLAM-PMILL simulations used in this work consist of 1,000 realizations with and another 1,000 without BAO. Figure 2.1 shows the power spectrum used for the initial conditions of the GLAM simulations. The initial power for the pair of simulations was identical, except the "withBAO" instances had periodic BAO wiggles in them, and the "noBAO" instances didn't have the feature.

The GLAM simulations are only capable of correctly resolving distinct halos (not subhalos), with virial masses greater than $\sim 10^{12}\,h^{-1}\mathrm{M}_\odot$ for the case of the GLAM-PMILL suite. The resulting dark matter halos were populated by galaxies using GALFORM semi-analytical model of galaxy formation[130], but in this work we only use the halo catalogs, both for BAO and noBAO realizations.

[i] https://www.skiesanduniverses.org/Products/MockCatalogues/GLAMDESILRG/

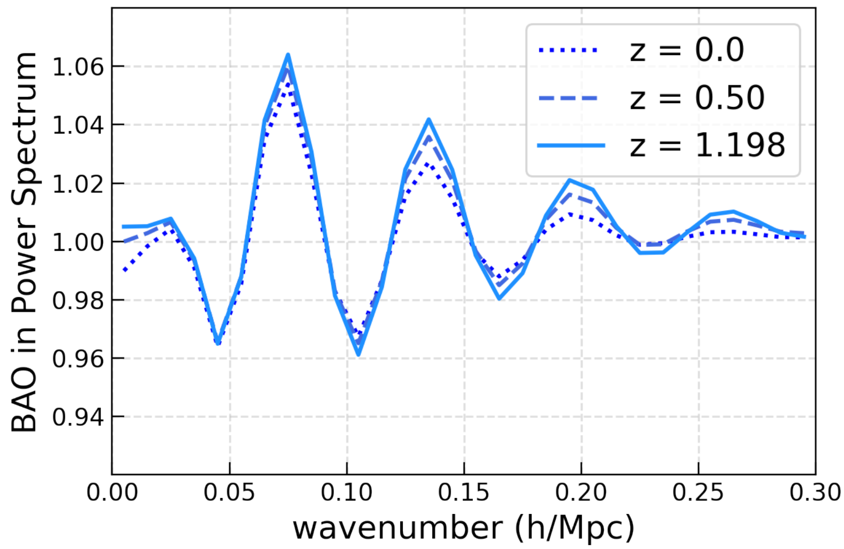

Figure 2.2: BAO signature in the power spectrum of `GLAM` simulations at different redshifts.

We will treat the measurements from these simulations as the ground truth for the rest of our work. The theoretical models will be evaluated based on how well they recover the measurements of `GLAM` simulations.

2.2.2 BAO in the power spectrum

We measure the power spectrum from both the *withBAO* and *noBAO* simulations, $P_{\rm w}$ and $P_{\rm s}$ respectively (see Appendix 2.A for the details). The power spectrum measurements are averaged in bins of width $\Delta k = 0.01h$ Mpc^{-1} starting from $k_{\rm min} = 0.005h$ Mpc^{-1} up to $k_{\rm max} = 0.3h$ Mpc^{-1}, resulting in 30 bins. This wavenumber range contains most of the linear and semi-linear modes of galaxy overdensity. There is very little BAO signal above our maximum wavenumber.

Figure 2.2 shows the ratio of $P_{\rm w}$ to $P_{\rm s}$ measured from `GLAM` at different redshifts. This is by definition the BAO signal in the power spectrum. The BAO signal looks exactly as expected. The amplitude of wiggles decays with decrease in redshift. The decay is wavenumber-dependent. BAO in high wavenumber modes decays faster with redshift as expected.

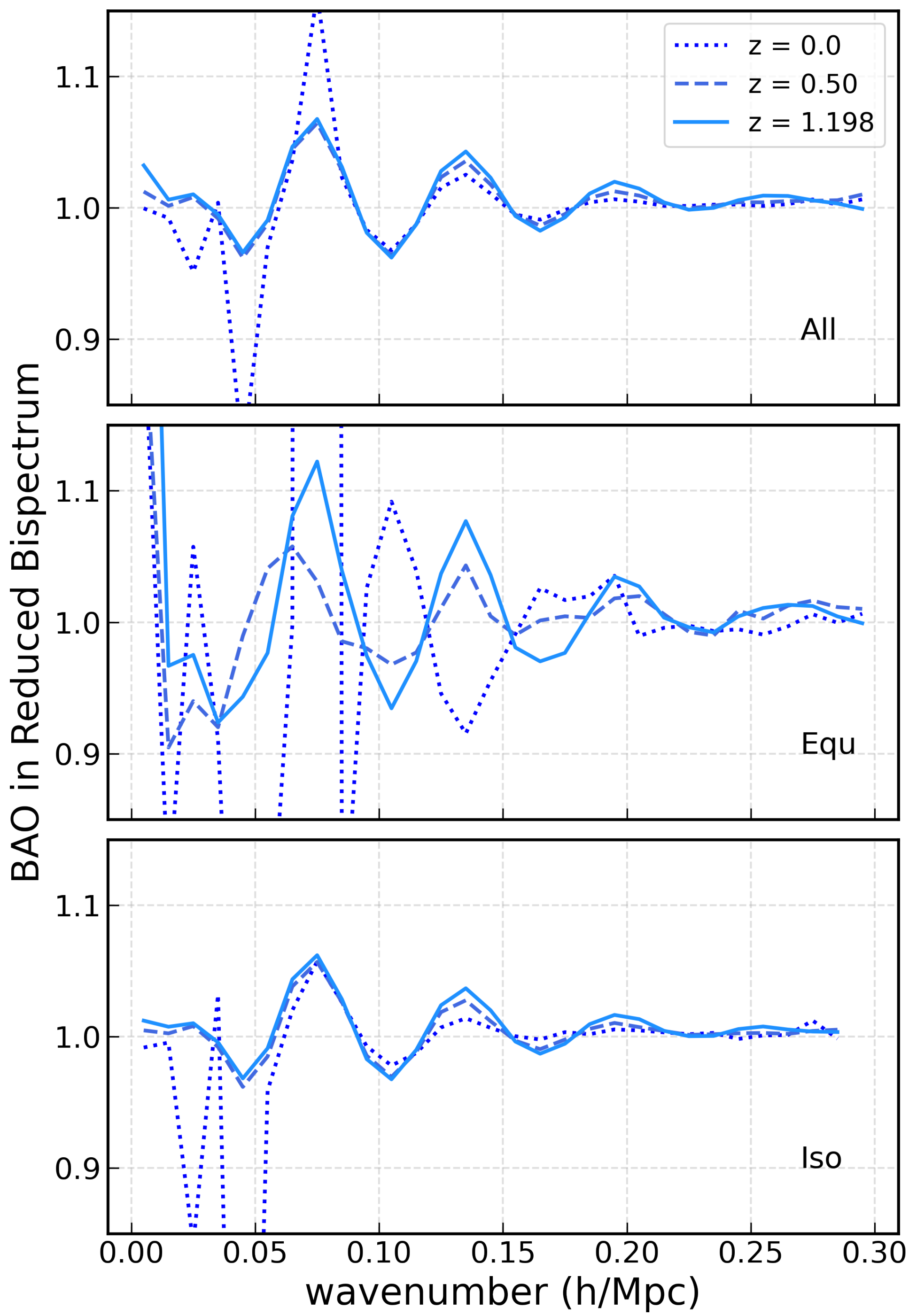

Figure 2.3: BAO signature in the bispectrum of `GLAM` simulations at different redshifts. Labels All (top panel), Equ (middle panel), and Iso (lower panel) refer to averages over all, equilateral, and isosceles configurations respectively.

2.2.3 BAO in the bispectrum

We define the BAO feature in the bispectrum similarly to that of the power spectrum. We measure the angularly averaged bispectrum - $B(k_1, k_2, k_3)$ - from *withBAO* and *noBAO* `GLAM` simulations, B_w and B_s respectively (see Appendix 2.A for the details). We bin each wavevector in 30 bins similar to how we did it for the power spectrum. This results in a total of 2600 bins for the (k_1, k_2, k_3) triplet. To map the 3D configuration space of triangles into a 1D index for analysis, we employ a "triangle index" that orders the triangles so that $k_1 \leq k_2 \leq k_3$. This ensures each unique triplet is only counted once without duplication since the bispectrum is invariant to any permutation of the indices.

Figure 2.3 shows the ratio of B_w to B_s at different redshifts. For better visualization, the measurements are averaged over k_2 and k_3 for different triangular configurations. We look at three different averages. In $B^\mathrm{all}(k)$ the value of k_1 is kept within a certain bin while all other values of k_2 and k_3 are averaged over. In $B^\mathrm{iso}(k)$ we average over bispectra that have the value of k_1 within a certain bin and $k_2 = k_3$. In $B^\mathrm{equ}(k)$ we average over all bispectra that have all three wavenumbers within a certain bin.

$$B^\mathrm{all}(k) = \sum_{k_2,k_3} \frac{B(k, k_2, k_3)}{N_k}, \tag{2.3}$$

$$B^\mathrm{iso}(k) = \sum_{k_2,k_3} \frac{B(k, k_2, k_3)_{k\neq k_2=k_3}}{N_{k\neq k_2=k_3}}, \tag{2.4}$$

$$B^\mathrm{equ}(k) = \sum_{k_2,k_3} \frac{B(k, k_2, k_3)_{k=k_2=k_3}}{N_{k=k_2=k_3}}, \tag{2.5}$$

where, N_k, $N_{k\neq k_2=k_3}$ and $N_{k=k_2=k_3}$ are the number of triangles that fall into the relevant configurations. We only do this to make the comprehansion of the plots easier. All our analysis are performed on the original (pre-averaged) 2600 bispectrum measurements.

Even though this averaging is not optimal for the retention of the BAO feature, it is still clearly visible in all configurations. The equilateral configuration has the most noise since each point in the k axis corresponds only to a single equilateral triangle. In the bispectrum, the BAO feature looks qualitatively similar. It is a decaying oscillation which is more pronounced at higher redshifts.

2.3 BAO modeling

2.3.1 Power spectrum

To model the BAO in the power spectrum we start with the linear model $P_{\rm lin}(k)$ that we compute using computer code `CAMB`[24] in a fiducial cosmological model. We split this power spectrum into BAO and smooth parts, $P_{\rm lin}^{\rm BAO}$ and $P_{\rm lin}^{\rm s}$ respectively, using `cosmoprimo` package. `cosmoprimo` has many options for performing this split. We use the option based on the work in[132] and[133], but we verified that this choice does not affect our results. We reduce the BAO feature by damping it with a Gaussian term $D(k,\Sigma) = \exp(-\Sigma^2 k^2/2)$ so that our final power spectrum template is

$$P^{\rm w}(k;\alpha,\Sigma) = P_{\rm lin}^{\rm s}(k)[(P_{\rm lin}^{\rm BAO}(\alpha k) - 1)D(k,\Sigma) + 1] \tag{2.6}$$

The dilation parameter α is only applied to the BAO wiggles. The superscript w denotes the fact that the template has BAO wiggles in it. We use just the smooth part of the linear power spectrum as the no-wiggle template.

$$P^{\rm nw}(k) = P_{\rm lin}^{\rm s}(k) \tag{2.7}$$

Formally, this is a limit $\Sigma \to 0$ of the wiggled template.

2.3.2 Bispectrum

Our bispectrum BAO model is inspired by the leading order perturbation theory with linear local bias given by[43].

$$B(\boldsymbol{k}_1, \boldsymbol{k}_2, \boldsymbol{k}_3) = 2\boldsymbol{Z}_1(\mu_1)\boldsymbol{Z}_1(\mu_2)\boldsymbol{Z}_2 P(\alpha k_1)P(\alpha k_2)+ \\ \text{cyclic terms} + [P(k_1) + P(k_2) + P(k_3)]S_1 + S_0, \tag{2.8}$$

where,

$$\boldsymbol{Z}_1(\mu) = (b_1 + f\mu_1^2), \tag{2.9}$$

$$\boldsymbol{Z}_2 = \left[\frac{b_2}{2} + b_1 F_2(\boldsymbol{k}_1, \boldsymbol{k}_2) + f\mu_3^2 G_2(\boldsymbol{k}_1, \boldsymbol{k}_2) \right. \\ \left. - \frac{f\mu_3 k_3}{2}\left(\frac{\mu_1}{k_1}(b_1 + f\mu_2^2) + \frac{\mu_2}{k_2}(b_1 + f\mu_1^2)\right)\right], \tag{2.10}$$

$\mu_i = \boldsymbol{k}_i \cdot \hat{\boldsymbol{z}}/k_i$ with $\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_3 = 0$ to account for the triangular condition.

$$F_2(\boldsymbol{k}_1, \boldsymbol{k}_2) = \frac{5}{7} + \frac{\boldsymbol{k}_1 \cdot \boldsymbol{k}_2}{2k_1k_2}\left(\frac{k_1}{k_2} + \frac{k_2}{k_1}\right) + \frac{2}{7}\left(\frac{\boldsymbol{k}_1 \cdot \boldsymbol{k}_2}{k_1k_2}\right)^2, \tag{2.11}$$

and

$$G_2(\boldsymbol{k}_1, \boldsymbol{k}_2) = \frac{3}{7} + \frac{\boldsymbol{k}_1 \cdot \boldsymbol{k}_2}{2k_1k_2}\left(\frac{k_1}{k_2} + \frac{k_2}{k_1}\right) + \frac{4}{7}\left(\frac{\boldsymbol{k}_1 \cdot \boldsymbol{k}_2}{k_1k_2}\right)^2, \tag{2.12}$$

The cyclic terms can be derived by replacing indices 1 and 2 in the first term with 2 and 3, and 1 and 3 respectively. f, b_1, b_2, S_0 and S_1 are free parameters of the fit.

Gagrani and Samushia[134] showed that most of the information is in the azimuthal averages of the first three even multipoles with $m = 0$ in the multipole expansion containing most of the information relevant to the derivation of cosmological constraints. In this work, we only consider the angularly averaged bispectrum (or the monopole) defined as

$$B_0(k_1, k_2, k_3) = \int_{-1}^{+1} d(\cos\theta_1) \int_{0}^{2\pi} d\phi B(k_1, k_2, k_3, \theta_1, \phi). \tag{2.13}$$

We integrate the five-dimensional bispectrum numerically to obtain the three-dimensional bispectrum monopole.

The model depends on wavevector lengths k_1, k_2 and k_3, a dilation parameter α and nuisance parameters b_1, b_2, and f. In perturbation theory, the last three parameters have a physical meaning (e.g. f is interpreted as a growth rate of structure). We treat these parameters as true nuisance parameters. Their only purpose is to give us a reasonable range of bispectrum shapes to separate the BAO feature. We do not assign to them any physical meaning.

We compute B by using P^{w} and B^{s} by using the no-wiggled template P_{s}. The ratio B/B^{s} then gives a model for the BAO signature in the bispectrum.

In the end, the model depends on parameters α, b_1, b_2, f, S_0, S_1, and Σ.

This single universal scaling model is not accurate at the sub-percent precision[135] but we ignore the higher order effects in this work since they do not really affect any of our main conclusions.

2.4 BAO fitting

We fit the model described in section 2.3 to the `GLAM` measurements described in section 2.2 by running a Monte Carlo Markov Chain (MCMC) using the `emcee` Python package[136]. The MCMC was run with 70 walkers for 2000 steps, discarding the first 800 steps as burn-in to ensure convergence. The chain is run over α and 6 nuisance parameters (f, b_1, b_2, S_0, S_1, Σ). The first three parameters in that list control the relative amplitude at different wavenumber triplets, the next two provide wavenumber-dependent and wavenumber-independent additive terms to remove the effect of shot noise, and the last parameter controls how pronounced the BAO feature is. The parameter priors are wide enough to encompass all areas of high likelihood. We find the best-fit values by minimizing

$$\chi^2 = \sum_{ij} X_i W_{ij} X_j, \tag{2.14}$$

where X is a vector of measurements (either $P_\mathrm{w}/P_\mathrm{s}$ at different wavenumber or $B_\mathrm{w}/B_\mathrm{s}$ at different wavenumber triplets) and W_{ij} are weights that upweight or downweight a relative contribution from different wavenumbers to the overall fit. The most optimal weights are the ones given by the inverse covariance of our measurements. Unfortunately, we don't have a reliable way of estimating the inverse covariance. We use the inverse of a sample variance over 1,000 `GLAM` mocks as our weights. We can not use the posterior of our likelihood to approximate the true constraining power of the BAO signal in the bispectrum since the sample covariance computed from a small number of mocks is a very noisy estimator of a true covariance and the Hartlap correction factor[137], commonly used to unbias the inverse of the covariance matrix, was not applied in this case due to the insufficient number of realizations. The minimum of the Eq. 2.14 is still an unbiased estimator and we can use it to estimate the systematic offsets in our fits. Our measurements are computed from a very large cumulative volume of $1{,}000 h^{-1}$ Gpc and we expect the uncertainty in the systematic offset to be negligible.

α is the only parameter of interest to us, the other six are nuisance parameters that we marginalize over. We have another hidden degree of freedom in this fit, which is the fiducial cosmology that we choose to produce the original P_lin template. We are hoping that no matter which template we start with, it can be made to fit the measurements by an appropriate rescaling of the oscillation frequency with α and damping the amplitude of oscillations at high wavenumbers with Σ. The dilation parameter is expected to be

$$\alpha = \frac{r_\mathrm{d}^\mathrm{tmp}}{r_\mathrm{d}}. \tag{2.15}$$

We do not have scaling with D_V like in Eq. (2.1). This is because we are measuring our bispectra in cubic boxes with known physical size and we are not affected by Alcock-Patczynski scaling. We will show the results of fitting to the redshift $z = 0.5$, but we also verified that the same conclusions hold for redshifts $z = 0$ and $z = 1.198$.

2.4.1 Aligned Template

We first check what happens when the template is computed in the actual cosmology of `GLAM` simulations. In this case, the BAOs in the template and the measurements are aligned and the expected value of α is one. The results from running our MCMC chains for the aligned template for the redshift of $z = 0.5$ are shown in Figure 2.4. The results for other redshifts look qualitatively similar. The likelihood contours are reasonably close to Gaussian. While some of the nuisance parameters are correlated with each other, the α_{V} parameter is uncorrelated with any of the nuisance parameters, suggesting that its maximum likelihood values are weakly affected by possible posterior volume effects. The f parameter is allowed to take on negative values, under the standard interpretation of its meaning this would be unphysical. We, however, do not interpret f as a growth rate of structure but rather use it as a pure nuisance parameter that is used to help us span the range of possible shapes for the smooth part of the bispectrum, so this is not a problem for our analysis. We checked that restricting f to have only positive values does not affect constraints on α, which is not surprising since f and α are very weakly correlated. Our recovered value of α is consistent with unity to 0.03 percent precision. It's worth noting that since we do not have reliable estimates of the covariance matrix, the widths of our MCMC chains can not be interpreted as errors in the measurement. We do not have at present a good way of estimating the variance in our measurements but we expect it to be subdominant to the systematic error from the fit since we are fitting the mean measurement of 1,000 one cubic Gigaparsec mocks.

Figure 2.5 shows the bispectrum BAOs from `GLAM` measurements at redshifts 0.0, 0.5, and 1.198, and their corresponding best-fit models evaluated at the best values derived from MCMC chains. The models match the measurements well and by visual inspection

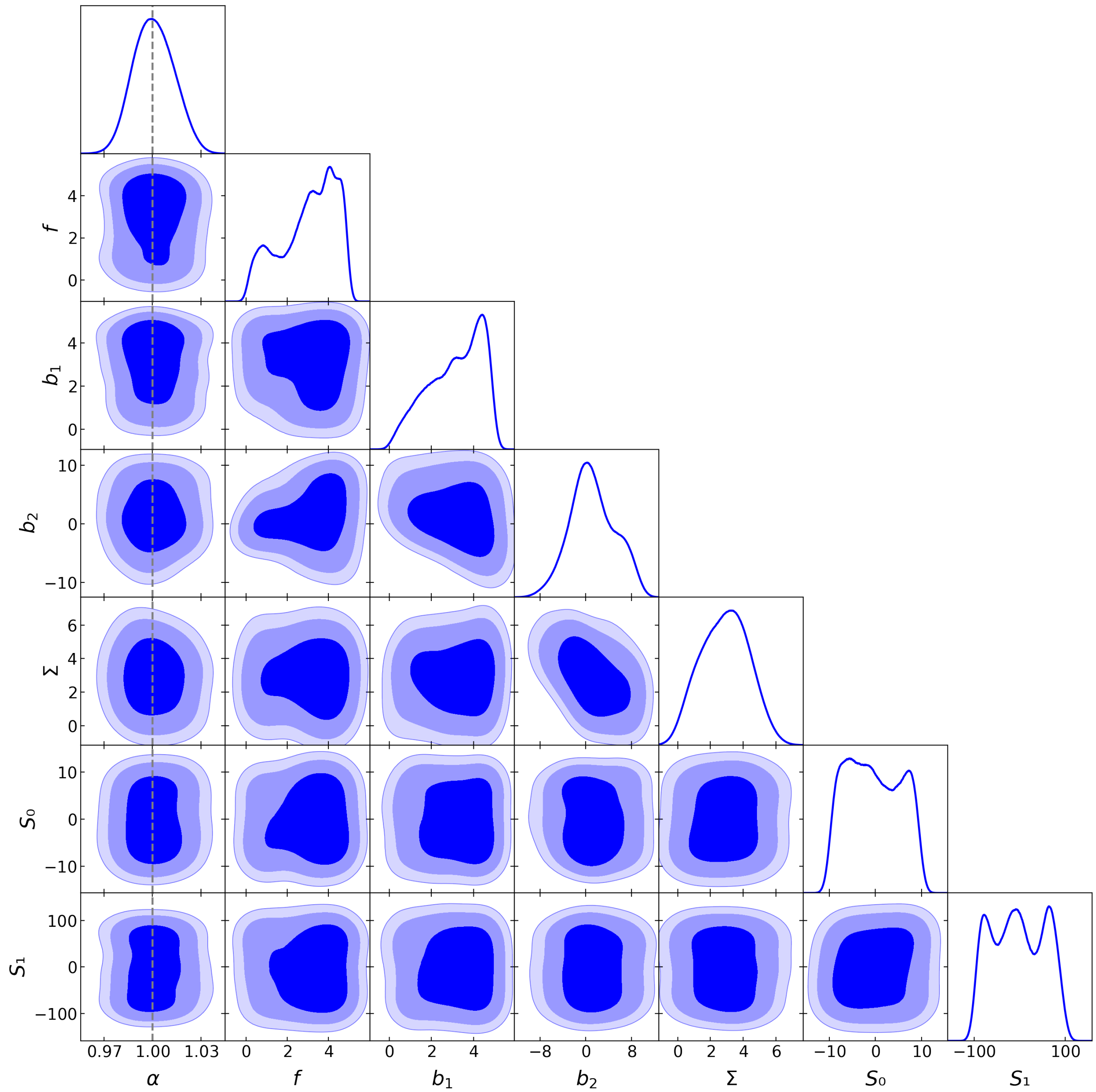

Figure 2.4: Equipotential contours of the MCMC posterior from fitting the BAO in `GLAM` bispectrum at $z = 0.5$. The grey dashed line shows the expected value of $\alpha = 1$.

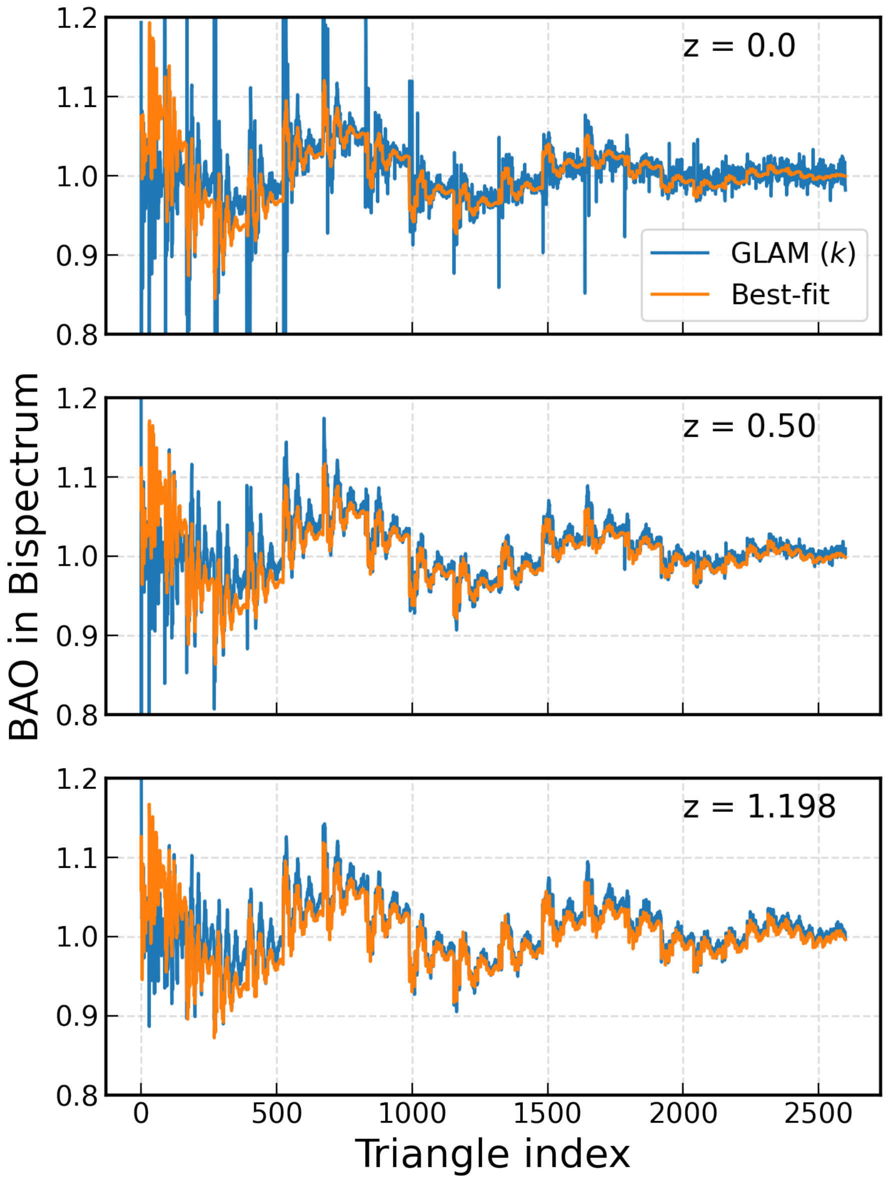

Figure 2.5: BAO signature in the bispectrum as a function of triangle index (see definition in Sec 2.2.3) at different redshifts. Measured bispectrum BAOs from the `GLAM` simulations (the blue line) is compared to our best-fit model (the orange line).

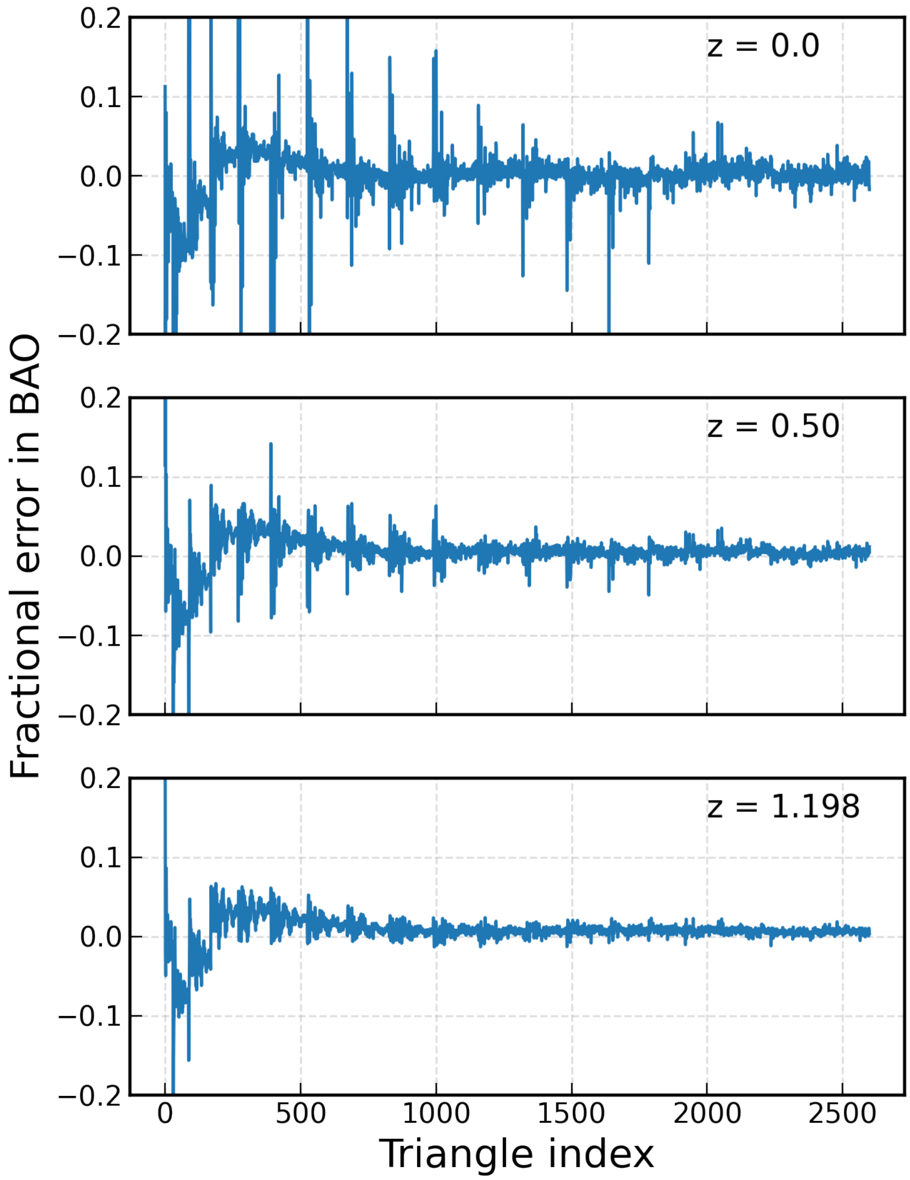

Figure 2.6: Fractional deviation between the measured bispectrum BAOs from the GLAM simulations and our best-fit model as a function of triangle index (see definition in Sec 2.2.3) at different redshifts.

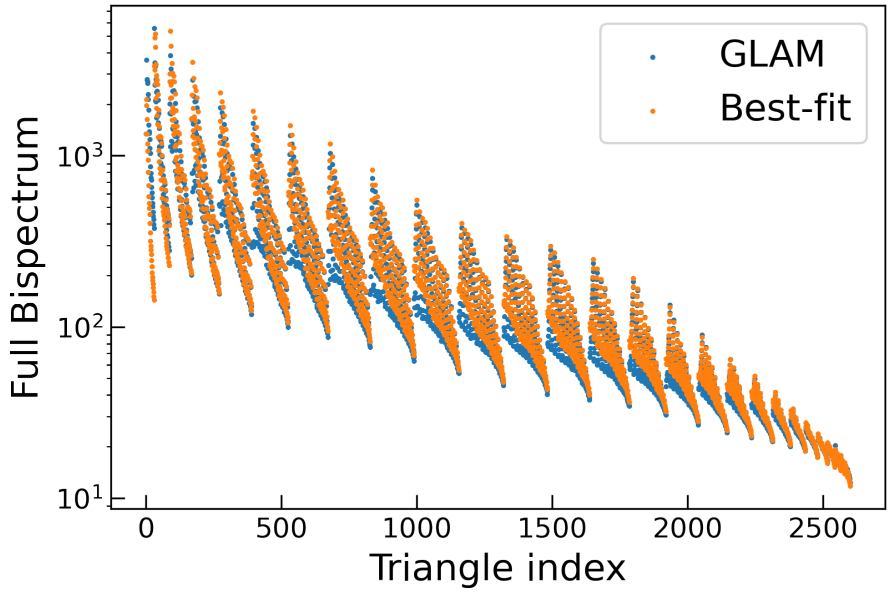

Figure 2.7: Full bispectrum as a function of triangle index (see definition in Sec 2.2.3) at redshift 0.5. Measured bispectrum from the `GLAM` simulations (the blue line) is compared to our best-fit model (the orange line).

successfully recover all features in the measurements. Figure 2.6 shows the same information but now as a ratio of the predicted and measured bispectra. At $z = 0.0$ there are some bispectrum triplets that have a large fractional deviation between the measurement and the theory. This is likely the failure of the simple model adopted in our work which does capture the overall frequency of the BAO oscillations well but fails to predict the amplitude of the BAO for some triplets.

Figure 2.7 shows a similar best fit to the entire bispectrum (not the BAO only like in the previous case). As expected the simple model can not fit the full shape of the bispectrum well, there are obvious and irreconcilable offsets between the measured and theoretical bispectra at a wide range of wavenumbers and triangular configurations.

To aid visual inspection we also plot the measured and best-fit model bispectra for averaged, isosceles, and equilateral triangles as defined by Eqs. (2.3)-(2.5). Figure 2.8 shows this comparison for the bispectrum measured at $z = 0.5$. The top panel shows the `GLAM` BAOs along with the model computed for the original templates derived from `CAMB`. The middle row shows the effect of the damping parameter Σ on the model for the best-fit Σ from the

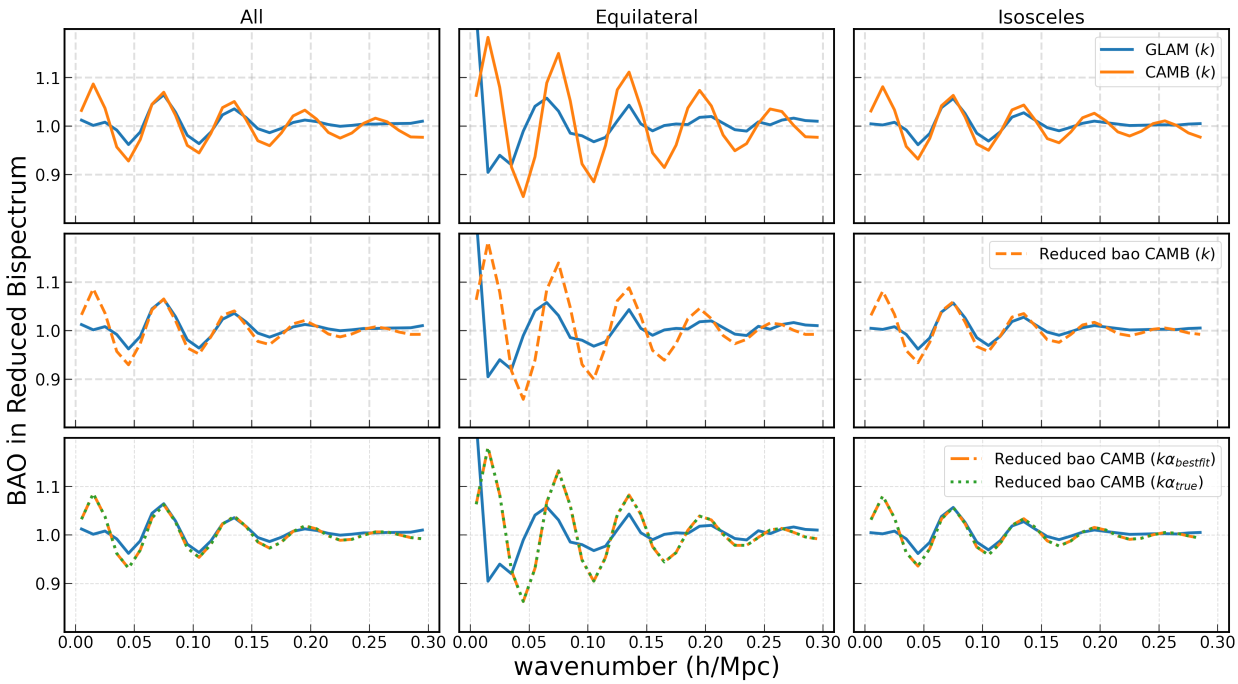

Figure 2.8: Measured BAO in the bispectrum compared to the model. The columns show bispectrum averaged over all (left), equilateral (middle), and isosceles (right) configurations. The blue lines denote the measurements and are identical along the columns. The solid orange lines in the top row denote the model computed based on a CAMB linear power spectrum. The dashed orange lines in the middle row denote the model with the BAO damping applied, which reduces the amplitude of the BAO wiggles. The dot-dashed yellow line and the dotted green line in the bottom row denote the model stretched by the best fit and true α values respectively.

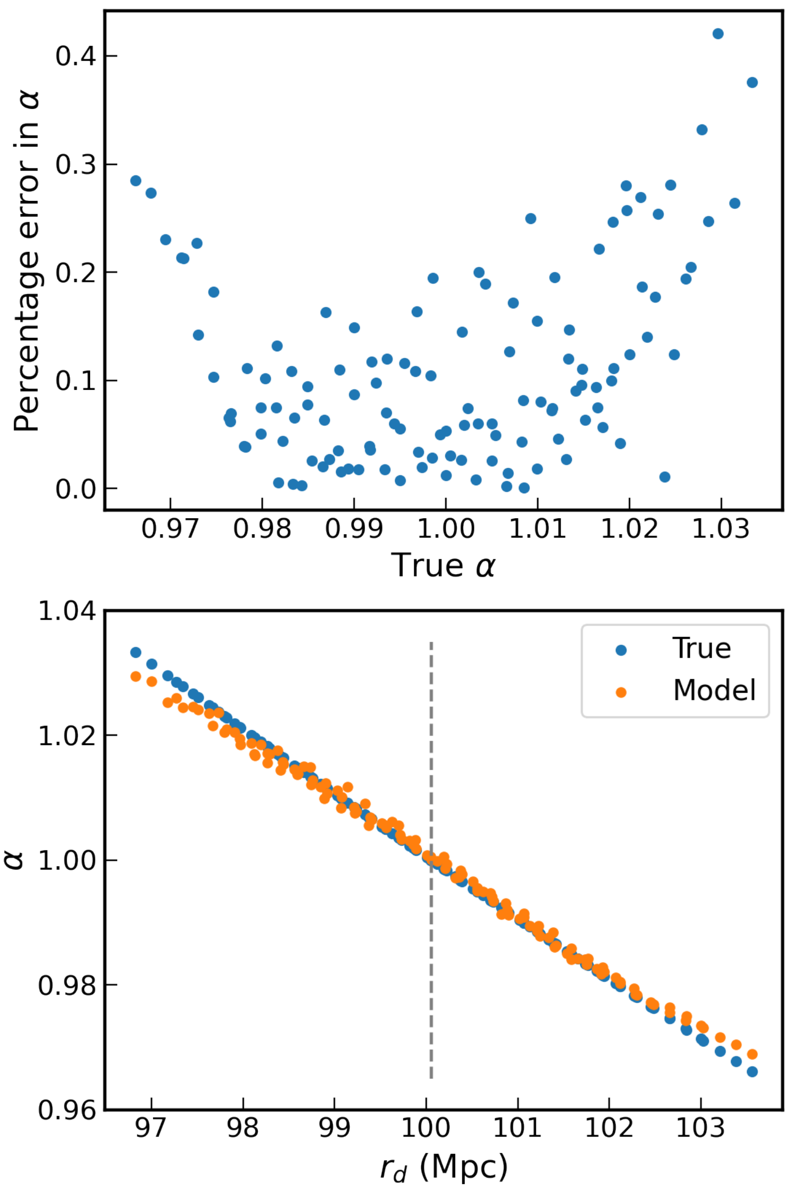

Figure 2.9: Percentage systematic error in the recovered value of α as a function of the expected alpha.

MCMC chains. The bottom row shows the result of stretching it by the best-fit value of α. Since our best-fit is very close to one, the best-fit BAO template is difficult to distinguish from the "true" BAO template (evaluated at $\alpha_V = 1$), which we also plot on the same figures for comparison. Visual inspection confirms that the best-fit template is indeed a good fit for the harmonic feature imprinted in the measurements. The template seems to be offset for the equilateral triangles. This can be explained by high noise in this specific measurement. The averaged and isosceles measurements are averages over many triangular configurations, while the equilateral measurement is averaged over very few triangular configurations.

2.4.2 Systematic effects due to offset template

We now check whether the model performs equally well when the template is computed in a different cosmology from the `GLAM` so that the position of the BAO peak is offset. To perform this test we generate templates in 100 different cosmologies around `GLAM` cosmology in a range $0.277 \leq \Omega_m \leq 0.327$. We then perform the fitting as before. Figure 2.9 shows the results of this exercise. Each point on the top panel corresponds to a template used in the fits. The horizontal axis shows the expected value of α and the vertical axis shows the percentage offset of the best fit from that expectation. In the ideal case, we want the points to be as close as possible to zero. The bottom panel shows the same data from a different point of view. The points are again individual templates, the horizontal axis is the value of r_d in the template and α is the amount of dilation needed to bring the template to be identical to the true BAO signal. Blue dots show our expectations and the orange dots show the actual results. We expected the systematic offset between the two to increase as the offset between the template and the simulations increases, which seems to be the case. The systematic offsets are mostly within 0.3 per cent as long as the frequency of the BAO in the template does not differ by more than 3 per cent from the frequency in the data.

2.5 Conclusions

We explored the possibility of modeling the BAO feature in the galaxy bispectrum. Even though some of this information is recovered during the reconstruction of the cosmological fields, the measurements of the BAO in the galaxy bispectrum have the potential to strengthen our constraints on key cosmological parameters. In the measured bispectrum the BAO signal is visible at wavenumber as high as $k \sim 0.3h^{-1}$ Mpc at redshifts of $z = 0.5$ and above. At lower redshifts, the BAO signature at higher wavenumber is damped by nonlinear evolution. We showed that even though simple perturbation-based models of bispectrum fail to model it accurately at higher wavenumbers, they work reasonably well when the smooth component is subtracted and only the BAO feature is modeled. We validated this on a suit of `GLAM` simulations that have been run with and without the BAO feature in the initial conditions. The ratio of the bispectra measured from those simulations is by definition the BAO feature in the bispectrum. Our six-parameter model was able to recover the unbiased value of α parameter when fit to the bispectrum monopole measured from a cumulative volume of $1,000\,h^{-3}\mathrm{Gpc}^3$. In real analysis, we don't know a priori what cosmological model to use when producing BAO templates. It is therefore important that the procedure results in unbiased results for a range of reasonable templates. We checked the robustness of the model by fitting the same measurements with the templates computed in offset cosmologies. By performing similar validation tests we demonstrated that the bias in the recovered values of α is at most 0.3 per cent for deviations up to 3 per cent from the true value of the sound horizon. This is good enough when fitting to currently existing data but may need improvement when future more precise data arrives (e.g. from the DESI and Euclid experiments). Possible avenues for decreasing the bias are computing the BAO template with more sophisticated modeling of nonlinear physics or using simulations for calibrating the templates.

Acknowledgements

We are grateful to Francisco Prada and Anatoly Klypin for their help with `GLAM` simulations.

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231

This research has made use of NASA's Astrophysics Data System and the arXiv open-access repository of electronic preprints and postprints.

JB, MR, and LS are grateful for support from the US Department of Energy via grants DE-SC0021165 and DE- SC0011840. JB is partially supported by the NASA ROSES grant 12-EUCLID12-0004. JE acknowledges financial support from the Spanish MICINN funding grant PGC2018-101931-B-I00. We thank Instituto de Astrofísica de Andalucía CSIC for hosting the `Skies & Universes` site for cosmological simulation products. GLAM A-series simulations were performed at the DIRAC@Durham facility managed by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility in the UK.

Data Availability

Any data product generated during this research will be shared upon request. The `GLAM` simulations are available for download from https://www.skiesanduniverses.org/Products/MockCatalogues/GLAMDESILRG/.

"ସୂକ୍ଷ୍ମ ଲହରୀ ମଧ୍ୟରେ ରହିଥାଏ ସୃଷ୍ଟିର ଗାଥା — ଜେଉଁଠି ପ୍ରତି ଘାତ ହେଉଛି ଏକ ନିମିତ୍ତ।"

— Inspired by Vedic sound metaphysics

"Within subtle waves lies the story of creation — each oscillation a cause."

Appendix

2.A Measuring Power Spectrum and Bispectrum

We divide every simulation cube into a 1024^3 grid. We assign halos to grid points based on the triangular shaped cloud scheme described in[138] to compute the number of particles in each grid cell - n_{ijk}. We then compute the overdensity field

$$\delta_{ijk} = \frac{n_{ijk} - \overline{n}}{\overline{n}}, \tag{2.16}$$

where $\overline{n}$ is the average number density

$$\overline{n} = \frac{1}{1024^3} \sum_{ijk} n_{ijk}. \tag{2.17}$$

We then perform a 3D discrete Fourier transform of the overdensity field

$$\tilde{\delta}_{ijk} = \mathcal{F}_{ijk}^{\ell mn}[\delta_{\ell mn}]. \tag{2.18}$$

The $\tilde{\delta}_{ijk}$ numbers can be arranged in a 3D cube so that each measurement corresponds to a certain wavevector $\mathbf{k} = (k_i, k_j, k_k)$[139–141]. The average power spectrum within a bin $(k_{\text{min}}, k_{\text{max}})$ is computed as the average value of all $\tilde{\delta}_{ijk}\tilde{\delta}^{\star}_{ijk}$ such that the length $|\mathbf{k}| = \sqrt{k_i^2 + k_2^2 + k_3^2}$ falls within the bin. The average bispectrum within a bin $(k_{i,\text{min}}, k_{i,\text{max}})$, where $i = (1, 2, 3)$ is similarly computed as the average of all $\tilde{\delta}_{i_1 j_1 k_1}\tilde{\delta}_{i_2 j_2 k_2}\tilde{\delta}_{i_3 j_3 k_3}$ such that wavevectors associated to the three $\tilde{\delta}$ fall within the corresponding bins and their sum is zero.

Chapter 3

Galaxy-Halo Connection

3.1 Cosmological Context and Structure Formation

Modern cosmology is built on the framework of the Λ Cold Dark Matter (ΛCDM) model, which describes a universe that is spatially flat, currently composed predominantly of dark energy ($\sim 70\%$) and cold dark matter ($\sim 25\%$), with only a small fraction contributed by baryonic matter and radiation[94]. Under this model, the universe began in a nearly uniform state after the Big Bang, with tiny fluctuations in the primordial density field seeded by quantum processes during inflation.

These fluctuations, observable today as anisotropies in the Cosmic Microwave Background (CMB), serve as the initial conditions for the formation of large-scale structure as in Fig. 3.1.1. Over cosmic time, gravity amplifies overdense regions, causing matter to collapse into a hierarchy of structures, from dark matter haloes to galaxies, groups, clusters, and ultimately the filamentary cosmic web. Understanding galaxy formation requires modeling how baryons evolve within this dark matter-dominated framework.

A crucial outcome of this gravitational evolution is the assembly of dark matter haloes, which act as the scaffolding for galaxy formation. Gas accretes into haloes, cools, and condenses into stars, leading to the emergence of the diverse population of galaxies observed today. This hierarchical process, where smaller systems merge to form larger structures,

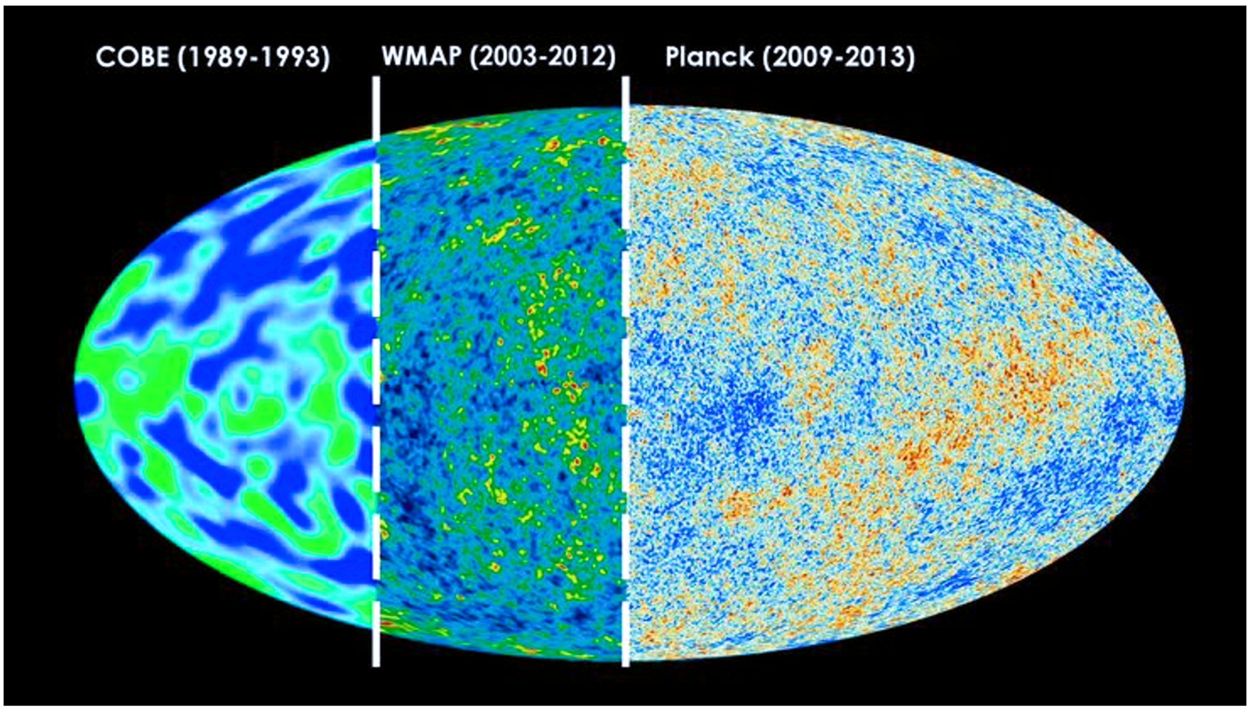

Figure 3.1.1: Temperature fluctuations in the Cosmic Microwave Background, as measured successively by the COBE[3], WMAP[4], and Planck[5] satellites, illustrate the increasing precision of cosmological measurements over time. Although the CMB appears nearly homogeneous and isotropic, tiny anisotropies at the level of one part in 10^5 encode the seeds of the large-scale structure observed in the Universe today. These fluctuations are remnants of quantum perturbations amplified during inflation and later evolved via gravitational instability. The figure illustrates this progression in observational capability. Image courtesy of NASA.

underpins our understanding of galaxy evolution[104;142].

Structure Growth and Cosmic Expansion

The growth of cosmic structure is modulated by the expansion history of the universe, encoded in the Hubble parameter $H(z)$. During the matter-dominated era, structure formation proceeds efficiently, but the onset of dark energy domination at $z \lesssim 1$ slows down gravitational collapse. This leaves an imprint on the abundance and properties of haloes and galaxies at different epochs.

Quantifying this growth requires both theoretical modeling and observational probes, including weak lensing, galaxy clustering, and the cosmic microwave background. Large-scale simulations such as IllustrisTNG[143–147], Uchuu[148], EAGLE[149;150], and Millennium[104] have played a vital role in connecting theory to observation, enabling the prediction of galaxy properties from initial conditions in a cosmologically consistent manner.

Galaxy Formation in a Cosmological Framework

Within this evolving cosmological background, galaxy formation is governed by a complex interplay of processes:

- **Gas accretion and cooling**: baryons fall into haloes and radiatively cool.
- **Star formation and feedback**: stars form in dense regions; supernovae and AGN inject energy back into the interstellar and circumgalactic media.
- **Chemical enrichment**: stellar evolution produces metals that are recycled into the gas phase, enriching future generations of stars.
- **Mergers and interactions**: hierarchical assembly leads to dynamical interactions that reshape galaxy morphology and trigger starbursts or quenching.

These processes are inherently nonlinear and span a vast range of spatial and temporal scales. While semi-analytic models (SAMs) and hydrodynamical simulations have advanced our understanding significantly, they remain computationally expensive and model-dependent, motivating the development of alternative techniques, including those based on machine learning.

This cosmological perspective sets the stage for the remainder of this chapter. In the following sections, we review how galaxies are modeled from the properties of their host haloes, how galaxy formation histories are predicted using neural networks, and why a stochastic treatment of unresolved variability is required to improve their realism.

3.2 The Galaxy–Halo Connection and Modeling Approaches

In the framework of hierarchical structure formation, dark matter haloes form through the gravitational collapse of overdense regions in the early universe. These haloes serve as the sites where baryonic matter accumulates, cools, and forms galaxies. Understanding the connection between galaxies and their host haloes is central to both astrophysical modeling and cosmological inference.

Observationally, galaxies trace the underlying matter field in a biased and stochastic manner. This bias arises because galaxy properties depend not only on the mass of the host halo but also on its assembly history, environment, and baryonic processes such as feedback. The theoretical challenge is to develop models that can predict galaxy populations, including stellar mass, star formation rate (SFR), metallicity, morphology, and color, based on halo properties.

Modeling the Galaxy–Halo Relationship

Several classes of models have been developed to capture the galaxy–halo connection:

- **Halo Occupation Distribution (HOD)** models[151–158] specify the probability distribution of galaxy number and type as a function of halo mass. These are typically calibrated against clustering measurements.

- **Subhalo Abundance Matching (SHAM)**[159–162] assumes a monotonic mapping between galaxy stellar mass (or luminosity) and a halo property such as maximum circular velocity or mass. SHAM models are simple and flexible but often lack predictive power for galaxy histories.

- **Semi-analytic models (SAMs)**[163–168] solve differential equations describing baryonic processes on top of dark matter merger trees. These include gas cooling, star formation, feedback, and chemical enrichment, and can predict galaxy evolution over time.

- **Hydrodynamical simulations** solve for both dark matter and baryonic physics in a cosmological volume. Simulations such as IllustrisTNG[169], EAGLE[149], and SIMBA[170] provide detailed, self-consistent predictions of galaxy properties and morphologies. Hydrodynamical simulations were used to study and help building the model of the halo occupations[171–177], understand the semi-analytic models[178], and test the assumptions of the subhalo abundance matching approach[176;179;180].

Each of these approaches involves trade-offs between physical realism, computational cost, and interpretability. While hydrodynamical simulations offer the most detailed predictions, they are expensive to run and often limited in volume or resolution. SAMs offer greater speed and tunability but rely on simplifying assumptions. HOD and SHAM methods are efficient but lack explicit time evolution.

The Emergence of Machine Learning Models

In recent years, machine learning (ML) methods have emerged as a promising alternative for modeling galaxy formation. These models aim to learn a mapping from dark matter halo properties—such as mass accretion history, environment, and merger history—to galaxy observables. When trained on high-fidelity simulations, ML models can generate large mock galaxy catalogs at minimal computational cost[181;182].

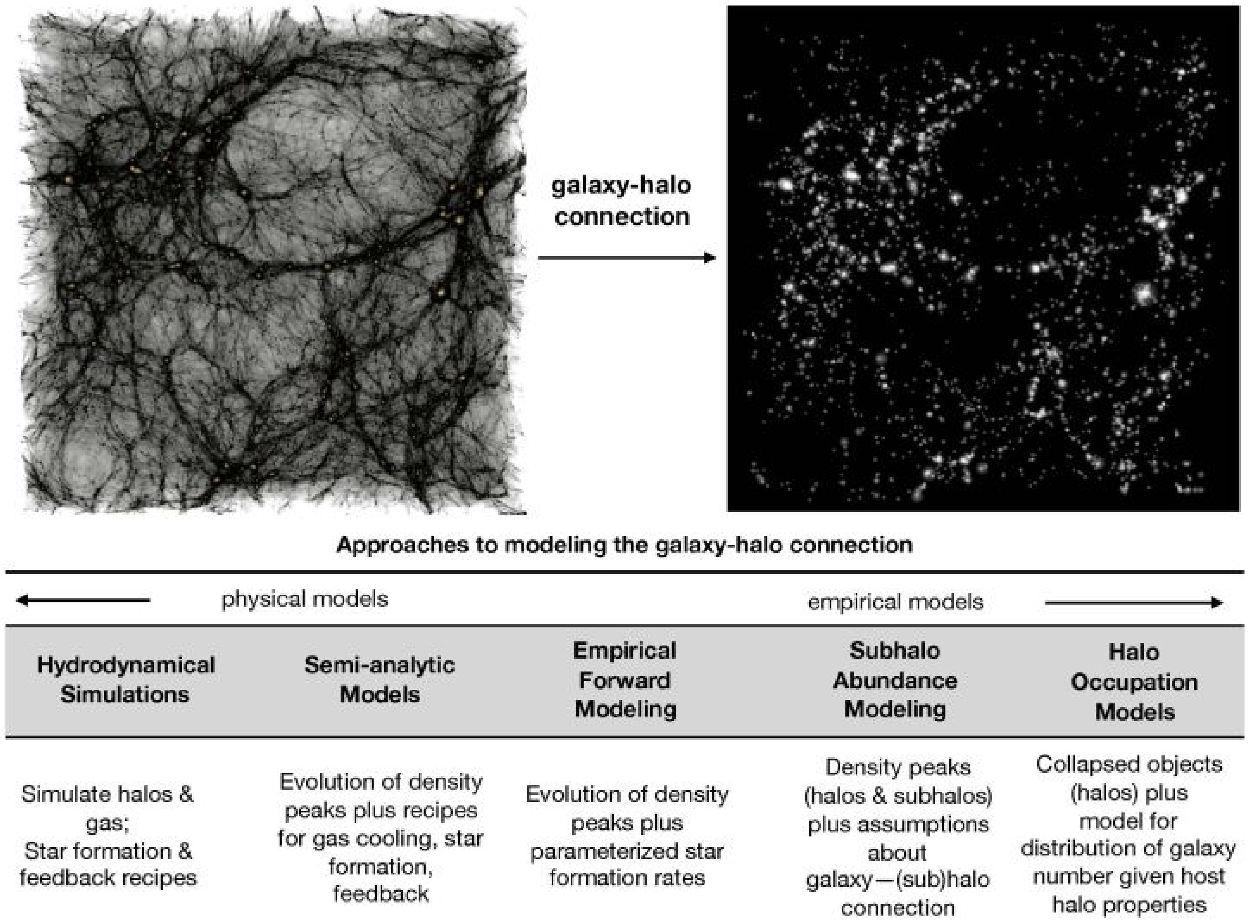

Figure 3.2.1: Illustration of various modeling approaches across physical scale and computational cost. Hydrodynamical simulations are physically rich but computationally expensive. SAMs, HOD, SHAM, and machine learning-based models provide efficient alternatives at varying levels of complexity and predictive power.

A key advantage of ML is its flexibility: it can capture non-linear, high-dimensional relationships that are difficult to encode in analytical prescriptions. However, this comes at the cost of interpretability and generalizability. Moreover, most ML models trained on simulations struggle to reproduce the full diversity and stochasticity of galaxy formation histories, especially at short timescales.

Recent efforts have also explored ML-driven emulators for constraining halo occupation parameters using clustering statistics[182–185]. These approaches demonstrate that ML can significantly accelerate inference pipelines, although care must be taken to avoid bias introduced by model training and parameter space limitations[185].

In this context, hybrid models that combine the predictive power of ML with physically or statistically motivated corrections are gaining attention. The following sections examine in more detail the structure of galaxy formation histories, their variability across scales, and why modeling this variability is essential for accurate mock catalogs and scientific inference.

3.3 Galaxy Formation Histories and Their Variability

The evolution of a galaxy is encoded in its formation history, a time-resolved record of its star formation rate (SFR), chemical enrichment, and stellar mass assembly. Among these, the star formation history (SFH) and metallicity history (ZH) are particularly important, as they determine the galaxy's spectral energy distribution (SED), color, and emission line properties. Recovering accurate galaxy histories is therefore critical for understanding both galaxy formation physics and for producing realistic mock observables.

Star Formation and Metallicity Histories

In a cosmological setting, star formation is driven by the accretion of gas into dark matter haloes and is regulated by internal and external feedback processes. The SFH of a galaxy captures the rate at which gas is converted into stars as a function of cosmic time or redshift:

$$\mathrm{SFH}(t) = \frac{dM_\star(t)}{dt}, \tag{3.1}$$

where $M_\star(t)$ is the stellar mass of the galaxy at time t. Similarly, the metallicity history traces the evolving abundance of heavy elements in the stellar population or the interstellar medium (ISM), often expressed as:

$$\mathrm{ZH}(t) = Z_\star(t), \quad \text{or} \quad Z_{\mathrm{gas}}(t), \tag{3.2}$$

depending on whether stellar or gas-phase metallicity is being tracked. These histories are not smooth but exhibit significant structure due to stochastic processes such as mergers, starbursts, gas accretion episodes, and feedback-driven quenching.

Significance for Observables

The integrated properties of a galaxy, including color, luminosity, and spectral shape, are direct consequences of its formation history. For example, even small bursts of recent star formation can significantly alter a galaxy's UV-optical color, while metallicity variations affect line strengths and continuum shape. In this sense, accurate SFHs and ZHs are necessary not only for theoretical modeling but also for generating forward-modeled observables such as photometry, spectra, and emission lines[186;187].

Moreover, short-timescale variations in SFH can lead to significant scatter in observable quantities. These variations are especially important in dwarf galaxies and starburst systems, and are increasingly relevant for interpreting JWST, LSST, and DESI observations. Models that average over such fluctuations risk underestimating the diversity and stochasticity of galaxy populations.

Sources of Variability

Variability in SFHs and ZHs arises from both physical and numerical factors:

- **Physical sources**: stochastic gas accretion, feedback loops (e.g. supernova- or AGN-driven), galaxy interactions, mergers, and environmental effects.
- **Numerical artifacts**: limited resolution in time or mass, smoothing from time-averaging in simulations or models, or limitations in ML architecture.

For instance, in hydrodynamical simulations like IllustrisTNG, individual galaxies exhibit rich and noisy SFHs (and ZHs) with variability on 100 Myr timescales and below (Figures 3.4.1, 4.4.1, 4.B.1). These fluctuations are not random noise but contain real information about the internal dynamics of galaxies and their cosmological environment.

Implications for Modeling

Most current modeling approaches, including classical semi-analytic models and neural networks, tend to produce smoothed or averaged histories, especially when trained to minimize global loss functions. While this may capture large-scale trends, it suppresses physically important small-scale variability.

As we demonstrate in the next section, this lack of high-frequency structure can be diagnosed and quantified using the power spectrum of the SFH or ZH time series. This motivates the development of hybrid models that combine deterministic accuracy on large scales with stochastic corrections on small scales.

3.4 Machine Learning for Galaxy Evolution Modeling

The increasing complexity of simulations and the sheer volume of data from cosmological surveys have made machine learning (ML) a natural tool for modeling galaxy formation. By learning non-linear mappings from halo properties to galaxy observables, ML offers a fast, scalable, and flexible alternative to semi-analytic or physically motivated models.

Motivation and Use Cases

ML has been applied in various astrophysical contexts, from classifying galaxy morphologies to emulating cosmological simulations. In the context of galaxy formation, ML models aim to predict stellar mass, SFR, metallicity, and even full SFHs and ZHs from halo merger trees or time-series data.

Such models are typically trained on outputs from hydrodynamical simulations like IllustrisTNG, where the ground truth includes both halo and galaxy properties. Once trained, the ML model can be applied to halo catalogs from dark matter-only simulations, enabling the generation of mock galaxy populations across vast volumes at negligible cost[6;181;182].

Model Architectures

Various ML architectures have been explored for this task:

- **Fully connected networks** (dense MLPs): map time-aggregated halo features to integrated galaxy properties.
- **Recurrent neural networks (RNNs)**: process halo histories as sequences, capturing temporal evolution of accretion and environment[6].
- **Convolutional and graph networks**: learn hierarchical or topological features, especially in spatial clustering tasks.

Among these, RNN-based models have shown success in predicting SFH/ZH time series by ingesting temporally resolved halo properties (e.g. mass, environment, merger events). However, while these models accurately reproduce large-scale trends, they consistently under-predict short-timescale variability, a key limitation for generating realistic observables.

Fourier Analysis of Predicted Histories

This discrepancy becomes more apparent in Fourier space. When comparing the power spectrum of predicted vs. true SFHs, ML models reproduce low-k modes (long timescale trends) but systematically underestimate power at high k, corresponding to short timescale fluctuations. This indicates a loss of phase and amplitude information on small temporal scales (Figures 4.2.2, 4.2.3, 4.4.2).

This smoothing effect is due in part to the training objective: loss functions like mean squared error naturally prioritize large-scale features that dominate the variance. It is also a result of the architectural bias toward smooth functions in most neural networks. As a consequence, predicted SFHs and ZHs often appear as denoised versions of the true histories, accurate in integrated quantities but lacking in temporal fidelity.

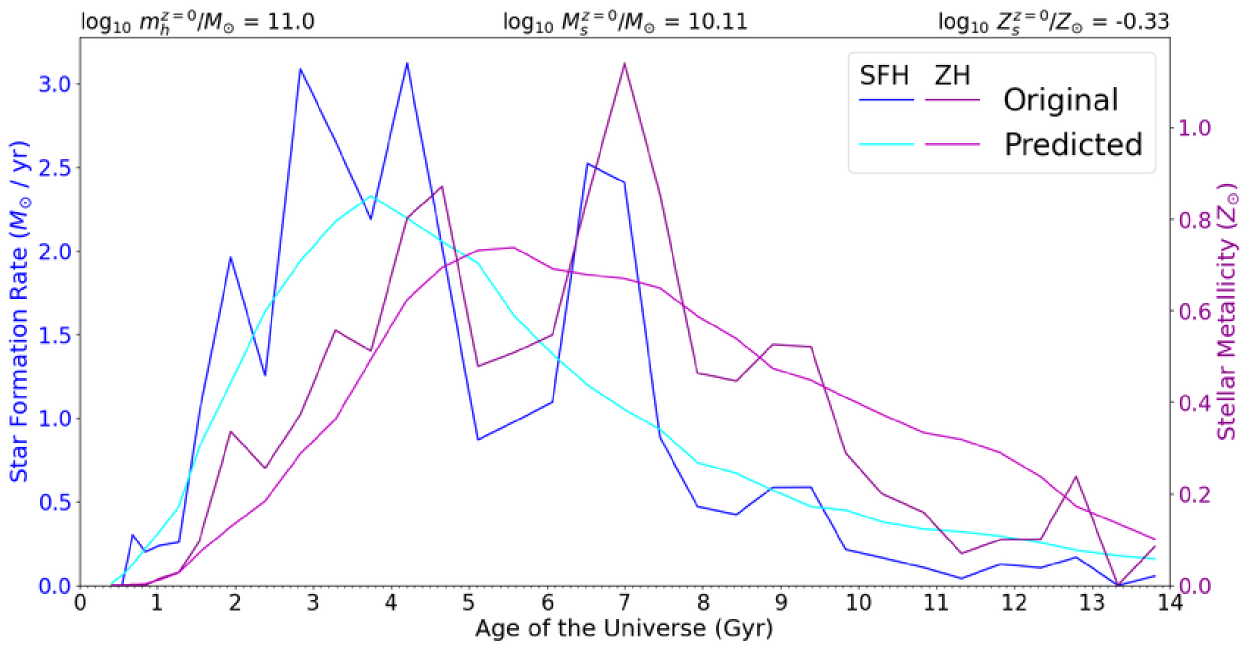

Figure 3.4.1: An example formation history for a satellite galaxy of intermediate mass in IllustrisTNG simulation. The true star formation history (blue) and the corresponding prediction (cyan) are shown alongside the true stellar metallicity history (purple) and the predicted metallicity history (magenta). Numerical estimates of the subhalo and predicted stellar mass and metallicity are shown above the graph. The neural network model modestly captures the overall shape of SFH and ZH, but underestimates high-frequency fluctuations present in the true (simulated) histories. Source : Chittenden and Tojeiro[6].

Toward Hybrid Stochastic Models

To address this, we require modeling strategies that preserve the strength of ML on large-scale predictions while restoring the short-timescale structure lost in training. This leads naturally to hybrid models that augment neural network predictions with statistically principled stochastic corrections, a concept developed in the following chapter.

3.5 Motivation for Semi-Stochastic Corrections

The limitations of deterministic machine learning models in recovering short-timescale variability in galaxy formation histories motivate the need for hybrid approaches. These limitations are not merely cosmetic; they directly affect the realism of synthetic galaxy populations and the observables derived from them, including spectral energy distributions (SEDs), color distributions, and star formation rates.

As discussed in the previous section, most neural networks trained to predict star formation and metallicity histories learn smoothed representations that match low-frequency trends but systematically suppress high-frequency variability. This results in time series that are accurate in an average sense but do not reflect the full stochastic richness observed in high-resolution simulations like IllustrisTNG.

Correcting for Lost Variability

A promising approach is to treat this discrepancy as a scale-dependent modeling failure: ML models succeed on long timescales but underperform on short ones. This perspective allows for a natural correction mechanism, retain the deterministic neural network prediction as the large-scale backbone, and inject small-scale fluctuations in a way that reproduces the true statistical properties of the underlying simulation.

This motivates a *semi-stochastic correction* strategy. The key idea is to compute the power spectrum of each predicted SFH/ZH and compare it to that of real simulated histories. Where power is lacking, typically at high frequencies, statistically consistent perturbations are introduced to restore the correct amplitude of fluctuations. Phase information is randomized under the assumption that short-timescale features are not strongly correlated across halos of similar mass or formation history.

Design Goals and Justification

This correction method is built around the following goals:

- **Accuracy**: preserve the fidelity of neural network predictions on resolved scales.
- **Realism**: match the statistical structure (power spectrum) of the simulated data.
- **Generality**: apply across diverse halo masses, environments, and cosmic times.
- **Efficiency**: computationally inexpensive and easily integrated into existing pipelines.

From a cosmological perspective, capturing short-timescale fluctuations is essential for generating realistic mock catalogs. These catalogs are used in forward modeling pipelines for interpreting galaxy colors, star formation rates, and emission lines, all of which are sensitive to SFH variability.

Overview of the Next Chapters

The next two chapters explore the use of machine learning for modeling galaxy formation histories and the development of correction strategies to improve their physical realism.

Chapter 4 introduces a semi-stochastic correction framework designed to recover short-timescale variability that is typically lost in standard neural network predictions. This method is applied to galaxies from the IllustrisTNG simulation and shown to enhance the fidelity of predicted star formation and chemical enrichment histories. The corrections lead to improved agreement with simulation-based observables such as the stellar-to-halo mass relation, spectral energy distributions, and photometric color distributions, particularly in recovering bimodality and scatter across galaxy populations.

Chapter 5 extends this approach to predictions made from pure dark matter simulations, including TNG-Dark and Uchuu, where baryonic information is absent from the training data. The predictive performance is evaluated across different halo properties, such as mass accretion history, half-mass radius, and cosmic environment. The semi-stochastic corrections improve the match to hydrodynamical results in terms of star formation rates, metallicities, and observable properties like spectra and colors, thereby demonstrating the framework's utility in generating realistic galaxy catalogs from dark matter-only simulations.

“ଅଦୃଶ୍ୟ ଆକୃତିରେ ନିହିତ ଥାଏ ଦୃଶ୍ୟ ଶୃଙ୍ଖଳା — ଯେଉଁଠି ଗୁରୁତ୍ୱ ଓ ଦୈବତା ବିକାଶ ପାଏ।”

— Inspired by Upanishadic cosmology

“The invisible form holds the visible sequence — where gravity and divinity unfold.”

Chapter 4

Optimised neural network predictions of galaxy formation histories using semi-stochastic corrections

This chapter originally appeared in the literature as :

Jayashree Behera, Rita Tojeiro and Harry Chittenden, Optimised neural network predictions of galaxy formation histories using semi-stochastic corrections, arXiv:2409.16548[188].

Abstract

We present a novel methodology to improve predictions of galaxy formation histories by incorporating semi-stochastic corrections to account for short-timescale variability. Our paper addresses limitations in existing models that capture broad trends in galaxy evolution, but fail to reproduce the bursty nature of star formation and chemical enrichment, resulting in inaccurate predictions of key observables such as stellar masses, optical spectra, and colour distributions. We introduce a simple technique to add a stochastic components by utilizing the power spectra of galaxy formation histories. We justify our stochastic approach by studying the correlation between the phases of the halo mass assembly and star-formation histories

in the IllustrisTNG simulation, and we find that they are correlated only on timescales longer than 6 Gyr, with a strong dependence on galaxy type. We demonstrate our approach by applying our methodology to the predictions on a neural network trained on hydrodynamical simulations, which failed to recover the high-frequency components of star-formation and chemical enrichment histories. Our methodology successfully recovers realistic variability in galaxy properties at short timescales. It significantly improves the accuracy of predicted stellar masses, metallicities, spectra, and colour distributions and provides a practical framework for generating large, realistic mock galaxy catalogs, while also enhancing our understanding of the complex interplay between galaxy evolution and dark matter halo assembly.

4.1 Introduction

According to the current model of galaxy evolution and structure formation, galaxies are formed by the contraction of baryonic gas that is gravitationally bound to a dark matter halo. The evolution of a galaxy is therefore expected to be at least partly determined by the properties and evolution of its halo and surrounding environment. Although this relationship has not yet been fully understood, studies in simulations and observations point to strong links between the two. For example, in cosmological hydrodynamical simulations, the halo formation history has been shown to impact on a galaxy's present-day stellar mass (e.g. Alarcon et al. [189]; Cui et al. [190]), star-formation rate or colour (e.g. Montero-Dorta et al. [191]), circumgalactic medium[192], and morphology (e.g. Davies et al. [192]), at fixed stellar or halo mass or using controlled experiments. Although there is broad agreement that halo growth and galaxy growth are linked, we still lack clarity in the details with different simulations and observations arguing for different — sometimes directly opposing — effects of how halo assembly histories impact galaxy properties. The difficulty comes from the complexity of this galaxy-halo connection and its dependence on physical processes that span decades in physical and time scales, as well as the interaction between baryons and dark matter. In observations, we remain limited by the fact that halo properties — and halo assembly in particular — remain notoriously hard to measure.

Recent machine learning techniques have shown promise in emulating complex astrophysical processes that connect the properties of dark matter halos to those of the galaxies they host. These efforts have two broad goals: one is to enable the creation of much larger mocks with realistic galaxy populations than what is possible with hydrodynamical simulations, in a fast and efficient way; and another is to study the impact of different halo properties on the build-up of the galaxy-halo connection. And, indeed, such efforts have been shown to offer a way to efficiently populate large N-body simulations with mock galaxy catalogs while encapsulating key relationships between halos and galaxies[182;193–195].

In a recent study,[6] (CT23 henceforth) developed neural networks to predict the star formation and chemical enrichment histories of central and satellite galaxies based solely on the properties of their dark matter halos and environments. The networks were trained on hydrodynamical simulations from the IllustrisTNG project[143–146;196] and successfully recovered key observational benchmarks like the stellar mass-halo mass relation, mass-metallicity relation, and colour bimodality. The model utilizes a semi-recurrent neural network (rNN) algorithm to predict the time-resolved star formation history (SFH) and metallicity (or chemical enrichmeny) history (ZH) of central and satellite galaxies from the historical evolution of their dark matter halos and local dark matter environment. From these properties, one can then self-consistently predict observables such as optical spectra and broadband colours. The model is fairly practical and can be used in high fidelity N-body simulations, statistically reproducing key relationships of the galaxy-halo connection, and producing realistic mock surveys of unprecedented size and complexity[197]. While the CT23 networks broadly matched time-averaged trends, the predicted histories lacked variability on short timescales compared to the original hydrodynamic histories. By correlating the lack of power on short timescales of the formation histories with residuals in predicted vs. true galaxy properties, CT23 concluded that the lack of variablity on short timescales led to a systematic underestimation of stellar masses, luminosities, and emission line strengths for some galaxies. They also noted a systematic underestimation of the scatter in fundamental relations such as the mass-metallicity relation. The discrepancy was more significant for star-forming central galaxies than for quenched satellites.

Modelling and measuring short-timescale variability in star-formation histories and its connection to fundamental relations is an active area of research. Studies on cosmological hydrodynamical simulations have linked long timescales ($\gtrsim 1$ Gyr) with halo-related processes (e.g. accretion) and shorter timescales with variability associated with feedback processes (e.g. Sparre et al.[198]; Iyer et al.[199]). The lack of a direct physical link between halo and star-formation histories on short timescales then becomes a natural and tempting explanation for the shortcomings of the networks in CT23. Purely stochastic methods to model the star-formation histories of galaxies have been successful at describing the scatter of the star-forming main sequence[200] and identifying different physical contributions to variability on different timescales by looking at different simulations of galaxy evolution (e.g. Iyer et al.[199]). Within these frameworks, for example, it is possible to explicitly model variability arising from different physical processes, such as gas inflow rates and the formation and destruction of giant molecular clouds, as well as their impact on observables, laying the ground for observational constraints on these models (e.g. Tacchella et al.[201]; Iyer et al.[202]).

In this follow-up paper to CT23, we implement a stochastic correction method to account for unmodelled short-timescale variability in the predicted star formation and metallicity histories. We have two main goals. One is to directly improve the predictions of the CT23 networks. Along with predicting the histories directly, we also train auxiliary neural networks to *directly* predict their power spectra as a function of halo properties. The power spectrum captures the degree of variability on different timescales, which can be attributed to physical drivers like stellar feedback, fluctuations in gas accretion, and mergers. This method utilizes randomized realizations of the histories by inverse Fourier transforming the predicted power spectra with quasi-random phases. This stochastic approach self-consistently introduces variability on short timescales while preserving the large-scale trends predicted by the original networks. In this paper we limit our corrections to timescales longer than 860 Myr, as that was the limiting frequency of the star-formation and chemical enrichment histories used by the neural network. Future work will study the connection between halos and baryons using a larger range of frequencies. Our other goal is to understand the connection between halo assembly, star-formation and chemical enrichment histories by looking at the correlations in phases of these two components. Effectively, we ask the question: *on what timescales can the phases of star-formation and chemical enrichment histories be treated as purely stochastic, given the mass assembly history of a host halo?*

The added variability introduced by our quasi-stochastic corrections produces stellar masses, spectra and metallicities that show substantially better agreement with the reference hydrodynamic simulation, confirming that the missing short-timescale variability was a primary cause of the original discrepancies.

Our paper is organised as follows. In Section 4.2 we summarise the CT23 model and data, as well as the Fourier transforms data used here. In Section 4.3 we explore the phase correlations between star-formation, metallicity and halo mass assembly histories with the goal of informing the limits of our stochastic corrections, and we present our formalism and the details of their implementation. In Section 4.4 we present our results and finally in Section 4.5 we discuss our results and conclude.

4.2 Model and Data

We use the same SFHs, ZHs and mass assembly histories (MAHs) of the dark matter halos as in CT23, computed from the Illustris TNG simulation. SFHs and ZHs are defined from the mass-weighted age and mass-weighted metallicity distributions of star particles bound to each subhalo at $z = 0$, respectively. These properties are then binned on to a time grid that corresponds to every third snapshot of the TNG simulations, spanning a redshift range of $0 \leq z \leq 20$. The final grid therefore has 33 bins, each corresponding to a lookback time interval of around 200 to 400 Myr. For full details on the neural network model, targets, features, and predictions, see CT23. Briefly, a combination of temporal features (such as the halo mass accretion rate history) and non-temporal features (such as the $z = 0$ halo mass or infall velocity for satellites) are fed to a semi-recurrent neural network to predict time-dependent star-formation and chemical enrichment histories of galaxies in the same time grid. To increase the sampling of high-mass objects, CT23 combined data from TNG100 and TNG300 after correcting SFHs and ZHs for mass resolution effects.

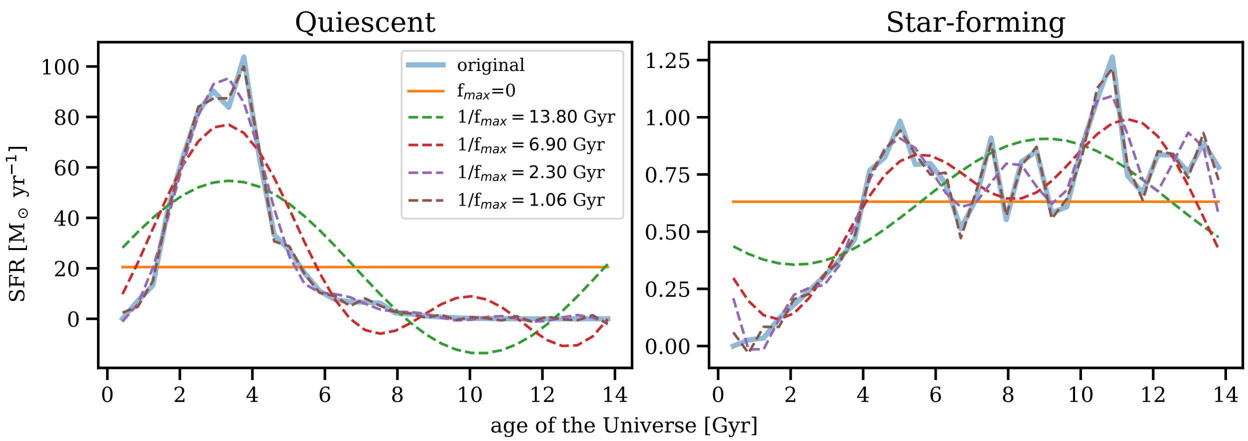

Figure 4.2.1: This figure demonstrates the decomposition of the star-formation history of two galaxies onto the frequency components of the DFT. In each panel, the thick blue line shows the original star formation history. The other lines show the reconstructed star-formation history using the frequencies up to (and including) the frequency listed on the label — we only use four frequencies here for clarity. The first component of the DFT is simply the mean the star-formation history over cosmic time, represented by the orange solid line. The left panel shows a typically quiescent high-mass galaxy ($M_* = 10^{11.2} M_\odot$) whereas the right panel shows a typically star-forming low-mass galaxy ($M_* = 10^{9.7} M_\odot$).

For the neural-network modeling, the TNG data is split into training and test datasets. The training set consists of 75 percent of the combined TNG100 and resolution-corrected TNG300 data. Separate training sets exist for central and satellite galaxy neural networks. 20 percent of the training data is held out for validation during training iterations. The rest of the 25 percent of the TNG galaxy data is reserved strictly for testing and is not used to update model parameters during training. Once trained, the neural networks are used to predict SFHs and ZHs from the halo properties of the test set. The predicted galaxy properties are compared with the true ones to evaluate to quantify model performance and check for overfitting. CT23 demonstrates that, on the whole, the recurrent neural network successfully recovers key properties of the galaxy population, namely the mean stellar to halo mass relation (SHMR, CT23-Fig 11) and the dependence of its scatter on halo assembly (CT23-Fig 15), stellar mass-magnitude relations (CT23-Fig 20), the relationship between galaxy SFH and halo mass (CT23-Fig 12) and its dependence of the stellar mass-weighted age with stellar mass (CT23-Fig 14). Not withstanding these key successes, the neural network fails to recover the scatter in the mass-metallity relation (CT23-Fig 16), the scatter in galaxies'

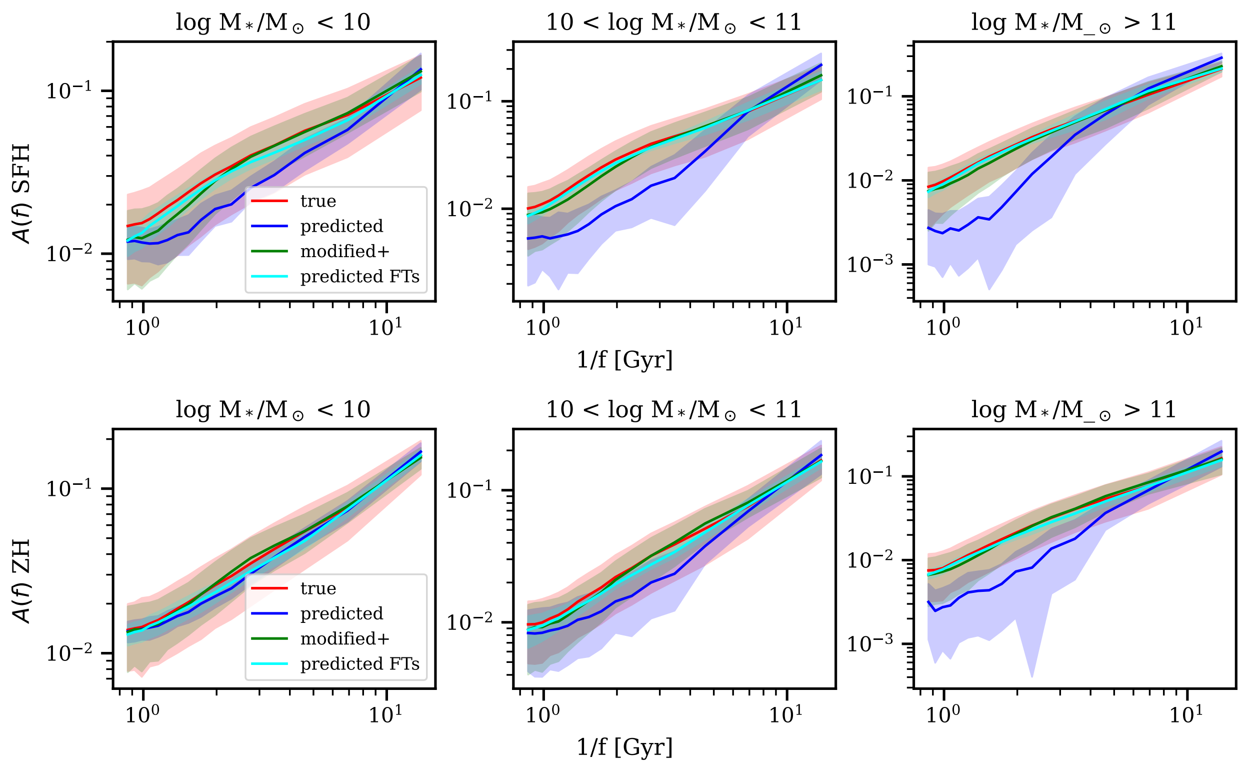

Figure 4.2.2: The mean amplitude of the DFT, $A(f)$, for the star-formation histories (top) and metallicity histories (bottom) of different samples of central galaxies. From left to right the different panels represent galaxies of low, intermediate, and high stellar mass, respectively. The different coloured lines show the mean of the true (red), predicted (blue), and modified+ (green) values. The cyan line shows the mean of the FFT amplitudes that were *directly* predicted by the rNN. The difference between the red and blue lines indicates the initial problem that this work aims to solve — the lack of power at small scales in the star-formation and metallicity histories predicted by the rNN (see Section 4.2). The similarity between the cyan and the red line motivates the approach proposed here — to use the predicted DFT amplitudes as an estimate of the true amplitudes (see Section 4.3). The difference between the green and red lines indicates that the methodology presented here is successful at recovering the power on all timescales (see Section 4.4). In all cases, the shaded areas show the standard deviation around the mean.

spectral energy distributions (SEDs, CT23-Fig17), or H_α luminosity. The network also clearly fails to predict SFH and ZH features on short time-scales (CT23-Fig 13).

To analyse and mitigate this shortcoming of the rNN, we begin by comparing the amplitude of the Fourier transform (FT) of the true and predicted SF and ZHs. To do that, we compute the FT of each galaxy's SFH and ZH (and, later, MAH) using a NumPy[i] implementation of the discrete Fourier transform (DFT[ii]), allowing us to compare predicted and true histories as a function of time-scale. We will represent the FTs of time-domain function $h(t)$ as complex functions $H(f) = A(f)e^{-i\theta(f)}$, where f denotes frequency, and $A(f)$ and $\theta(f)$ are the amplitude and phase as functions of frequency, respectively. For each of our functions sampled in 33 bins in time (SFH, ZH, MAH), we obtain a DFT with 17 amplitudes and phases. The zeroth array element corresponds to the average function value, with phase $\theta = 0$ by default. Each of the other 16 elements i contains the DFT at a frequency that is $1/i^{th}$ of the period of the function, which is assumed to be age of the Universe at $z = 0$ in the TNG cosmology. In other words, the frequency corresponding the $i = 1$ frequency corresponds to a mode with a period $P = 13.80$ Gyr, $i = 2$ to a mode with a period of $P = 6.90$ Gyr and so forth. The highest sampled frequency is 1.16 Gyr^{-1} equivalent to a timescale of 0.86 Gyr.

For illustration purposes, we show in Fig. 4.2.1 the reconstruction of the true star-formation history of two galaxies in our training sample as a function of the maximum frequency (or, inversely, minimum timescale) used. Fig. 4.2.2 shows the average amplitude A(f) of the SFHs and ZHs in three stellar mass bins that we will call "low mass" ($\log M_*/M_\odot < 10$), "intermediate mass" ($10 \leq \log M_*/M_\odot < 11$), and "high mass" ($\log M_*/M_\odot \geq 11$). The magnitude of $A(f)$ depends on the stellar mass of the galaxy, so amplitudes are normalised to unity for each galaxy prior to averaging, which allows us to compare the shape of A(f) for the true (red) and predicted (blue) SFH and ZHs. In the case of the SFHs (top) we can see the clear deficit in power of the SFHs predicted by the CT23 rNN (blue line) on timescales shorter than approximately 5 Gyr, particularly at high stellar mass. In the case of the ZHs (bottom) this deficit is not seen in the lowest stellar mass bin.

[i] https://numpy.org
[ii] https://numpy.org/doc/stable/reference/routines.fft.html

We can also analyse how the phases of the SFH and ZHs predicted by the rNN compare to the true ones. Fig. 4.2.3 shows the predicted vs true phases the SFHs of central galaxies in the test sample, at three different frequencies and in three mass bins (the resulting figure for satellites is similar and not shown). Although the SFHs predicted by the rNN have phases that are well predicted on the lowest frequency, the correlation quickly deteriorates with increasing frequency, and phases for frequencies higher than $1/2.76$ Gyr^{-1} are uncorrelated across true and predicted SFHs. We also note that the distribution of $\theta(f)$ is not uniform in $[0, 2\pi]$, even for high frequencies — in other words, the range of allowed values of θ is determined by galaxy evolution processes, and carries physical meaning. This can also be seen through the dependence on stellar mass, in that the phases of the more massive galaxies are both shifted with respect to those of lower mass galaxies (reflecting the earlier epoch of star formation) and are slightly better predicted by the rNN, particularly towards lower frequencies. We will investigate some of the reasons for this in Section 4.3.2. For now, we point out that the peak of the SFH or ZH of a galaxy is largely determined by the phase of the longest timescale mode (see e.g. Fig. 4.2.1) and that these are well recovered by the rNN at all stellar masses. However, the limitations towards higher frequencies are evident. In this paper we propose a solution to this problem by modifying the SFHs and ZHs predicted by the rNN such that they have the similar FTs to the true SFHs, and we will refer to those as ”modified/modified+”, depending on the details of the modification. We will refer to quantities directly obtained from the TNG training samples as ”true” and to the quantities predicted by the rNN as ”predicted” (matching those in CT23).

This line of investigation is possible because the rNN is able to predict *directly* the shape of A(f) very well (i.e., A(f) is explicitly made a target of the NN), although with a lower scatter than those corresponding to the true SFHs — see cyan line in Fig. 4.2.2. In the next section we introduce a formalism to correct the SFHs and ZHs in the case that the true amplitude of the FT is known (i.e., assumed to be correctly predicted by the rNN) and, in Section 4.3.2, we consider how to treat the phases in detail.

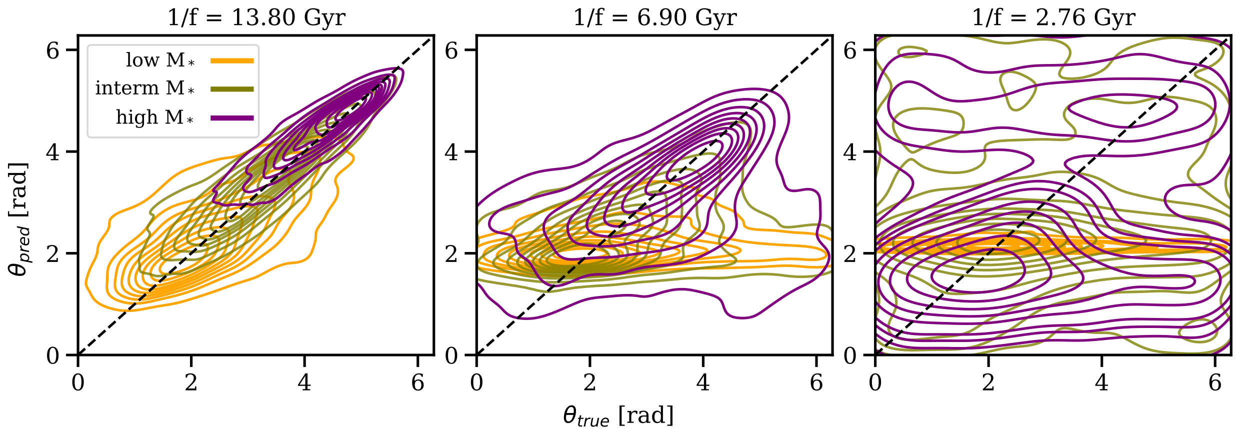

Figure 4.2.3: The phases of the predicted star-formation histories vs the phases of the true star-formation histories of central galaxies shown at three timescales: 13.8 Gyr (left), 6.90 Gyr (middle), and 2.76 Gyr (right).Each panel shows a kernel-density estimation visualisation of the phases for galaxies in the three mass bins, with contours representing levels of equal probability density, in 10 equal bins
. The black dashed line is the one-to-one line. As we move to shorter timescales and to lower stellar masses, the correlations between true and predicted phases breaks down.

4.3 Semi-Stochastic corrections

4.3.1 Formalism

We aim to modify the predicted star formation and chemical enrichment (metallicity) histories from our neural networks to better match the intrinsic variability of the original ones in the hydrodynamic simulations. We assume an additive relationship between the true and predicted histories and their respective FTs:

$$h_{\text{true}}(t) = h_{\text{pred}}(t) + h_{\text{stoc}}(t) \tag{4.1}$$

$$H_{\text{true}}(f) = H_{\text{pred}}(f) + H_{\text{stoc}}(f) \tag{4.2}$$

where $h_{\text{true}}(t)$ and $h_{\text{pred}}(t)$ are the true and predicted histories in real time, and $h_{\text{stoc}}(t)$ is the correction term we seek to compute.We denote the FT of the signal in the time domain $h_j(t)$ as $H_j(f) = \mathcal{F}[h_j(t)]$, where $\mathcal{F}$ is the FT operator and $\mathcal{F}^{-1}$ its inverse. Each component in frequency space can be expressed in complex form as:

$$H_j(f) = A_j e^{-i\theta_j} = A_j \cos\theta_j - - - iA_j \sin\theta_j, \quad j \in \{\text{true}, \text{pred}, \text{stoc}\}. \tag{4.3}$$

Substituting into equation 4.2 and equating the real and imaginary parts, we obtain:

$$\begin{aligned} A_{\text{true}} \cos\theta_{\text{true}} &= A_{\text{pred}} \cos\theta_{\text{pred}} + A_{\text{stoc}} \cos\theta_{\text{stoc}}, \\ A_{\text{true}} \sin\theta_{\text{true}} &= A_{\text{pred}} \sin\theta_{\text{pred}} + A_{\text{stoc}} \sin\theta_{\text{stoc}}. \end{aligned} \tag{4.4}$$

Solving these equations for A_{stoc} and θ_{stoc} gives:

$$\begin{aligned} A_{\text{stoc}} &= \Big[(A_{\text{true}} \cos\theta_{\text{true}} - A_{\text{pred}} \cos\theta_{\text{pred}})^2 \\ &\qquad + (A_{\text{true}} \sin\theta_{\text{true}} - A_{\text{pred}} \sin\theta_{\text{pred}})^2\Big]^{1/2}, \\ \theta_{\text{stoc}} &= \tan^{-1}\left(\frac{A_{\text{true}} \sin\theta_{\text{true}} - A_{\text{pred}} \sin\theta_{\text{pred}}}{A_{\text{true}} \cos\theta_{\text{true}} - A_{\text{pred}} \cos\theta_{\text{pred}}}\right). \end{aligned} \tag{4.5}$$

The amplitude, A_{stoc} represents the misfit between true and predicted transforms while the phase θ_{stoc} is randomized. This defines the complete complex correction term $H_{\text{stoc}}(f)$, that introduces missing stochasticity while preserving the predicted large-scale trends. We then obtain the correction in real space by taking the inverse FT:

$$h_{\text{stoc}}(t) = \mathcal{F}^{-1}[H_{\text{stoc}}(f)]. \tag{4.6}$$

Finally, the corrected star formation or enrichment history is computed by adding this term to the predicted one, as in equation 4.1.

To determine the true amplitudes A_{true} and true phases θ_{true} we make the key assumption that the neural network accurately predicts the FT amplitudes (see cyan and red lines in Fig. 4.2.2), which govern the distribution of variability timescales. Under this assumption, we substitute the FT amplitudes directly predicted by the neural network as the true amplitudes. Consequently, the only remaining parameter to be determined is θ_{true}.

4.3.2 Phase correlations and stochasticity

A purely stochastic correction would assume random θ_{true} for each frequency of the star-formation and chemical enrichment histories. However, given the relationship between halo mass assembly and star formation, for example, we expect that the phases of these two functions are correlated to some extent, even considering a time-delay (which would translate onto a shift in phase between the two). In this section, we investigate the correlations between phases of the MAH, SFH and ZH of different populations of galaxies in the training sample with the goal of understanding the appropriate limits of our stochastic corrections. We define three populations of galaxies according to their position in the star-formation main sequence, as shown in Fig. 4.3.1. The star formation rate (SFR) is computed by averaging the stellar mass in the last 1 Gyr. We split the galaxy population into star-forming and quiescent, and further split off the high-mass end of the star-forming main sequence — these galaxies sit below the main sequence, and we will refer to them as "quiescing", as they are potentially quenching now. We also define a distinct set of three populations according only to their stellar mass, the boundaries of which are also shown in Fig. 4.3.1 and are the same boundaries used in Fig. 4.2.3.

To investigate to what extent the phases of the different histories are correlated, we compute the Spearman Rank correlation coefficient, r_s, between MAH and SFH, SFH and ZH, and MAH and ZH, as a function of frequency, for each galaxy sample and separately for centrals and satellites (Fig. 4.3.2). In almost all cases, we see that the phases of the MAH and SFHs are correlated only on timescales longer than at least 6 Gyr. This likely contributes to the fact that the rNN only captures the nature of SFHs on long timescales — the rNN learns the connection between halo and stellar assembly, but Fig. 4.3.2 shows that these are uncorrelated on timescales shorter than 6 Gyr. More interestingly, Fig. 4.3.2 also reveals a dependency of the galaxy-halo connection on the type of galaxy. Specifically, galaxies in the lower-mass end of star-forming main sequence have the strongest correlation between the phases of their halo and stellar assemblies. These galaxies, sitting in the linear part of the star-forming main sequence, are often described as gas-regulated systems[203;204],

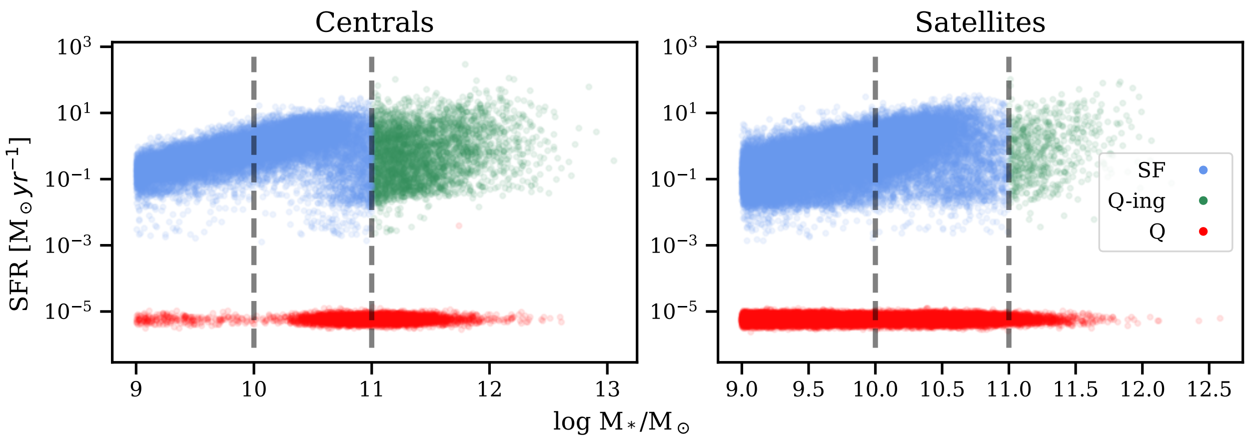

Figure 4.3.1: The star-formation rate vs stellar mass of centrals (left) and satellite (right) galaxies in the training sample. Quiescent galaxies have zero SFR in the most recent bin, spanning 475 Myrs, but they are plotted here with a dummy value for presentation values. The colours demonstrate how galaxies are classified in this diagram: quiescent, star forming, or high-mass star-forming. Also represented are the three stellar mass bins used, defined as low mass (log $M_*/M_\odot < 10$), intermediate mass ($10 \leq$ log $M_*/M_\odot < 11$), and high mass (log $M_*/M_\odot \geq 11$).

whereby galaxies are forming stars from infalling cold gas at a fixed efficiency. The expected coincidence of infalling cold gas and dark matter onto the halo then gives a plausible explanation as to why the phases of these galaxies' MAH and SFHs show the strongest correlation. At the high mass end of the star-forming sequence, however, the correlation between halo mass assembly and star formation decreases. r_s is very small at *all* timescales for central SF massive galaxies, and effectively zero is we select galaxies based on a SFR averaged over the last 475 Myr, instead of 1 Gyr. In these galaxies, star formation timescales have decoupled from those of mass assembly even on very long timescales. In the case of satellites, we correlate the phases of the star-formation histories with the subhalo assembly histories. The main difference is in the quenched galaxies, although we note a more complex relationship between halo and star-formation phases. Obviously our galaxy samples, defined according to either stellar mass or position in the star-forming main sequence, are not independent. The purpose of this paper is not to carefully disentangle the main drivers of the correlation between MAH, SFH and ZH phases, but rather to understand to what extent and in what conditions they correlate so that we may inform how we determine $\theta_{\rm true}$. Also clear from Fig. 4.3.2 is that the phases of star formation and chemical enrichment are substantially

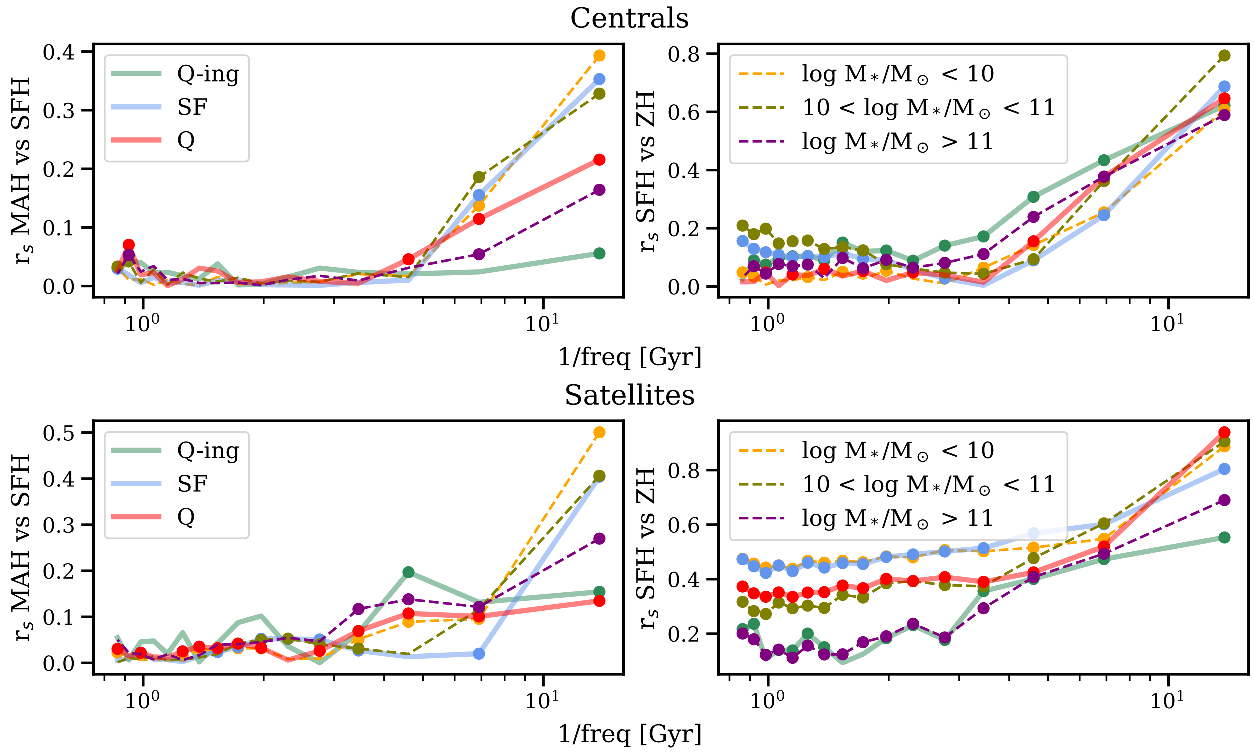

Figure 4.3.2: The Spearman rank correlation coefficient, r_s, between the phases of the halo mass assembly histories and the star formation histories (left) and between the phases of the star formation and chemical enrichment histories (right) as a function of frequency. In the case of satellites (bottom), we considered the subhalo mass assembly histories. Solid lines show r_s of different types of galaxies, as shown in Fig. 4.3.1. Dashed lines show r_s of galaxies of different stellar mass. Statistically significant correlations are identified by a solid symbol. In general we find that halo and stellar growth are most correlated on long timescales ($\gtrsim 6$ Gyr) and for low-to-intermediate stellar mass galaxies in the star-forming main sequence. We also find that star-formation and chemical enrichment are correlated to shorter timescales, with less dependence on galaxy type or stellas mass.

more strongly correlated, for all galaxy types and in particular satellite galaxies.

4.3.3 Implementation

Given the results from the previous section, we decide firstly not to modify the phases of the two longest frequency components as they are well predicted by the network[iii]. We also preserve the amplitudes of the longest component, but only for galaxies with $\log M_*/M_\odot < 9.5$, where predictions remain accurate (see Fig. 4.2.2). Secondly we decide to treat the $\theta_{\rm true}$ for the SFH as effectively stochastic (i.e., with no imposed correlation with the phases of the halo's accretion history) for all other frequencies.

That leaves two important matters to consider: 1) the fact that SFH and ZH phases are effectively not random in the training sample (i.e., they do not uniformly populate the $[0, 2\pi]$ interval); 2) the fact that the phases of SFH and ZHs are strongly correlated.

We address the first by randomly sampling phases from the training sample from galaxies matched in stellar mass, therefore insuring that — although stochastic — the phases are drawn from the same range spanned by the data [iv]. To achieve this, the predicted sample is first divided into 15 bins based on the percentile distribution of its stellar masses. The training sample is then binned according to the same stellar mass bins, and for each bin, a single galaxy is randomly selected from the corresponding training sample bin. This approach ensures that the selected training data spans the same range of stellar masses as the predicted sample while maintaining the stochastic nature of the phase selection. We address the second by drawing phases for SFH and ZH from the same galaxy, thereby maintaining any existing correlations between the two quantities in the training data. The subset of training sample picked this way is termed as $h_{\rm subtrain}(t)$ with the corresponding FT as $H_{\rm subtrain}(f)$, FT phases as $\theta_{\rm subtrain}$ and FT amplitudes as $A_{\rm subtrain}$.

[iii]Fig. 4.2.3 shows that at a timescale of 6.90 Gyr, this statement is no longer true for our lowest mass bin. However, drawing stochastic phases for these galaxies at this frequency makes no significant difference to our results, and we therefore adopt the scheme here for simplicity.

[iv]We tried other matching schemes between test and training data, such as a 2D match on halo and stellar mass, or a halo mass alone scheme, but none provided significant improvement on a match on stellar mass alone.

Therefore, in summary, we employ the equations in Section 4.3.1 with the following substitutions for θ_{true} and A_{true}:

$$\theta_{\text{true}}(f) = \begin{cases} \theta_{\text{pred}}(f) & \text{if } f = 0 \text{ or } 0.03735525 \\ \theta_{\text{subtrain}}(f) & \text{otherwise} \end{cases} \tag{4.7}$$

$$A_{\text{true}}(f) = \begin{cases} A_{\text{pred}}(f) & \text{if } f = 0 \text{ and } \log M_*/M_\odot < 9.5 \\ {}_{\text{pred}}A(f) & \text{otherwise} \end{cases} \tag{4.8}$$

where f is the frequency,$A_{\text{pred}}(f)$ refers to the amplitudes obtained by applying DFT to the predicted time-domain histories, $h_{\text{pred}}(t)$. In contrast, ${}_{\text{pred}}A(f)$ denotes the amplitudes directly predicted by the rNN in Fourier space.

This method brings substantial improvement to the SFHs and ZHs predicted by the rNN. However, although the mean of ${}_{\text{pred}}A(f)$ matches the mean of $A_{\text{true}}(f)$, the scatter in ${}_{\text{pred}}A(f)$ is substantially smaller. It is also the case that, in individual galaxies, ${}_{\text{pred}}A(f)$ can have occasional large residuals that, when paired with stochastic values for $\theta_{\text{subtrain}}(f)$, can lead to fluctuations in the SFHs and ZHs that are unphysical. For example, sharp transitions and flat plateaus in the SFHs and ZHs (such as quenching) rely on phase coherence, which our method does not maintain. We deal with these two issues — both stemming from the fact that ${}_{\text{pred}}A(f)$ is not a perfect match to $A_{\text{true}}(f)$ and the lack of phase coherence in our stochastic corrections — in two ways: a set of filters to avoid extreme deviations of the SFHs and ZHs from their predicted values (including negative values), and a normalisation to restore the expected standard deviation of the SFHs and ZHs. We detail these next.

Filters

We apply four post-processing filters to the modified SFH and ZHs. In summary, these are:

1. All negative values resulting from the stochastic modifications are replaced with the rNN-predicted values.

2. At very early times ($t < 2.8$ Gyr), modifications are restricted to a factor of two of the original predictions.

3. Galaxies identified as quenched remain unmodified after quenching.

4. Modifications to central galaxies at recent times ($t > 10.8$ Gyr) are not allowed to fall below 35 per-cent of the rNN predictions.

The justification for (i) is clear as negative values are unphysical.

The reason for (ii) comes from the fact that the rNN does well at early times. The impact on colours and SEDs from this filter is very small.

We justify (iii) because of the need to maintain the quenching of galaxies, which is well predicted by the rNN but clearly endangered by stochastic corrections which can create spurious SF events. We identify potentially quenched galaxies from their specific star formation rate (sSFR) values averaged over the last 1.008 Gyr (by requiring it to be less than 0.2 yr^{-1}). We confirm a galaxy as quenched through either the shape of their SFH (by requiring the logarithm of SFR at recent times to be at least two order of magnitudes below the logarithm of its peak value, SFR_{peak}) or through a fixed threshold value in sSFR which we set to 0.006 yr^{-1}. We then identify continuous "quenched time bins" by starting at the most recent time bin and moving towards earlier times. Once a bin has either a sSFR>0.006 yr^{-1} or a $\mathrm{SFR}>\mathrm{SFR}_{peak}/100$, then we stop. Time bins identified as "quenched" are reverted back to the values of SFH and ZH predicted by the rNN. Our method to identify quenched galaxies and quenched plateaus in their SFHs was developed through trial and error by visually inspecting predicted and true SFHs of dozens of galaxies. We note that the task of identifying quenched galaxies and quenched plateaus in the training data is substantially easier as the simulation completely shuts down star-formation so one needs only to look for contiguous zero-valued bins. However, the rNN has difficulty in predicting zeros and has much shallower quenching, meaning that a straight cut in SFR or sSFR fails to identify the features we are looking for.

We therefore adopt this more relaxed, empirically motivated threshold that better captures recently quenched systems in predicted histories. We found the above to be reasonably robust to small changes in the values we quote here. This filter has the strongest impact in population colours and SEDs.

Finally, the justification for (iv) comes from the observation that the stochastic corrections can cause a subtle but systematic decrease of the SFHs at recent times. This is due to the rNN having particular difficulty at predicting the long frequency components of ${}_{pred}A$ in (mostly) star-forming, low-mass galaxies. Therefore this filter offers a mild improvement for low-mass centrals.

Normalisation

The normalization procedure adjusts the corrected SFHs based on their stellar masses to match the standard deviation observed in the training set within specific timescale bins.

First, we partition the sample into 16 bins based on the logarithm of stellar mass of the predicted sample. The boundaries for these bins were determined using a linear spacing between the minimum and maximum logarithmic stellar mass values in the corrected dataset.

For each bin, and at each time step, the corrected SFHs were normalized. This involved adjusting the corrected SFHs to have the same standard deviation as the training SFHs within the same bin. Specifically, for each mass bin b and each time step t, the normalization was computed as follows:

$$h_{\text{norm-mod},b,t}(t) = (h_{\text{mod},b,t}(t) - - - \mu_{\text{mod},b,t}) \frac{\sigma_{\text{subtrain},b,t}}{\sigma_{\text{mod},b,t}} + \mu_{\text{mod},b,t} \tag{4.9}$$

where $\mu_{\text{mod},b,t}$ and $\sigma_{\text{mod},b,t}$ denote the mean and standard deviation of the modified set at time t in the stellar mass bin b. Similarly, $\sigma_{\text{subtrain},b,t}$ denotes the standard deviation of the subset of training sample (refer to Section 4.3.3 for more details) at time t in the stellar mass bin b.

Any resulting negative values in the normalized SFHs/ZHs were replaced with the unnormalized modified values to ensure non-negative star formation rates. After normalization,

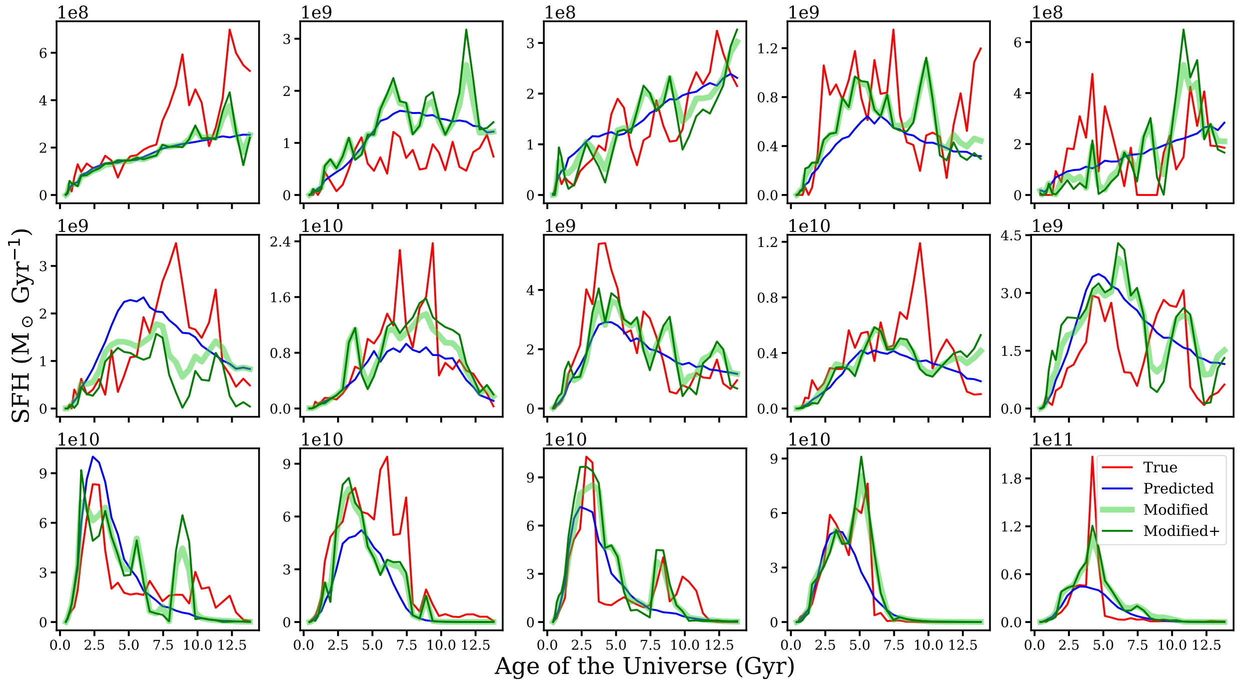

Figure 4.4.1: Star formation histories (SFHs) of 15 random central galaxies, grouped by stellar mass as a function of timescale. The top panel shows low-mass galaxies ($\log M_*/M_\odot < 10$), the middle panel shows intermediate-mass galaxies ($10 \leq \log M_*/M_\odot < 11$), and the bottom panel shows high-mass galaxies ($\log M_*/M_\odot \geq 11$). Blue lines represent the smooth rNN-predicted SFHs, red lines show the true SFHs, thick lightgreen lines show the stochastically modified SFHs and the sharp green lines show the results after normalization is applied to them. Corrections effectively restore short-timescale variability, though the specific features remain stochastic. The normalization step has a small effect on individual galaxies but ensures accurate scatter at the population level.

the FT was applied to each galaxy's SFH/ZH to derive the amplitude of the SFH/ZH in the frequency domain.

The impact of this normalisation procedure is small in individual galaxies, but is important in restoring the scatter of the training FT amplitudes (see next section), which, as we have seen, is underpredicted by the rNN.

4.4 Results

In this section we show the impact of the full set of corrections (i.e., including all filters presented in Section 4.3.3 and the normalisation step detailed in Section 4.3.3) on individual SFHs, mean optical spectra, optical colours distributions, and the mass-metallicity relation.

In figures we refer to results with all filters applied but no normalisation as "modified" and to results with all filters and normalisation as "modified+". In Appendix 4.A we show a subset of plots that demonstrates the specific impact of the filters and normalisation.

By construction, the mean FT amplitude and standard deviation of the corrected SFHs must match that of the true data — we show this as the green line in Fig. 4.2.2. The stochastic corrections applied to individual SFHs can be seen in Fig. 4.4.1 for 15 random central galaxies. The results for satellites are visually identical and not shown, and we show the same plots for ZHs in Appendix 4.B. The blue lines show the relatively smooth SFHs predicted by the rNN and the red lines show the true SFHs. The difference between the red and blue lines demonstrates well the lack of power on small scales — even though the broad shape of the SFH is well predicted by the rNN, short timescale features are entirely absent. The SFHs after the stochastic corrections are applied (green lines) are effective at adding these features back — the specific locations of short timescale features are, of course, not reproduced with our corrections as the phases are largely stochastic. However, statistically, the impact at population level is evident as we will see later. Fig. 4.4.1 also shows the impact of the normalisation step, detailed in Section 4.3.3. On individual galaxies the effect is minor compared to that introduced by the corrections, but it guarantees that population-level quantities have the correct standard deviations.

The optical spectra and mean colour distributions are shown in Figs. 4.4.2 and 4.4.3, respectively. Optical spectra and colours are computed according to the same methodology of CT23 — in summary we use the Flexible Stellar Population Synthesis package[186;205;206] to produce absorption and emission spectra according to the SFHs and ZHs in each case. For each galaxy, the optical spectra is then integrated over $g-$, $r-$, $i-$ and $z-$band filter response curves to produce rest-frame magnitude and colours. Emission lines are not shown in Fig. 4.4.2 for clarity, but are included in the computation of the colours shown in Fig. 4.4.3. Fig. 4.4.2 shows that mean and standard deviation of the optical spectra are improved after the stochastic corrections are applied (green line and shaded regions), at all mass ranges and for central and satellite galaxies. Changes in the shape of the spectra are hard to appreciate in this figure, but their impact can be seen in the colour distributions of Fig. 4.4.3. Here the

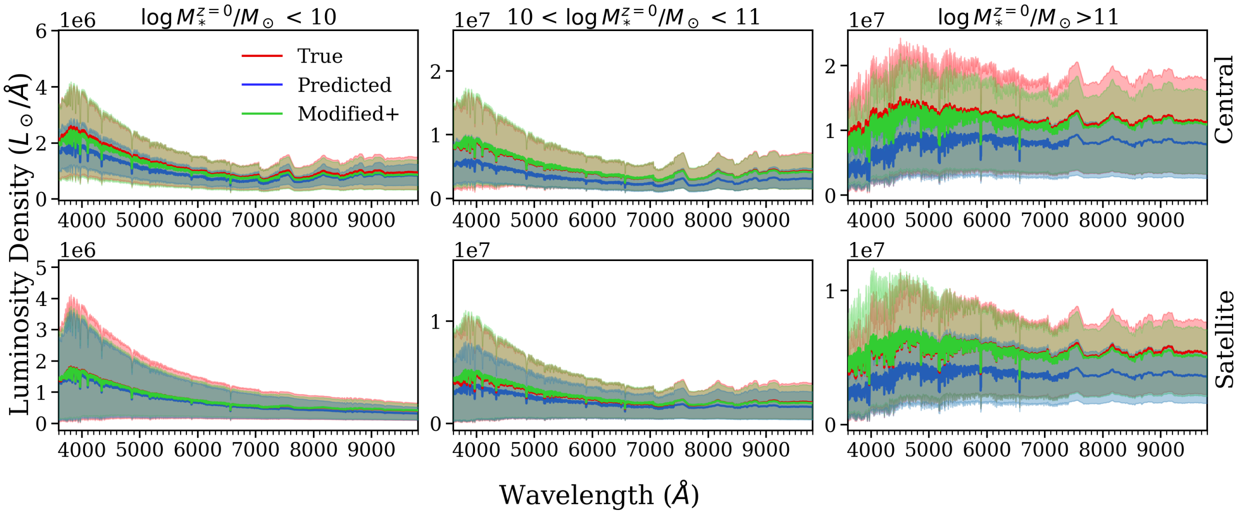

Figure 4.4.2: The median and 15th–85th percentile range for stacked central (top) and satellite (bottom) spectra in bins of stellar mass. Red, blue, and green represent spectra derived from the TNG, rNN-predicted, and modified star formation histories and metallicities. Emission lines are omitted for clarity. The corrections improve the overall match to the true spectra across all mass ranges, but the changes in spectral shape are subtle and better reflected in the corresponding color distributions (see Fig. 4.4.3).

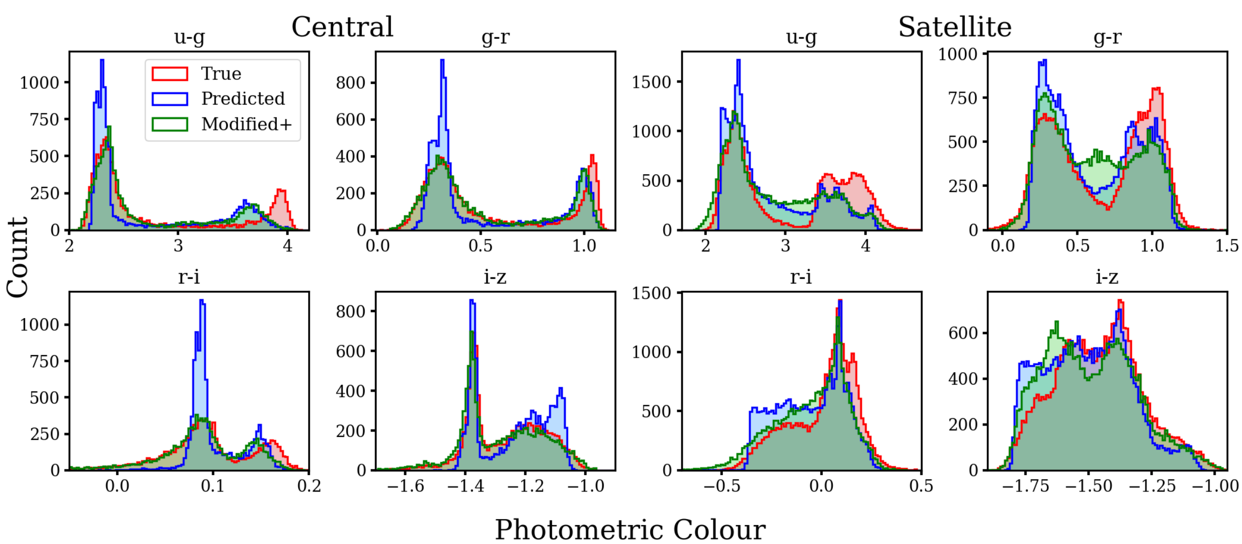

Figure 4.4.3: Photometric color distributions across five bands, showing differences between two consecutive bands. The bimodal distributions derived from TNG data are in red, rNN predictions in blue and modified spectral energy distributions in green. The corrections help recover the true color distribution for central galaxies, especially for the blue star-forming galaxies. However, in the case of satellite galaxies, the modifications slightly blur the galaxy bimodality especially in the g-r color.

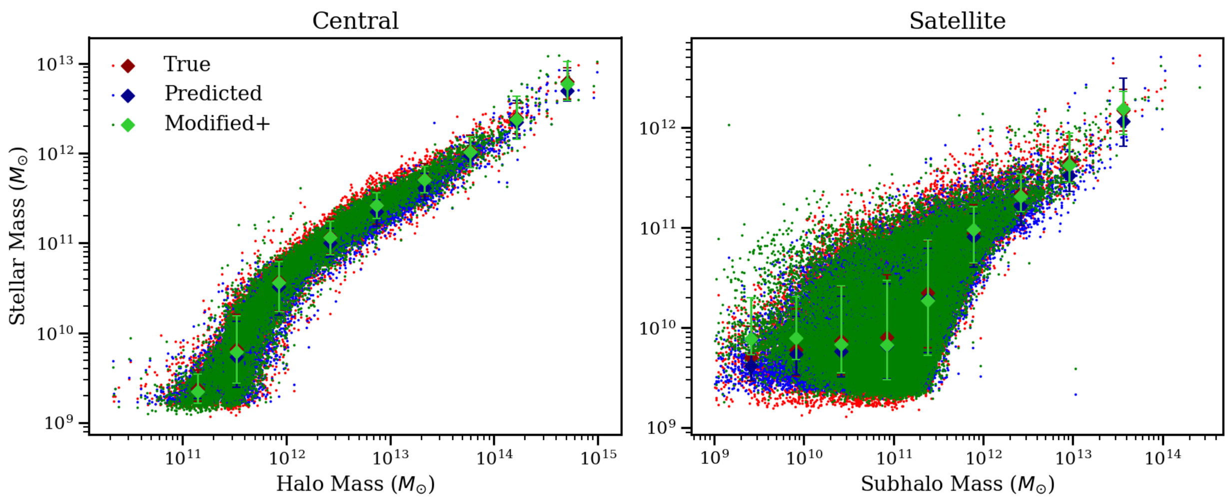

Figure 4.4.4: The stellar-halo mass relation evaluated using the true, predicted, and stochastically modified star formation rates for central galaxies (left) and satellite galaxies (right). Stellar mass is shown as a function of FoF halo mass for centrals and subhalo mass for satellites. Data points from the original TNG data are shown in red, the uncorrected predictions in blue, and the post-normalization modifications in green. Overlaid symbols represent the binned mean stellar mass in each halo mass bin, with error bars indicating the 15th to 85th percentile range. The stochastic modifications bring improvement to the mean and scatter of the SHMR on all mass ranges for centrals and for satellites — see Tables 4.4.1 and 4.4.2.

impact of the corrections is shown to be different for different colours and also for centrals and satellites. In the case of central galaxies, the stochastic corrections help recover the true colour distribution towards the red end, but fail to move the red sequence in $u-g$ to its true position. This seems to be associated with fast-quenching galaxies, whose SFHs are hard to predict by the rNN and hard to fix with stochastic corrections — a sharp transition to a flat (quenched) plateau requires coherent combination of phases, which our method cannot accommodate. In every other case for central galaxies, the stochastic corrections bring the colour distributions in good agreement with those of the true sample. For satellite galaxies, the situation is not as clear. Improvement is more marginal and, in $g-r$, the modifications blur the galaxy bimodality in satellites. The reason seems to be related specifically with low mass satellites and the difficulty in predicting reliable FT amplitudes in that regime[197].

Fig. 4.4.4 shows the stellar-to-halo relation (SHMR) for centrals and satellites. Tables 4.4.1 and 4.4.2 show the mean stellar mass and scatter, in bins of halo mass, for the true, predicted and modified+ datasets. Although the original predictions were already very good

Table 4.4.1: The mean (top three rows) and scatter (bottom three rows) of stellar mass in halo mass bins for true, predicted, and modified+ data for centrals galaxies. The modifications bring a small improvement to the mean and scatter (which were already well predicted by the rNN) in all halo mass bins.

Halo Mass	**True**	**Pred**	**Mod+**
		$\log(\langle M_*/M_\odot \rangle)$	
$\log(M_h/M_\odot) < 12$	10.15	10.08	10.13
$12 < \log(M_h/M_\odot) < 13$	11.17	11.09	11.16
$\log(M_h/M_\odot) > 13$	11.90	11.84	11.90
		$\sigma_{log(M_*/M_\odot)}$	
$\log(M_h/M_\odot) < 12$	0.45	0.43	0.45
$12 < \log(M_h/M_\odot) < 13$	0.30	0.28	0.29
$\log(M_h/M_\odot) > 13$	0.30	0.30	0.29

Table 4.4.2: Same as Table 4.4.1 but for satellite galaxies. The original predictions for satellites are marginally poorer than those for centrals, leading to larger improvements from the modifications.

SubHalo Mass	**True**	**Pred**	**Mod+**
		$\log(\langle M_*/M_\odot \rangle)$	
$\log(m_h/M_\odot) < 11$	10.16	10.07	10.17
$11 < \log(m_h/M_\odot) < 12$	10.61	10.53	10.58
$\log(m_h/M_\odot) > 12$	11.41	11.32	11.39
		$\sigma_{log(M_*/M_\odot)}$	
$\log(m_h/M_\odot) < 11$	0.42	0.38	0.41
$11 < \log(m_h/M_\odot) < 12$	0.52	0.50	0.53
$\log(m_h/M_\odot) > 12$	0.32	0.29	0.30

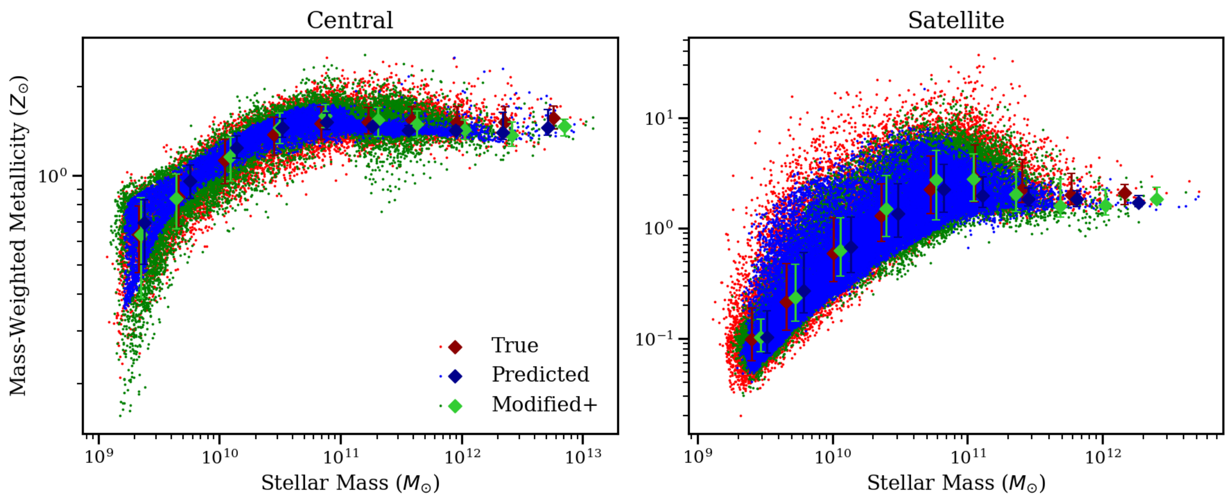

Figure 4.4.5: The mass-metallicity relation evaluated using the true, predicted, and stochastically modified chemical enrichment histories, for central galaxies (left) and satellite galaxies (right) and shown as a function of stellar mass. Data points from the original TNG data are shown in red, predictions in blue, and the stochastically modified results in green. The overlaid markers represent the binned mean stellar metallicity, and error bars denote the 15th to 85th percentile range within each stellar mass bin. While the predicted MZR captures the overall trend, it underestimates the scatter, particularly at intermediate and high stellar masses. The stochastic modifications improve the scatter for both centrals and satellites, though the full extent of variability still remains slightly underrepresented in satellites at lower masses — see Tables 4.4.3 and 4.4.4.

Table 4.4.3: The mean (top three rows) and scatter (bottom three rows) of the stellar metallicity of central galaxies, in bins of stellar mass, for the true, predicted and modified+ datasets. The predictions significantly underestimate the scatter of the MZR at intermediate and high stellar masses (by around 35% to 60% respectively), but the stochastic modifications are able to restore the scatter at these masses.

Stellar Mass	**True**	**Pred**	**Mod+**
		$\langle Z/Z_\odot \rangle$	
$\log(M_*/M_\odot) < 10$	0.81	0.82	0.78
$10 < \log(M_*/M_\odot) < 11$	1.38	1.41	1.42
$\log(M_*/M_\odot) > 11$	1.55	1.46	1.52
		$\sigma_{Z/Z_\odot}$	
$\log(M_*/M_\odot) < 10$	0.21	0.20	0.23
$10 < \log(M_*/M_\odot) < 11$	0.22	0.14	0.23
$\log(M_*/M_\odot) > 11$	0.17	0.07	0.18

Table 4.4.4: As Table 4.4.3 but for satellite galaxies. The predictions underestimate the mean and scatter at all masses (between around 20% to 75%, depending on the mass). The modifications bring substantial improvements in all cases except to the scatter of low mass satellites. Unlike the case with centrals, the stochastic corrections are not able to completely restore the mean and scatter of the true MZR.

Stellar Mass	**True**	**Pred**	**Mod+**
		$\langle Z/Z_\odot \rangle$	
$9.5 < \log(M_*/M_\odot) < 10$	0.47	0.33	0.35
$10 < \log(M_*/M_\odot) < 11$	2.05	1.61	2.01
$\log(M_*/M_\odot) > 11$	3.40	2.09	2.81
		$\sigma_{Z/Z_\odot}$	
$9.5 < \log(M_*/M_\odot) < 10$	0.69	0.37	0.34
$10 < \log(M_*/M_\odot) < 11$	1.99	1.18	1.66
$\log(M_*/M_\odot) > 11$	2.74	0.62	1.46

(within 0.1 dex), the stochastic corrections bring further (although modest) improvement to the mean and scatter of this fundamental relation. Finally, we consider the impact of the stochastic corrections on the mass-metallicity relation (MZR), which is shown in Fig. 4.4.5. One of the biggest limitations of the SFHs and ZHs recovered by the rNN was the difficulty in predicting the mean and scatter in the MZR. The MZR is plotted as a mass-weighted stellar metallicity as a function of stellar mass (in turn computed by integrating the SFH for each galaxy), so it depends on the accuracy of *both* the SFH and the ZHs. Tables 4.4.3 and 4.4.4 show the mean metallicity and its scatter in bins of stellar mass. The stochastic corrections bring substantial improvement to the mean and to the scatter of the MZR in centrals and satellites. The scatter in the MZR, in particular, was underpredicted by around 40% at intermediate masses (more at higher masses) in centrals and in satellites. Our stochastic corrections alleviated this issue (almost entirely in the case of centrals), demonstrating the importance of short-time scale events in order to get the scatter of this fundamental relation correctly. In the case of satellites, the modifications were not able to completely restore the true mean and scatter.

4.5 Summary and Conclusions

In this paper, we introduce an investigation of the galaxy-halo connection as a function of frequency with the intention of producing stochastic contributions to SFHs and ZHs. Our work is done in the context of the rNN predictions of CT23 that, although being successful at predicting the broad shape of SFHs and ZHs, were unable to predict short-term variability. We argued here that the cause must be, at least partly, that the phases of the mass assembly and star-formation/chemical enrichment histories are themselves decoupled on short timescales. We capitalized on the success of the rNN to directly predict the FT amplitudes of the SFHs/ZHs in order to introduce a formalism to compute semi-stochastic corrections.

Our main results can be summarised as follows:

- In TNG-100, the phases of the halo mass assembly and star-formation histories are correlated only on timescales greater than around 6 Gyr (Fig. 4.3.2, left). This is in agreement with a body of literature that associates short timescale contributions to star-formations histories with feedback processes or giant molecular cloud timescales, and intermediate to long timescales with gas accretion and depletion times (e.g. Sparre et al.[198]; Iyer et al.[199]; Caplar and Tacchella[200]). Here we attempt a more direct link between halo and star-formation histories by correlating the phases of these two components.

- The correlation between the halo mass assembly and star-formation history phases depend on stellar mass, galaxy type, and on whether a galaxy is a central or satellite, being almost non-existent for massive quenching galaxies (Fig. 4.3.2, left). The reason for the decoupling between halo and stellar growth on these galaxies, even on long timescales, is not clear. On long timescales, star-forming galaxies can be approximated as gas regulated systems, forming stars at fixed efficiency. A coherent accretion of dark matter and cold gas should then result in a correlation of the phases of halo and stellar growth. A decoupling might signify a break in star-formation efficiency or complex dependencies (for example, as a function of environment) that conspire to give uncorrelated phases on the whole. Studying this decoupling at higher temporal resolution, as a function of environment, and on different simulations will be subject of future work.

- The phases of the star-formation and chemical-enrichment histories are correlated over a wide range of the timescales probed, with a notable dependence on stellar mass, particularly at shorter timescales (Fig. 4.3.2, right). In low-mass galaxies, chemical enrichment and star-formation are coupled on all timescales, with the correlation decreasing with stellar mass.

- The stochastic corrections are successful at introducing power on short timescales and producing more realistic star-formation and chemical enrichment histories (Fig. 4.4.1 and 4.B.1).

- The stochastic corrections improve optical spectra and colour distributions, especially for central galaxies (Figs. 4.4.2 and 4.4.3). The stochastic corrections are unable to shift the red sequence in $u - g$ for central galaxies, which remains too blue. This is associated with fast-quenching galaxies which the rNN is unable to reproduce and the stochastic corrections (by construction) are unable to address.

- The improvement in the MZR afforded by the stochastic corrections (Fig. 4.4.5 and Tables 4.4.3 and 4.4.4) points to the importance of short timescale contributions to the scatter in this fundamental relation. Unlike the gas-phase metallicity, a mass-weighted stellar metallicity is sensitive to the full star-formation history of a galaxy, potentially making the stellar MZR a useful probe of the limits of the galaxy-halo connection. In the case of satellites, the semi-stochastic corrections were not able to fully restore the true mean and scatter, which may be related to difficulties in predicting the FT amplitudes of the ZHs for satellites.

The stochastic modeling technique presented here provides a flexible way to incorporate unresolved variability into models of galaxy formation that do not physically model those scales. We demonstrate our method on the predictions of a machine-learning model, but the technique is broadly applicable to empirical and semi-analytic models (SAM) as well. The time-resolution of a SAM is tied to the time-steps used to compute the baryonic properties of galaxies. This resolution is dependent on the dark matter simulation, but if we take the Millennium simulation as an example[104], the time resolution is given by 58 snapshots × 20 time steps in each, at an average time step of 10^7 years[207]. Most SAMs will compute SEDs on the fly and store only the "current" SED at a specific snapshot. Without full SFHs/ZHs, it would be difficult to implement our method. L-Galaxies[208] is an exception in that it saves SFH and ZH of every galaxy, and produces SEDs in post-processing. The time resolution at which SFHs and ZHs are saved is chosen in a way that balances computational restrictions and the accuracy of galaxy magnitudes[207], using a scheme that maintains high resolution at young ages but degrades it at older stellar ages. At $z = 0$, this results in 19 bins with a width between 1.5×10^7 years at young ages (which matches the time step resolution) and 2.1×10^9 years at old ages (which is a degradation of the time step resolution). The approach proposed in our paper could be used to add stochasticity on time-scales shorter than the timestep width, or introduce additional stochasticity on longer timescales. Our approach requires an estimate of the true amplitude of the power spectra of the SFHs/ZHs over the timescales of interest. These could be estimated directly from cosmological hydrodynamical simulations, making our method a potentially efficient way to combine information from expensive hydrodynamical simulations with fast SAM models.

The properties of *individual* galaxies are, in general, not significantly improved. That is expected given the semi-stochastic nature of our corrections. The key point from this study is that, on timescales shorter than around 6 Gyr, stellar and chemical enrichment histories have decoupled from halo growth. Therefore from a galaxy-halo connection perspective, it becomes irrelevant when short-timescale events happen. Provided that we have an estimate for the frequency of decoupling, we can always stochastically correct SFHs and ZHs, within the constrains shown in this paper (namely the correlation between SFH and ZH phases, and the physical range of phase values, which depends on frequency).

Understanding in detail how the phases of MAHs decorrelate from the phases of SFHs/ZHs can help GHC models produce more realistic predictions, which our method demonstrates (though there will be other ways to implement this). We see this as a first step only — Fig. 4.3.2 clearly shows that this decoupling happens differently for different types of galaxies. We do not know why this is yet, and we have not taken that into account in our corrections, which we leave for future work.

Finally, our formalism for the stochastic correction can be cast as a simple empirical semi-stochastic model that parametrizes the galaxy-halo connection as a correlated time-series between halo assembly, star formation and chemical enrichment histories. Our results suggest that fundamental relations, such as the MZR, might have the sensitivity to *observationally constrain* the timescale at which the phases of halo assembly and star formation decouple, for example, and how that might change with galaxy type. Such a measurement would be an entirely new way to test hydrodynamical models of galaxy formation. We leave that investigation for future work.

Acknowledgements

We would like to thank the anonymous reviewer for their careful and detailed report, which led to several clarifications and improvements throughout the paper.

JB is grateful for support from the US Department of Energy via grants DE-SC0021165 and DE-SC0011840. JB is partially supported by the NASA ROSES grant 12-EUCLID12-0004. The UKRI Science and Technology Facilities Council supported HGC under grant ID ST/T506448/1, which the authors gratefully acknowledge.

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231.

We wish to thank the IllustrisTNG project for access to their data.

This research has made use of NASA's Astrophysics Data System and the arXiv open-access repository of electronic preprints and postprints.

JB thanks Lado Samushia for useful discussions and comments. RT acknowledges helpful conversations with Amelie Saintonge that partially influenced the direction of this paper.

Data Availability

The data used for this research are currently available in the CT23 zenodo repository, mentioned in CT23. The data supporting the findings will be added to it when the paper is accepted. The code, and documentation of the code and data, will also be made publicly available in an online GitHub repository.

Appendix

4.A Analysis of semi-stochastic corrections

In this appendix, we analyze supplementary figures to illustrate the impact of applying our formalism (Section 4.3.1 and 4.3.2) under different cases of implementation (Section 4.3.3 and 4.3.3) and emphasize the importance of each filter and normalization in improving the model's ability to capture variability and scatter. We present four sets of results to justify the need for each:

- Case I: Method + (i)
- Case II: Method + (i) + (ii) + (iii) + (iv)
- Case III: Method + (i) + (ii) + (iv) + Norm
- Case IV: Method + (i) + (ii) + (iii) + Norm

We focus only on central galaxies (except for Case II that presents both centrals and satellites) where the effects of each filter and normalization are more pronounced. Only figures that clearly highlight differences and justify the corrections are included.

4.A.1 Case I

In this case, we apply stochastic corrections with only Filter(i), which replaces any negative values generated during the stochastic modifications on the original rNN-predicted values. This serves as a basic setup to investigate the effects of corrections without more refined filtering or normalization steps.

Fig. 4.A.1 presents the star formation histories (SFHs) for 15 random central galaxies, categorized by stellar mass similar to Fig 4.4.1. But in this case the modified SFHs start

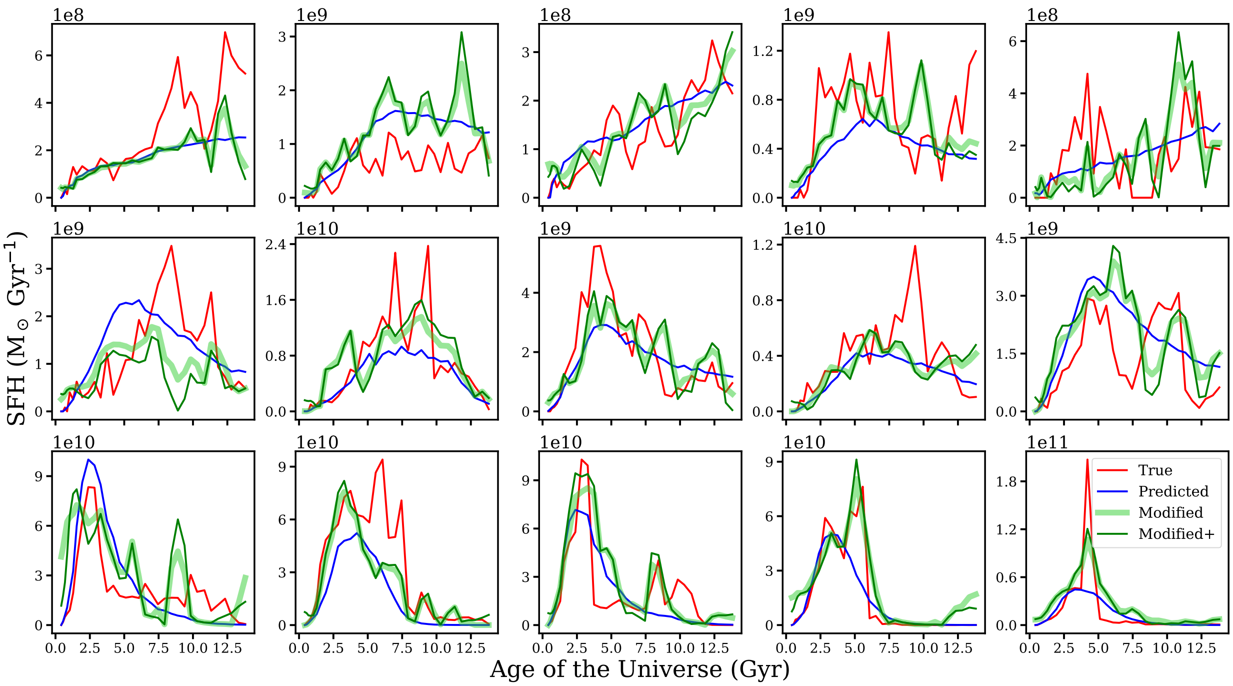

Figure 4.A.1: Same as Fig. 4.4.1, but when the correction method excludes all filters(except filter(i)) and normalization. The top panel shows low-mass galaxies ($\log M_*/M_\odot < 10$), the middle panel shows intermediate-mass galaxies ($10 \leq \log M_*/M_\odot < 11$), and the bottom panel shows high-mass galaxies ($\log M_*/M_\odot \geq 11$). The modified SFHs start from unphysically high values, particularly at early times, due to how the Fourier transform and its inverse operate without constraints. This highlights the necessity of filter (ii) to prevent such unrealistic initial conditions.

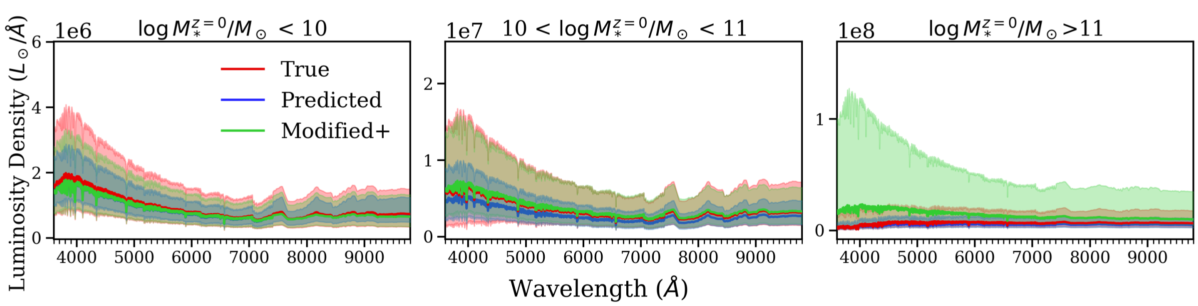

Figure 4.A.2: SEDs for central galaxies when the correction method excludes all filters(except filte(i)) and normalization. The true spectra are shown in red, the rNN-predicted spectra in blue, and the stochastically modified spectra in green. For low-mass galaxies ($\log M_*/M_\odot < 10$), there is negligible improvement as the corrections cannot significantly alter the underestimated rNN predictions. In high-mass galaxies ($\log M_*/M_\odot \geq 11$), the modifications introduce excessive variability, leading to unrealistically high spectra due to unaddressed quenching effects. Intermediate-mass galaxies show balanced spectra on average, but this balance results from offsetting over- and undercorrections, masking individual inaccuracies.

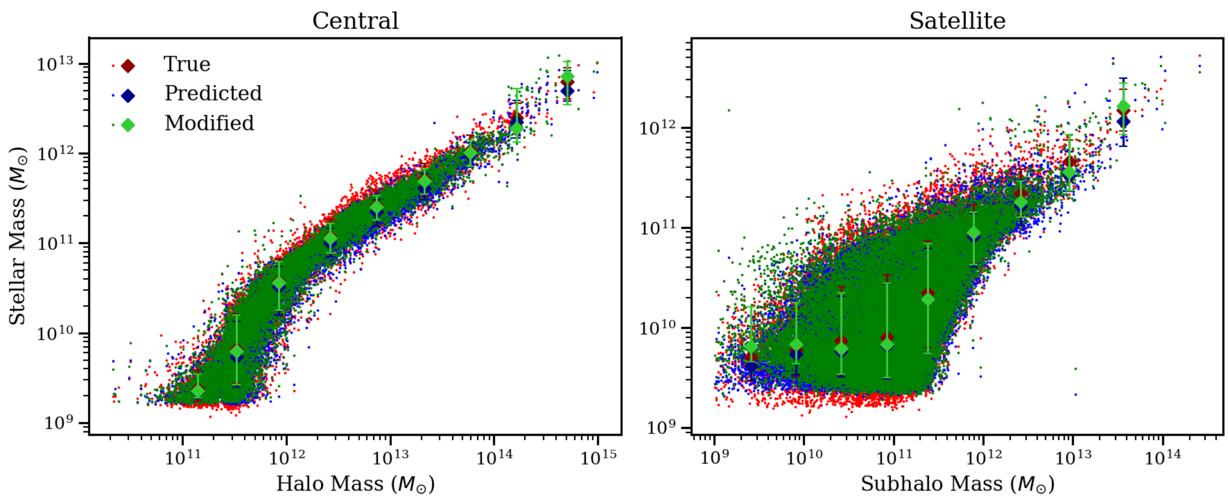

Figure 4.A.3: The stellar-halo mass relation (SHMR) for central (left) and satellite (right) galaxies similar to Fig. 4.4.4 but here the correction method excludes normalization. The corrections show significant improvement in the SHMR where the modifications better capture the scatter and alignment with true values, although slight discrepancies remain for especially for high-mass ($\log M_*/M_\odot \geq 11$) satellite galaxies, where variability is still not fully recovered compared to Fig. 4.4.4.

from unphysically high values due to the limitations of the FT and inverse FT methods, which fail to address the early-time star formation features effectively. This highlights the necessity of Filter (ii), which is designed to mitigate this issue.

Fig. 4.A.2 shows the spectra distributions for central galaxies. For low-mass galaxies, the corrections show negligible improvement while in high-mass quenched galaxies, they introduce excessive variability leading to excessive overestimation. For intermediate-mass galaxies, the corrections seem to match the true distribution at first glance, but this is misleading. The apparent improvement in the average spectra results from balancing over- and under-corrections across different galaxies. While the mean spectra looks reasonable, the variability introduced is inconsistent, leaving individual galaxies either over- or under-corrected. Thus, without proper filtering, the overall improvement remains superficial and doesn't reflect true accuracy.

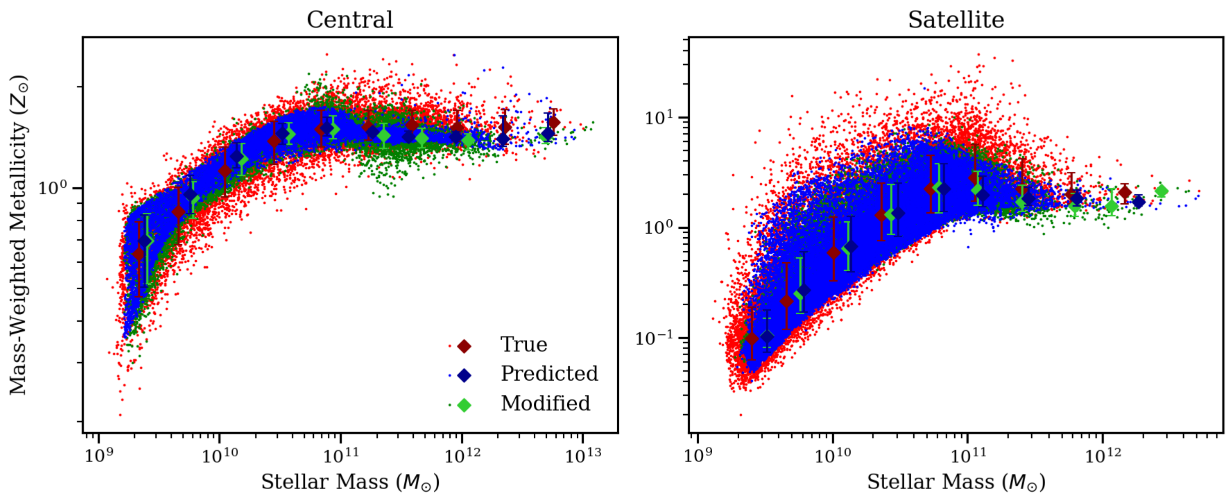

Figure 4.A.4: The mass-metallicity relation (MZR) for central (left) and satellite (right) galaxies, similar to Fig. 4.4.5, but with the correction method excluding normalization. Although these corrections show significant improvement in the stellar mass (Fig. 4.A.3), the improvement is not equally reflected in the metallicity. There is a notable lack of recovery in both the mean and scatter in all mass ranges, compared to the normalized results in Fig. 4.4.5

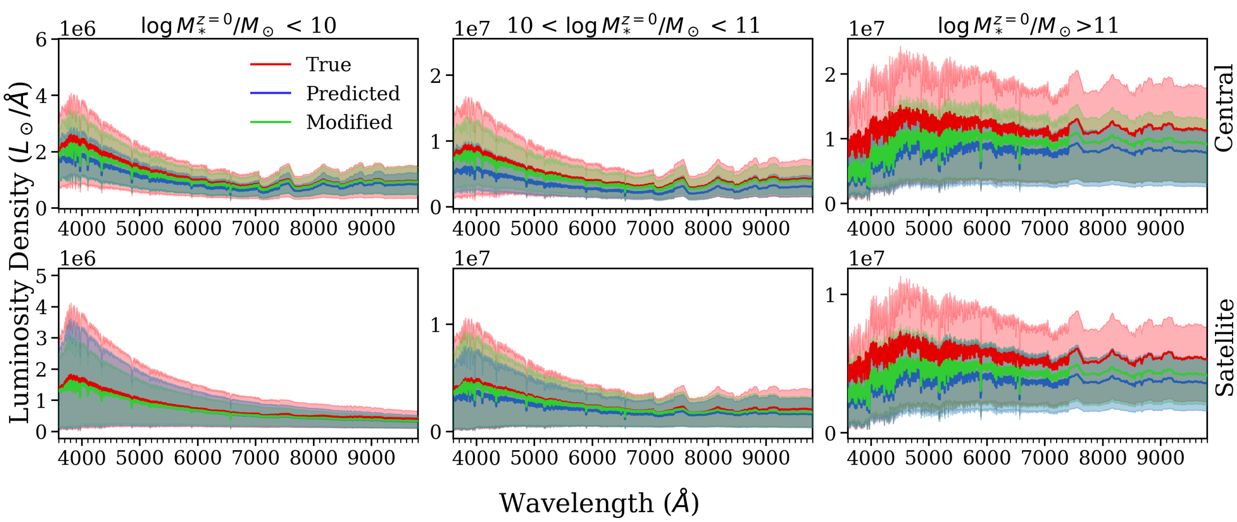

Figure 4.A.5: Spectral energy distributions (SEDs) for central (top) and satellite (bottom) galaxies, similar to Fig. 4.4.2, but with the correction method excluding normalization. Although the corrections significantly improve the match to the true spectra, particularly in low-intermediate mass central galaxies and intermediate mass satellite galaxies, some discrepancies still remain in the amplitude and shape, especially for high-mass galaxies ($\log M_*/M_\odot \geq 11$). This highlights the need for normalization to properly recover both the mean and scatter in the spectra.

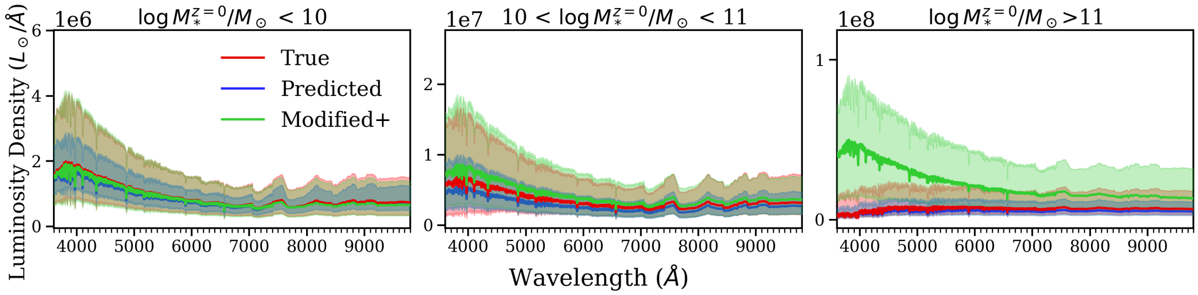

Figure 4.A.6: Spectral energy distributions (SEDs) for central galaxies when filter (iii) is excluded. The absence of this filter has minimal impact on low-mass galaxies due to the lack of quenched populations. However, for higher-mass galaxies, the correction method introduces artificial starburst events in quenched regions, leading to significantly overestimated spectra. This demonstrates the importance of filter (iii) for accurately modeling star formation in quenched galaxies.

4.A.2 Case II

Fig. 4.A.3 shows significant improvement in the SHMR, especially in the low and intermediate mass ranges, as the model better aligns with true values by introducing necessary variability in the SFH. However, the MZR (Fig. 4.A.4) remains less effective, particularly in the scatter. Similar trends are seen in the SEDs (Fig. 4.A.5) where corrections improve the match to the true spectra in both mean and scatter, though some discrepancies in amplitude and shape remain. These gains are achieved through normalization (Section 4.3.3) that aligns the variability and scatter with true data, without altering overall trends.

4.A.3 Case III

Fig. 4.A.6 shows the effect of stochastic corrections on the spectra for central galaxies when filter (iii) is excluded. Comparison to Fig. 4.4.2 shows that this filter has negligible impact in low-mass galaxies, due to the lack of quenched galaxies. However, for higher masses, the correction formalism introduces artificial starburst events in quenched regions, leading to significantly overestimated spectra. This is further reflected in Fig. 4.A.7, where this implementation improve the color distribution of blue, star-forming galaxies but worsen the red, quenched galaxies.

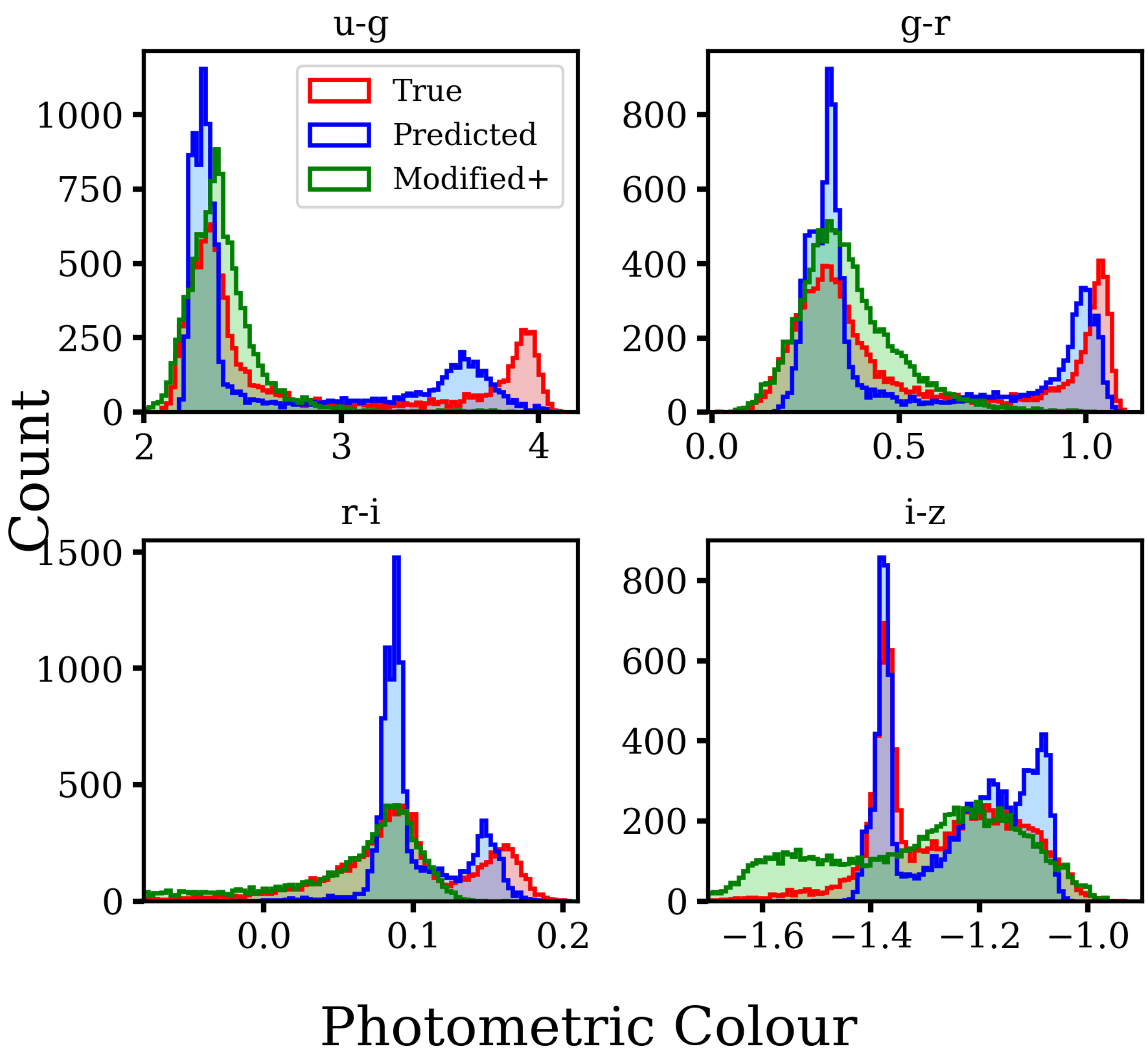

Figure 4.A.7: Color distributions for central galaxies where filter (iii) is excluded. While corrections improve the distribution for blue, star-forming galaxies, they worsen the red, quenched galaxy distribution. Without the restoration of quenched scales, artificial starburst events are introduced, resulting in incorrect colors for quenched galaxies and an increased scatter in the red color bands.

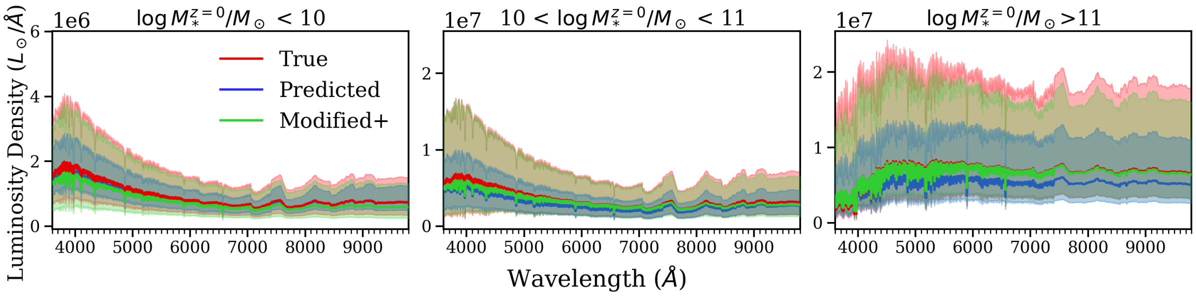

Figure 4.A.8: Spectral energy distributions (SEDs) for central galaxies where filter (iv) is excluded, results in a noticeable deterioration in the spectra for low-mass, star-forming galaxies, which show reduced accuracy compared to the results in Fig. 4.4.2 with the full set of filters.

4.A.4 Case IV

Fig. 4.A.8 presents the spectra for central galaxies where filter (iv) is excluded. Compared to Fig. 4.4.2, which includes this filter, the spectra for low-mass, star-forming galaxies are less accurate, showing a small but noticeable deterioration. This shortfall is also reflected in the color distribution (Fig. 4.A.9 compared to Fig. 4.4.3), where excluding filter(iv) leads to a slight lack of correction for blue galaxies.

4.B Corrections in Metallicities

The stochastic corrections applied to individual metallicity histories (ZHs) are presented in Figure B1, which shows the ZHs for 15 random central galaxies. These corrections, similar to those applied to star formation histories (SFHs) in Fig. 4.4.1, introduce short-timescale variability effectively, demonstrating the robustness of the method. By adding stochasticity, we are able to recover small-scale fluctuations in the ZHs that were not captured by the rNN predictions.

"ଯେଉଁଠି ଅନିୟମ ଅଛି, ସେଠି ରହିଛି ପ୍ରାଣ ଓ ନୂତନତାର ସୂତ୍ର।"

— From modern Odia poetic thought

"Where irregularity exists, there lies the seed of life and novelty."

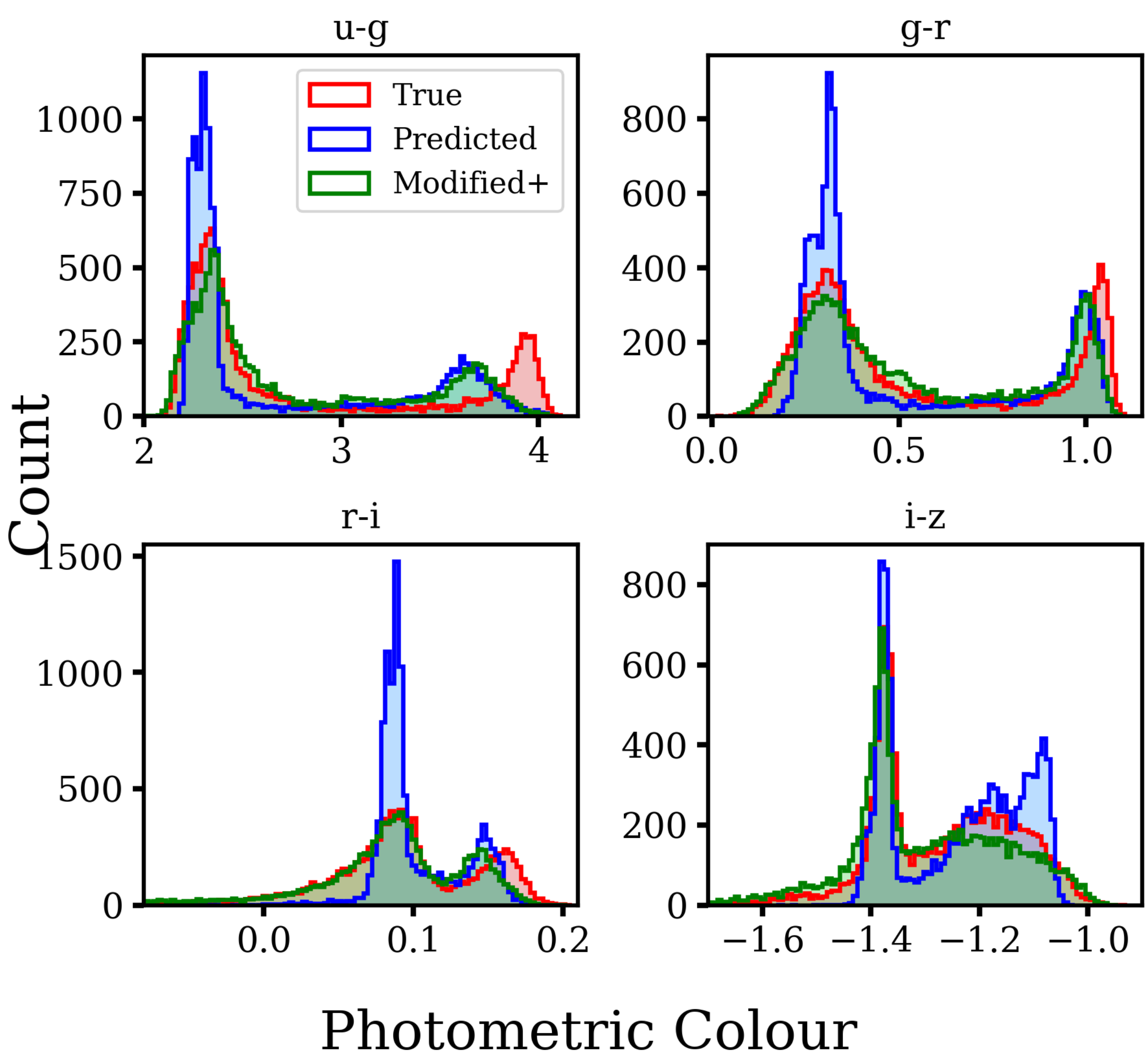

Figure 4.A.9: Color distributions for central galaxies where filter (iv) is excluded. While the overall color recovery is effective, it improves further when this filter is included, as seen in Fig. 4.4.3 because the exclusion of this filter leads to slight discrepancies, particularly in blue, star-forming galaxies, where the scatter is not fully captured.

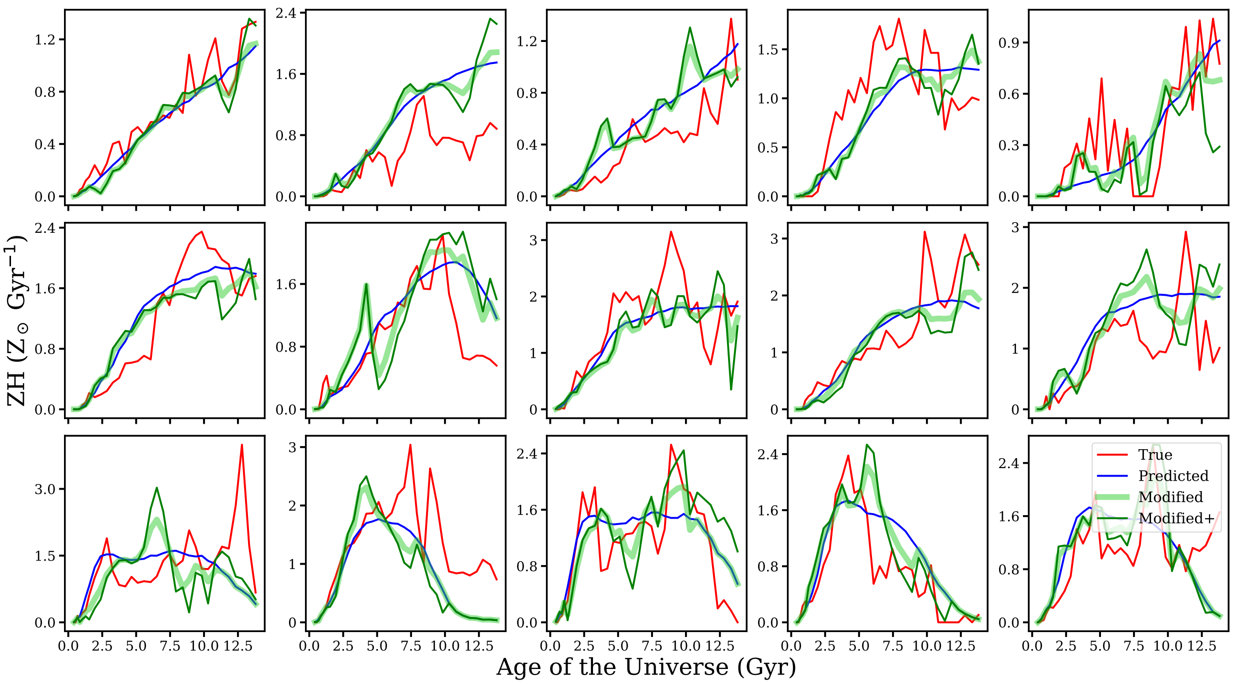

Figure 4.B.1: Similar to Fig. 4.4.1, but for the metallicity histories (ZHs) of 15 random central galaxies. The top panel shows low-mass galaxies ($\log M_*/M_\odot < 10$), the middle panel shows intermediate-mass galaxies ($10 \leq \log M_*/M_\odot < 11$), and the bottom panel shows high-mass galaxies ($\log M_*/M_\odot \geq 11$). Blue lines represent the smooth rNN-predicted ZHs, red lines show the true ZHs, light green lines show stochastically modified ZHs, and dark green lines show the results after normalization. Corrections effectively restore short-timescale variability in the ZHs as well.

Chapter 5

Evaluating the galaxy formation histories predicted by a neural network in pure dark matter simulations

This chapter originally appeared in the literature as :

Harry Chittenden, Jayashree Behera and Rita Tojeiro, Evaluating the galaxy formation histories predicted by a neural network in pure dark matter simulations, arXiv:2409.16079[197].

Abstract

We investigate a series of galaxy properties computed using the merger trees and environmental histories from dark matter only cosmological simulations, using the predictive semi-recurrent neural network outlined in Chittenden and Tojeiro[209], and using stochastic improvements presented in our companion paper: Behera, Tojeiro, and Chittenden[188]. We apply these methods to the dark matter only runs of the IllustrisTNG simulations to understand the effects of baryon removal, and to the gigaparsec-volume pure dark matter

simulation Uchuu, to understand the effects of the lower resolution or alternative metrics for halo properties. We find that the machine learning model recovers accurate summary statistics derived from the predicted star formation and stellar metallicity histories, and correspondent spectroscopy and photometry. However, the inaccuracies of the model's application to dark simulations are substantial for low mass and slowly growing haloes. For these objects, the halo mass accretion rate is exaggerated due to the lack of stellar feedback, yet the formation of the halo can be severely limited by the absence of low mass progenitors in a low resolution simulation. Furthermore, differences in the structure and environment of higher mass haloes results in an overabundance of red, quenched galaxies. These results signify progress towards a machine learning model which builds high fidelity mocks based on a physical interpretation of the galaxy-halo connection, yet they illustrate the need to account for differences in halo properties and the resolution of the simulation.

5.1 Introduction

Cosmic hydrodynamical simulations are a valuable tool for modelling the causal relationship between galaxies and haloes across cosmic time; known as the galaxy-halo connection (GHC). Modelling galaxy evolution in its full complexity, however, is a difficult task. Numerous physical processes such as the accretion and cooling of gas are necessary to fuel star formation and synthesise heavy elements[210;211], which in turn are influenced by the growth of the dark matter halo and its interactions with the surrounding cosmological structures[212]. Capturing the complete GHC requires details of large scale structure evolution on megaparsec scales to compute massive galaxies and clusters, and the physics internal to the galaxy to compute summary statistics such as mass-metallicity relations.

In Chittenden and Tojeiro[209], hereafter CT23, we designed a semi-recurrent neural network capable of reproducing the GHC in the Illustris: The Next Generation (TNG) simulations[213–218], in which we predict the star formation history (SFH) and metallicity history (ZH) of central and satellite galaxies, from which we recover physical relations such as the stellar-halo mass relation (SHMR) and mass-metallicity relation (MZR), reproduce observational statistics such as colour bimodality, and show the growth and internal properties of the halo to influence the predicted SFH and stellar mass, while environmental properties influence the ZH and stellar metallicity.

Machine learning models which can emulate galaxy properties from their dark matter component alone can be used to make predictions in cosmic N-body simulations, where the absence of galaxies allows for such simulations to computed with much greater size and resolution than their hydrodynamical counterparts. Many studies would benefit from a simulation containing galaxy evolution on these scales[219–221], however the limited computational resources required for such simulations make computing such a hydrodynamical suite highly impractical.

In this paper, we explicitly evaluate the performance of the trained galaxy evolution model on N-body simulations. Predictions are expected to be somewhat different in dark simulations due to the lack of baryonic effects on the halo mass function[222–225], and due to differences in mass and spacial resolution, and alternative definitions or calculations of key halo and environmental properties. We compare predictions in the hydrodynamical TNG simulations, discussed at length in CT23, with the equivalent dark simulations in the TNG suite (hereafter TNG-Dark), and the Uchuu dark matter simulation[226].

Uchuu is a high fidelity N-body simulation which assumes the same Planck Collaboration[227] cosmological parameters as all TNG simulations, therefore there are little to no effects to consider regarding mean matter density or Hubble expansion. By comparing predictions in TNG-Hydro with TNG-Dark, we isolate the effect that baryons have on the predicting power of the neural network. By comparing TNG-Dark with Uchuu, we show the effect of Uchuu's lower resolution and alternative definitions of halo properties. The predictions made in Uchuu are therefore the simplest results to be expected for a high fidelity galaxy catalogue produced using our model.

Further to this endeavour, we apply a stochastic modification to the N-body simulation predictions, whose methods and improvements to our TNG-Hydro-based galaxy-halo statistics are outlined in Behera, Tojeiro, and Chittenden[188], hereafter BTC24. It is shown in CT23 that the machine learning model alone is inadequate for predicting star formation events on short timescales, in spite of the success of predicting the general evolution of galaxies across all times. These "stochasti" features are well recovered by this correction, and the results of BTC24 illustrate that certain missing properties of our fiducial galaxy-halo statistics can be attributed to the physics of variable star formation. We apply this modification to the TNG-Dark and Uchuu data to assess the plausibility of reproducing this variability in N-body mocks.

In this paper, we investigate the quality of predictions between our original results and the predictions in TNG-Dark and Uchuu, explain systematic differences between the results in each of these simulation suites and discuss the suitability of our model for reproducing galaxies in a pure dark matter simulation. In Section 5.2 we discuss the definitions of the dark matter data in each of the simulations, and our methods of acquiring the necessary input data. We compare properties of the input data in Section 5.3, baryonic predictions in Section 5.4 and observational results in Section 5.5, eludicating their disparities each time, and exhibiting the improvements made by the stochastic amendment introduced in BTC24. We evaluate the successes and failures of the model, the stochastic correction, and their proficiency for modelling the GHC on gigaparsec scales in Section 5.6, and summarise our findings in Section 5.7.

5.2 Data Acquisition

In this section, we describe how the data from the pure dark matter simulations are processed for use in our neural network, and discuss the effects of fundamental differences between these simulations on the quality of our data.

Full details regarding data processing are given in CT23, however the data processing procedure can be summarised as follows:

- Calculate mass accretion histories by finite differencing the (sub)halo mass along the main progenitor branch (MPB).

- Interpolate the MPB over the time domain of every third snapshot in TNG, down to and including $z = 0$.

- For each of these snapshots, calculate overdensities and skews[i] using the Grid Search In Python (GriSPy) package[228].

- Compute the proxy for circular velocity history: the square root of the ratio of (sub)halo mass ($M_\odot$) to half-mass radius (Mpc).

- For satellites, compute infall parameters, e.g. scaled infall time, infall velocity, based on and including the scaled accretion time from Shi et al.[229]; from the MPBs of the satellite and its host halo. If the simulation contains baryonic data, discard any samples where the stellar mass of the host halo is greater than 10% of the halo mass.

- For centrals, acquire the $z = 0$ cosmic web properties calculated using the Discrete Persistent Structure Extractor (DisPerSE) code[230]. The TNG-Hydro results are publicly available[ii].

- Calculate the specific mass accretion gradients shown in Montero-Dorta et al.[231]. Fit a Gaussian function to the distribution of gradient values, and discard any samples whose gradient is over 5σ from the mean of the best fit Gaussian.

- Apply a Gaussian quantile transformation (GQT) to the input data where necessary, using time-independent transformations for most temporal variables. Apply this transformed data to the network.

This gives a list of quantities used in the neural network model, summarised in Table 5.2.1. As in CT23, we use the integrated stellar masses and mass-weighted metalllicities derived from predicted SFHs and ZHs to produce summary statistics such as the SHMR and MZR, and use the Flexible Stellar Population Synthesis (FSPS) code[232;233] to calculate SEDs and other observables from these network output quantities.

[i] A measure of the mass-weighted radial distribution of subhaloes exterior to the target subhalo, tracing subhalo interaction history, as discussed in CT23.

[ii] https://github.com/Chris-Duckworth/disperse_TNG/

	Network Data			
	Quantity	Notation	Units	Network
Temporal Features	Halo Mass Accretion Rate	$\dot{M}_h$	$M_\odot$/Gyr	Both
	Subhalo Mass Accretion Rate	$\dot{m}_h$	$M_\odot$/Gyr	Satellite
	1Mpc Overdensity	δ_1		Both
	3Mpc Overdensity	δ_3		Central
	5Mpc Overdensity	δ_5		Central
	Circular Velocity (proxy)	$\tilde{v}_{\mathrm{vir}}$	$\sqrt{(}M_\odot/\mathrm{Mpc})$	Both
	Dark Matter Half-Mass Radius	$R_{\frac{1}{2}}$	Mpc	Both
	1Mpc Radial Skew	μ_3		Satellite
	3Mpc Radial Skew	μ_3		Central
	Distance To Closest Subhalo	d_{μ_3}	Mpc	Both
Non-Temporal Features	Specific Halo Mass Accretion Gradient	β (c) β_{halo} (s)	$\log \mathrm{Gyr}^{-2}$	Both
	Specific Subhalo Mass Accretion Gradient	β_{sub}	$\log \mathrm{Gyr}^{-2}$	Satellite
	Scaled Infall Time	a_{infall}		Satellite
	Scaled Formation Time	a_{max}		Satellite
	Infall Mass Ratio	μ		Satellite
	Infall Velocity	v_{rel}	km/s	Satellite
	$z = 0$ Cosmic Web Distances	d_{CW}	kpc	Central
	Starting Time	t_{start}	Gyr	Both
	$z = 0$ Halo Mass	M_h	$M_\odot$	Both
	Maximum Absolute Halo Accretion Rate	$\mid \dot{M}_h \mid$	$M_\odot$/Gyr	Both
	$z = 0$ Subhalo Mass	m_h	$M_\odot$	Satellite
	Maximum Absolute Subhalo Accretion Rate	$\mid \dot{m}_h \mid$	$M_\odot$/Gyr	Satellite
Targets	Star Formation History	$\mathcal{S}$	$M_\odot$/Gyr	Both
	Metallicity History	$\mathcal{Z}$	$Z_\odot$	Both
	$z = 0$ Stellar Metallicity	Z	$Z_\odot$	Both
	$z = 0$ Stellar Mass	M_s	$M_\odot$	Both
	Mass Weighted Age	MWA	Gyr	Both

Table 5.2.1: Simplification of Table 1 in CT23, summarizing the quantities used in both neural networks, according to their input and output layers and their arrangement in the network. Additional columns show the variables' units, and the central and/or satellite networks which use them.

5.2.1 TNG-Dark

Table 5.2.2 shows the key differences between the simulations used in this work. The TNG-Dark simulations are of similar resolution to their hydrodynamical counterparts, and therefore resolution effects will be small if at all applicable. The time domains of their snapshots are also equivalent. The crucial difference between the two results is therefore due to the absence of baryonic effects on the haloes in the TNG-Dark simulations.

In order to interpret the differences between the dark and hydrodynamical TNG simulations, we select samples in the dark simulation by cross-matching SubLink trees with our original, hydrodynamical dataset[234;235], and samples with no cross-matched result in TNG-Dark are discarded. In Figure 5.2.1 we show that for central haloes this has little effect on the $z = 0$ halo mass distribution of our data, yet for satellite objects the number of samples begins to decline sharply below $\sim 4 \times 10^{10} M_{\odot}$.

Generally, objects in one simulation are not accurately matched to the other for two reasons. Firstly, objects in proximity to much larger haloes may pass within the virial radius of the larger halo in one of the two simulations, allowing for central haloes to be matched to satellites, and vice versa. Similarly, and particularly so for low mass objects, the SubFind algorithm may combine two distinct subhaloes into a single object, such that one low mass subhalo will cease to exist in the lower resolution simulation (TNG-Dark). These effects eliminate $\sim 39\%$ of our satellite subhaloes compared with $\sim 1\%$ of our central haloes.

Most TNG-Dark variables were computed in the same manner as the data from hydrodynamical simulations. The exception is the $z = 0$ DisPerSE cosmic web distances, which, due to the geometric correspondence of the hydrodynamical and dark simulations, are identical to their cross-matched counterparts. We therefore assign the cosmic web properties of TNG-Hydro samples to their cross-matched equivalents in TNG-Dark, in lieu of publicly accessible data for the latter.

5.2.2 Uchuu

Merger Forests

Uchuu merger trees are grouped into 2,000 "Forest": ensembles of merger trees which contain all haloes which have interacted with any given member of the forest, and occupy a volume of space separate from all other forests. Each forest can therefore be treated independently of another, as the Consistent-Trees code[236] used to define merger trees in Uchuu is run independently in groups which occupy a fixed volume, which are then concatenated if they interact or pass within 25 Mpc/h of each other[226].

In this work we use forest 1411, the largest Uchuu forest, and acquire similar samples to the TNG dataset by drawing from the TNG-Dark halo mass distribution and sampling the Uchuu forest at $z = 0$ accordingly, which is shown by Figure 5.2.1. This results in approximately 30,000 central and satellite haloes each.

As the chosen Uchuu forest has a geometric mean side length of approximately 234 Mpc, it is of similar volume to the TNG simulations, and thus for high mass objects, sampling the (sub)halo mass distributions effectively approximates the (sub)halo mass functions of Uchuu. At lower masses, sampling Uchuu according to the TNG-Dark distribution ensures that the properties in different mass ranges can be compared without sample size bias. While this selection consistutes our test data for the neural network, environmental quantities are computed using the complete Uchuu forest, and are considered unbiased due to the TNG and Uchuu simulations having the same cosmic matter density parameters.

We use the YTree Python package to extract the main progenitor branches from the Uchuu forest. The quantities we obtain directly from the forest include halo mass, half mass radius, positions and peculiar velocities. Cosmic web distances are obtained by applying DisPerSE to the forest, which being a self-consistent subset of Uchuu means that the algorithm does not need to be applied to the full Uchuu simulation. All other quantities, such as overdensities, are calculated using the same methods as in CT23.

Alternative Definitions

Unlike TNG, SubLink merger trees do not exist in Uchuu, and haloes and their substructures are defined using the Rockstar code[237]. There exists a flag in each Uchuu merger tree which indicates the ID of the halo which hosts the target halo. A "first order" satellite halo is one whose host is a central "zeroth order" halo, a "second order" halo is hosted by a first order halo, and so on. We therefore treat first order haloes as satellite haloes, and zeroth order haloes as central haloes.

In TNG, halo and subhalo masses are taken as the sum of all masses of dark matter particles bound to the group by the SubLink algorithm. This field does not exist in the Uchuu merger trees, however a number of definitions of mass exist. We find that the closest match to this field which exists in both suites is M_{200c}: the total mass enclosed within an overdense region equal to 200 times the critical mass density of the universe. These mass accretion histories show similar behaviour in TNG-Dark and Uchuu, and produce a similar overall mass distribution, and consequently the predictions of the nerual network are similar when trained using M_{200c} and with the SubLink mass. This is our choice of field representing halo masses in Uchuu. Halo mass dependent calculations such as overdensities are also calculated using this quantity.

Resolution Differences

Table 5.2.2 shows that the mass resolution of Uchuu is an approximate order of magnitude weaker than the TNG300 simulations. This results in very poor resolution of low mass haloes in Uchuu, and the absence of low mass first order satellite haloes, following quality cuts such as the omission of samples whose specific mass accretion gradients differ from the mean by more than 5σ, serves to limit the number of satellites below a subhalo mass of $\sim 2 \times 10^{11} M_{\odot}$, which we show in Figure 5.2.1. We are left with a modest sample of low mass satellite haloes but a clearly different distribution of subhalo masses below $\sim 2 \times 10^{11} M_{\odot}$.

Another important detail is the difference in time resolution of the Uchuu snapshots compared with TNG. Uchuu has half as many snapshots as TNG, as shown in Table 5.2.2,

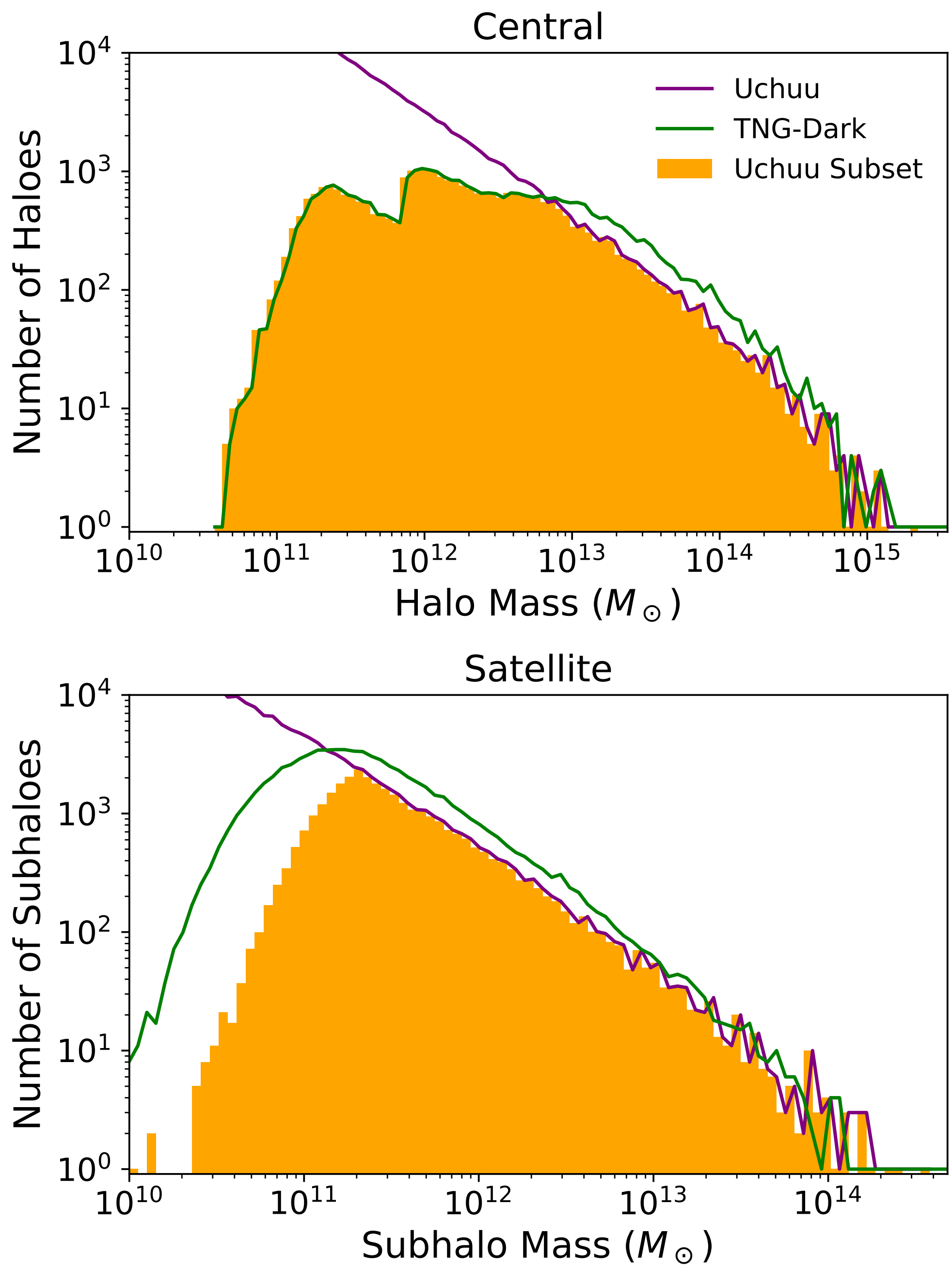

Figure 5.2.1: Distributions of central halo and satellite subhalo masses in the complete Uchuu forest (purple) and our cross-matched TNG-Dark sample (green). Sampling the former distribution according to the latter, we obtain the distribution of Uchuu haloes used in this work (orange). The distribution of central haloes is similar to the TNG-Dark data, however the lack of well-resolved satellite subhaloes at low mass serves to bias the sample distribution of satellite subhaloes in our Uchuu dataset.

Simulation Properties				
Suite	Simulation	$L_{\rm box}$ (Mpc)	$m_{\rm DM}$ ($M_\odot$)	$N_{\rm snap}$
TNG - Hydro	TNG100-1	110.7174	7.5×10^6	100
	TNG300-1	302.6277	5.9×10^7	100
TNG - Dark	TNG100-1-Dark	110.7174	8.9×10^6	100
	TNG300-1-Dark	302.6277	7.0×10^7	100
Uchuu	Uchuu	2952.4653	4.8×10^8	50

Table 5.2.2: Properties of the simulations used in this work, relating to their volume and resolution. $L_{\rm box}$ indicates the side length of the comoving cubic volume of the simulation, $m_{\rm DM}$ represents the mass of a single dark matter particle, i.e. the smallest possible mass in the simulation, and $N_{\rm snap}$ represents the total number of time snapshots in the simulation. Simulations are grouped in this table according to the suite of simulations (e.g. TNG-Dark) from which they originate. For all of these simulations, mass resolution is decreased as the volume of the simulation and number of particles are increased.

however the temporal separation of snapshots is smaller than TNG for $2 \lesssim z \lesssim 4$ and larger otherwise. All temporal properties, regardless of their simulation, are linearly interpolated over the same 33 snapshots in TNG, yet the sparse nature of these snapshots means that information on short timescales may be lost.

5.3 Halo Properties

For each simulation used in this study, we apply the same algorithm to acquire the input data, however there exist key differences between the definitions of these properties, depending on the model in question; and differences in their statistics due to the removal of baryons or the resolution of the data. In this section, we discuss the similarities and differences between key input properties of the neural network in each simulation, and how this may influence the predictions of the neural networks.

We compare properties for central and satellite haloes in bins of different mass and accretion history to investigate these differences in separate evolutionary regimes. For centrals, we characterise halos of different accretion histories according to the specific mass accretion gradient, denoted β, which Montero-Dorta et al.[231] show to be clearly correlated with gas fraction, quenching timescale and assembly bias in TNG300. We use this quantity for quality cuts in both central and satellite datasets, however it is not so useful for modelling satellite subhalo histories due to the distinct nature of the modes of accretion in their central and satellite phases. For satellites, we use the scaled accretion time a_{max}, which, like β, is derived explicitly from the subhalo mass accretion history, and is shown by Shi et al.[229] to have similar connections with satellite galaxy properties.

In the following figures which compare haloes of different histories, subplots are arranged such that each column is a quintile of (sub)halo mass, with higher mass bins towards the right; while each row is a quartile of accretion gradient, with the steepest accretion histories on the top row. For satellites, higher a_{max} values are placed on the top row, such that in both cases, haloes with the earliest half-mass formation times are placed on the top row. These percentiles are taken from the TNG-Hydro data. This convention is adopted for all figures of this format, such that in each case the earliest-forming, gas-poor halos or subhalos appear on the top row. For simplicity, we show only odd-numbered mass quintiles, and only the first and fourth quartiles of mass accretion gradient.

5.3.1 Mass Accretion History

Figures 5.3.1 and 5.3.2 show the median and $15^{\mathrm{th}} - 85^{\mathrm{th}}$ percentile ranges of mass accretion histories in bins of final (sub)halo mass and mass accretion gradient. In most bins, these mass accretion histories are very similar. The differences between simulations arise in low mass and shallow gradient cases, where the amplitude of Uchuu mass accretion histories are reduced, and in the low mass regime is increased for TNG-Dark haloes at early times.

The enhancement of mass for early, low mass haloes is likely due to the abscence of baryonic driven outflows, which will have a substantial effect on haloes of a high gas fraction. The weaker accretion rates in the Uchuu simulation, however, are a result of the lower mass resolution of the simulation. In shallow gradient bins, there is a clear flat profile to Uchuu accretion histories, being somewhat reminiscent of TNG samples which were discarded by

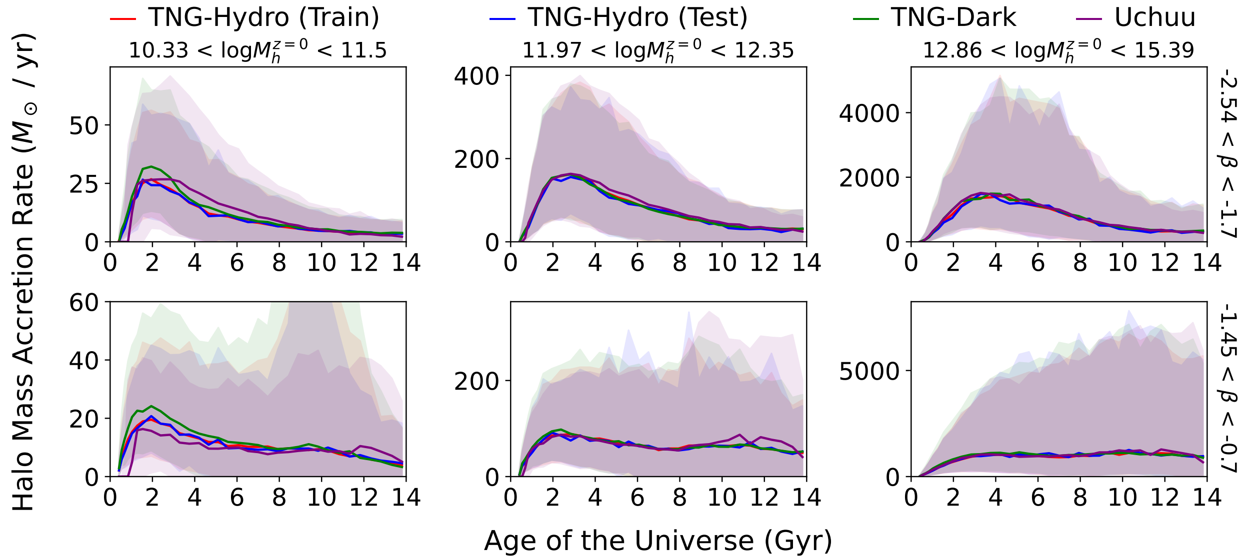

Figure 5.3.1: Mass accretion histories of central haloes in bins of halo mass, increasing along the horizontal axis, and specific mass accretion gradient, decreasing in steepness down the vertical axis. Solid lines show the median mass accretion history per bin, while shaded regions show the 15$^{\text{th}}$ and 85$^{\text{th}}$ percentiles. Mass accretion histories are shown for the training (red) and testing (blue) datasets of the TNG-Hydro simulations, alongside TNG-Dark (green) and Uchuu (purple).

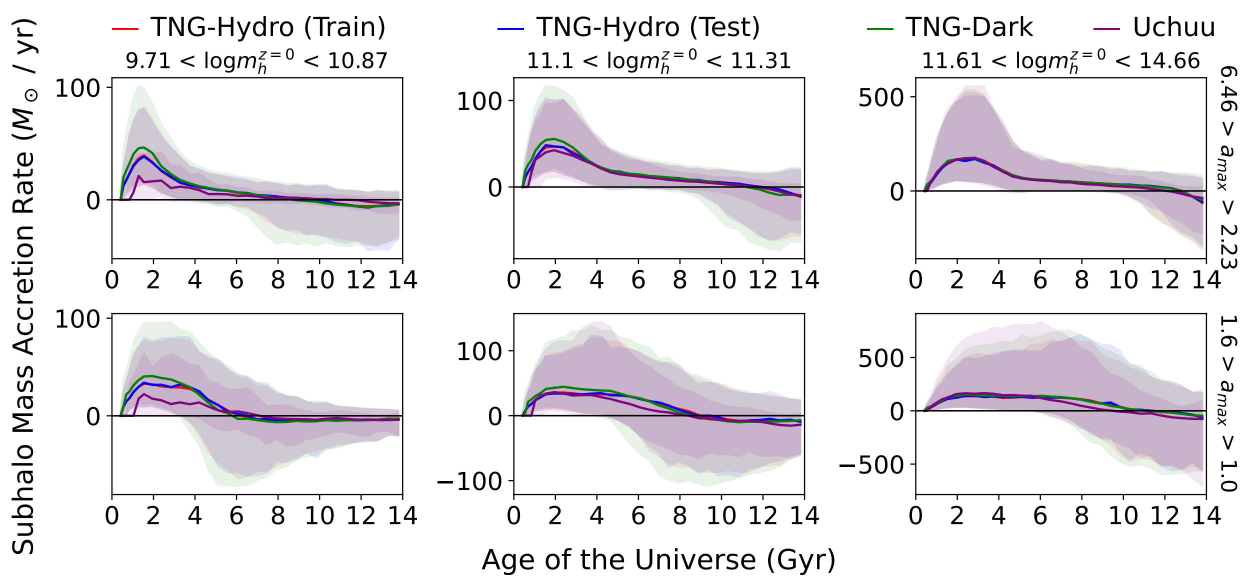

Figure 5.3.2: Mass accretion histories of satellite subhaloes, binned in the same manner as in Figure 5.3.1, using satellite subhalo mass in place of central halo mass, and scaled accretion time in place of specific mass accretion gradient. One key difference in the properties of these accretion histories is their approach towards zero or negative accretion, which rarely happens for central haloes. Subplots show the different times at which subhalo growth stops in different growth regimes.

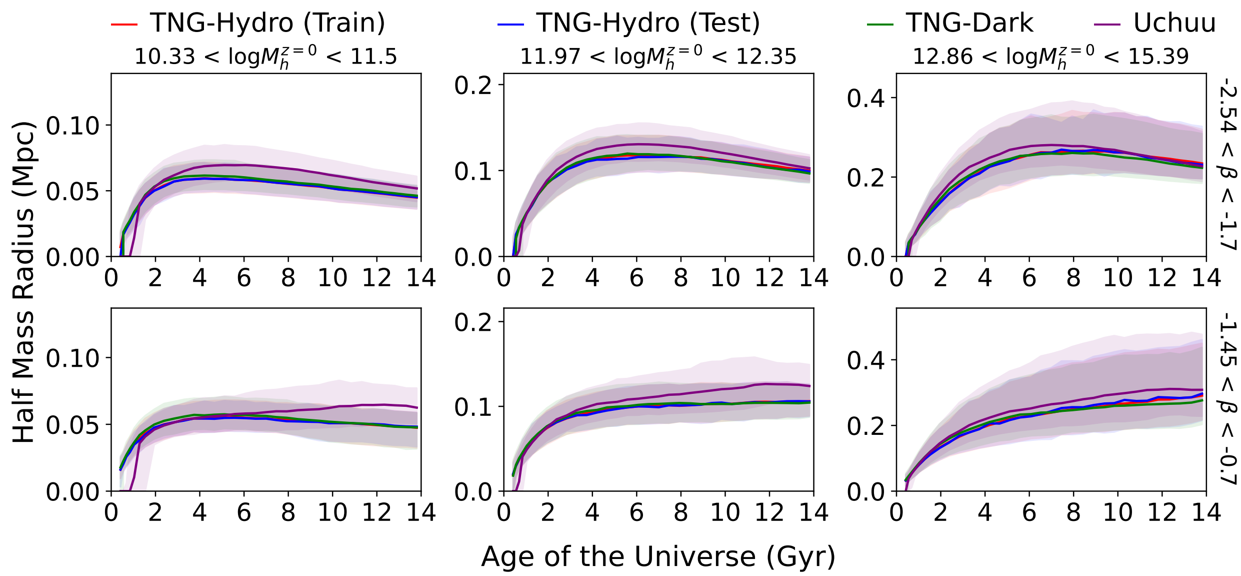

Figure 5.3.3: Growth of the half-mass radius of central haloes over time, shown in the same tabular format as Figure 5.3.1, with the same bins of halo mass and accretion gradient.

our quality cuts.

With Uchuu being of lower mass resolution than any of the TNG simulations, the resolution of low mass haloes will be poor. In the lowest mass quintile, galaxy evolution may be poorly predicted from the start of the galaxy's evolution due to the sensitivity of the star formation algorithm in TNG to the number of dark matter particles[238]. Haloes with shallower accretion gradients generally form at later times[231], as is the case for our Uchuu data, and therefore those in any given mass bin will be of lower mass at any nonzero redshift. They are therefore subject to a similar resolution effect. For satellites with shallow accretion, there is a noticeable decline in the median accretion rate compared with TNG, probably due to the absence of low mass objects being accreted onto the halo.

5.3.2 Half-Mass Radius

The quantity which differs most starkly in the Uchuu data, however, is the halo half-mass radius, shown for central and satellite haloes in Figures 5.3.3 and 5.3.4 respectively. There is a clear deviation in the size evolution of haloes in Uchuu compared with TNG in low mass and shallow gradient bins. However, they are also subtly but noticeably larger in most other

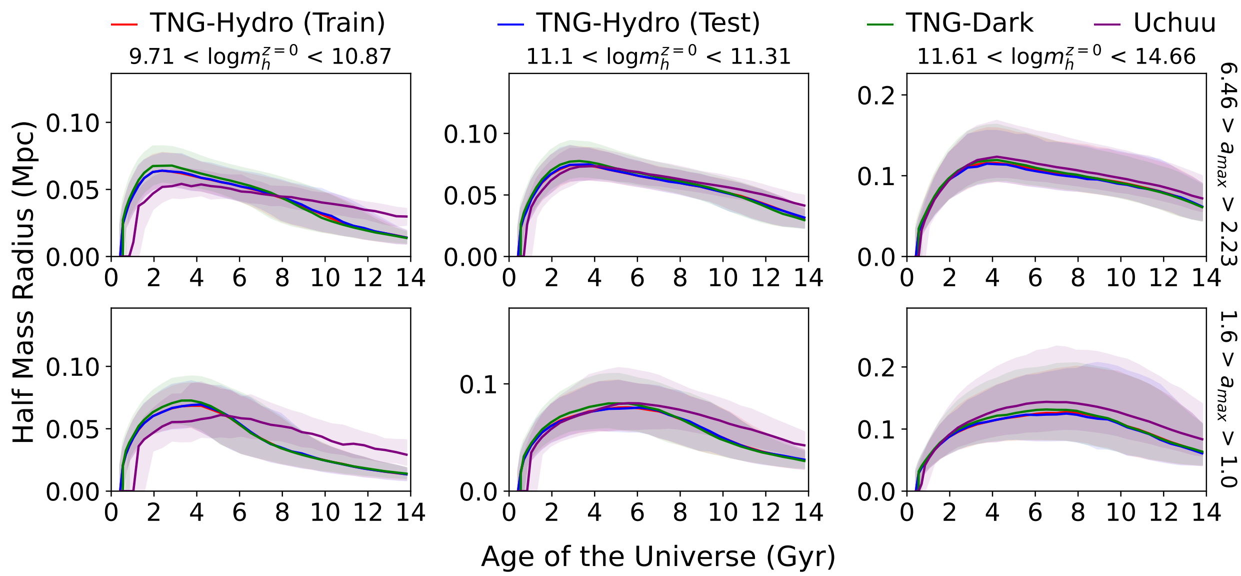

Figure 5.3.4: Growth of the half-mass radius of satellite subhaloes over time, shown in the same tabular format as Figure 5.3.2, with the same bins of halo mass and scaled accretion time.

bins and at most times.

In TNG-Dark, the half-mass radii of central haloes are subtly smaller in high mass bins than in the other simulations. This may be explained by the results of Haggar et al.[224] and Riggs et al.[225], who show that the number density of haloes which are gravitationally bound to a galaxy group or cluster is underpredicted in dark simulations with respect to their hydrodynamical counterparts, within two virial radii of the cluster or galaxy group. This is caused not only by the high density of baryons in the large object's centre but the effect this has on its own density profile.

Chua et al.[239],[240] show that the radial profiles and asymmetries of haloes in TNG, and in the original Illustris simulation, are affected significantly by baryons. According to our data, however, this does not affect the size history in TNG-Dark when compared with the full physics simulation. In CT23 we thoroughly compared quantities in hydrodynamical and dark TNG simulations at all times to ensure that they could be used in a dark matter simulation. In Uchuu, however, the half-mass radius is defined in terms of virial mass, unlike the SubFind result in TNG.

The SubFind algorithm used in TNG defines the boundaries of a subhalo according to a contour of constant density meeting a saddle point in the local density field[241], and the half-mass radius is defined as the radius which encloses half of the mass enclosed by this boundary. In Uchuu, a spherical region whose radius encloses the virial mass is assumed[226], and therefore the calculation of half-mass radius is dependent on the halo density profile. The Rockstar halo finder[237] used in Uchuu calculates this profile from the cumulative binding energies of halo particles, and therefore is sensitive to the NFW profile[242] concentration parameter.

Zhao et al. [243] show that, in N-body simulations, the increase of the NFW concentration parameter over time is scaled by the time of formation of 4% of the final halo mass; while Prada et al. [244] show that halo concentration is additionally sensitive to fluctuations in the linear density field on the scale of the halo's mass. These are both quantities which will depend on the resolution of the simulation, however this effect on the growth of halo concentration is most likely reflected in the more extended mass distributions at late times, as well as the smaller radii of low mass haloes at early times. Ishiyama et al. [226] also show that in the smaller, higher resolution Shin-Uchuu simulation, while the halo mass functions do not differ significantly between the two Uchuu models, the mass-concentration relation does grow differently between the two simulations, illustrating that this difference in structure is in fact a resolution effect. Morphological halo quantities such as virial velocity and axis ratios are additionally dependent on the gravitational softening scale, which, being larger in Uchuu than in TNG, would result in a flatter $M_h - v_{\max}$ relation at low mass, and thus a similarly distorted mass-concentration relation[245].

For both central and satellite haloes, it is the youngest, smallest haloes which are most affected by the difference in halo concentration owing to Uchuu's lower resolution. In each case these haloes are growing from low mass progenitors and thus more likely to suffer from delayed growth in the Uchuu data. Furthermore, Hoffmann et al. [246] argue that the subhalo finder algorithm makes little difference to the inferred halo shape, provided that the subhalo is comprised of adequately many particles, which low mass, low resolution haloes do not possess. We therefore stress that the delayed growth of the half-mass radius is a resolution effect and not dependent on the algorithm which defines the halo. Yet on the other hand, Onions et al.[247] argue that Rockstar is superior to most algorithms in identifying halo substructures in dense halo centres, which can affect the half-mass radius value. The fact that some graphics in Figures 5.3.3 and 5.3.4 show a larger range of radii in Uchuu for all times, suggests that this is also a relevant factor.

5.3.3 Circular Velocity

The proxy for the virial circular velocity used in our model depends on the above two quantities. It was defined specifically due to extreme differences between the maximum orbital velocities in TNG-Hydro and TNG-Dark, particularly at early times.

Due to the discrepancies we have shown for these quantities in the Uchuu data, the median values for this proxy are slightly too small, but the shape of this median curve remains congruous to its TNG counterparts. In low mass bins, however, this proxy is overestimated in TNG-Dark; most likely as a result of excess mass accretion. This also happens to high mass centrals, however to a smaller extent and owing instead to smaller half-mass radii.

5.3.4 Cosmic Environment

In CT23 we characterised the environmental histories of central and satellite haloes using subhalo overdensities at each TNG snapshot, and devised a radial skewness quantity which characterised the interaction histories of these haloes. While the mass and structure quantities discussed thus far were deemed important to predicting the star formation histories of the haloes' galaxies, these environmental properties were shown to predict their chemical enrichment.

As previously discussed, the definitions of halo substructure differ in the Uchuu catalogue, using Rockstar haloes in the place of subhaloes, and thus the overdensity and skew calculations are based on halo tracers. Overdensities in Uchuu are larger than those in TNG-Dark, which can be seen in Figure 5.3.5. In a lower resolution simulation, the calculation of environmental quantities is more sensitive to edge effects of the calculation volume, and the radially symmetric definition of a halo in Uchuu can affect the centre of mass of a given ob-

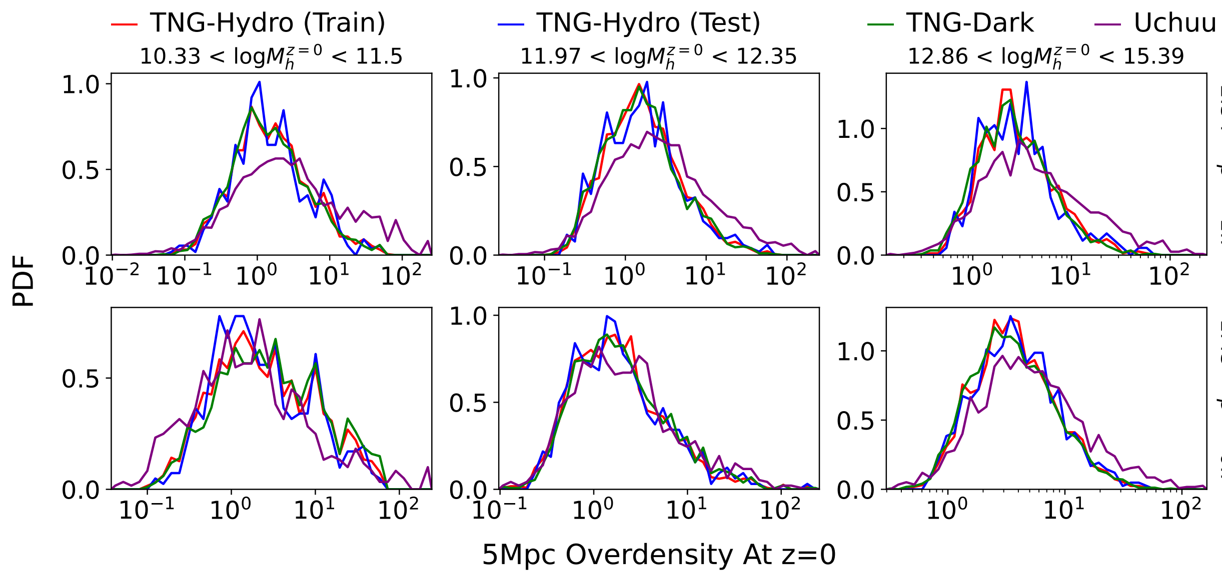

Figure 5.3.5: PDFs of the dark matter overdensities enclosed in a 5Mpc spherical region surrounding zero-redshift haloes in the four simulation datasets. These histograms show similar distributions of overdensity for a given mass and accretion history, except for typically larger densities in the Uchuu simulation.

ject, which may both be contributing factors to the differences in overdensity. Skews, on the contrary, are not noticeably affected by the resolution difference. As this is a mass-weighted quantity and thus independent of any scaling, it is likely that the difference in overdensities is a result of the mass content of the relevant density tracers.

5.4 Galaxy Predictions

Having outlined the causes of differences between haloes and environments in each simulation, we explain in this section how the implementation of these distinct variables into the neural network changes important results. We discuss the effects that the simulation differences have on the direct predictions and derived halo-galaxy relations, providing a physical explanation into differences in the results.

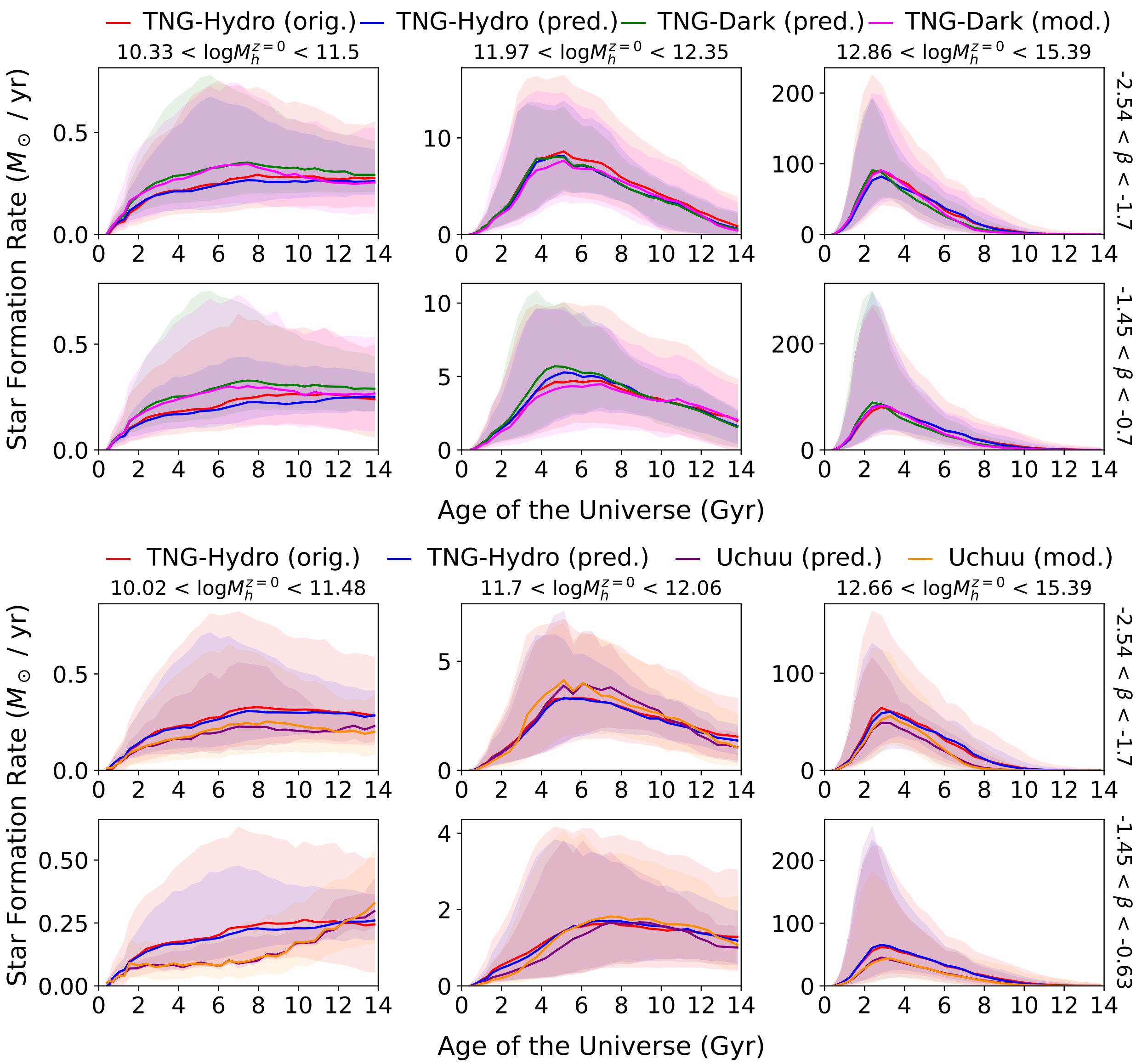

Figure 5.4.1: Star formation histories of central galaxies in bins of halo mass and specific mass accretion gradient. In low mass bins, there is an overprediction of star formation rates in TNG-Dark, and an under-prediction in Uchuu. In higher mass bins, this difference becomes smaller. The stochastic modification has shown here to reshape the shape and variance of N-body predictions to more closely resemble the orginal TNG data, improving upon the predictions in the hydrodynamical simulation as well. However, the modification cannot amend the poorly predicted star formation histories in low mass objects, as the predicted Fourier Transforms used by the modification are similarly affected by mass resolution.

5.4.1 Star Formation History

Figures 5.4.1 and 5.4.2 show the median and 15^{th} – 85^{th} percentile ranges of predicted star formation histories in bins of final (sub)halo mass and mass accretion gradient, including the modified star formation histories in the N-body simulations. Each of these figures show that galaxies in high mass and fast accretion bins are typically well matched, but in low mass bins the differences are stark. Star formation rates in TNG-Dark are over-predicted, while in Uchuu they are underpredicted.

From Figure 5.4.2 it may appear that the majority of Uchuu galaxies are severely underpredicted in their stellar mass. This is misleading due to the absence of low mass haloes, discussed in Section 5.2.2. The two lowest mass quintiles of TNG haloes contain $\sim 12\%$ of satellite galaxies in Uchuu. However, these star formation histories are nonetheless severely underpredicted. Their halo histories have already shown that these objects accumulate their mass later and more slowly than their TNG counterparts. This can be seen clearly for satellites in Figures 5.3.2 and 5.3.4.

In Table 5.4.1 we show Spearman correlation coefficients between halo and galaxy properties in a narrow, low mass bin for central and satellite objects in Uchuu. Within individual bins at low mass, the final stellar mass is correlated strongly with the proxy for circular velocity, which in the Uchuu data is subtly smaller than its TNG equivalents. Weaker correlations exist with stellar mass and half-mass radius and overdensity, which are also somewhat different in Uchuu.

However, particularly for low mass satellites in Figure 5.4.2, underpredicted star formation histories are clearly correlated with similarly undermined mass accretion histories in Figure 5.3.2, which will have had a causal effect on their galaxy growth. Mass accretion at early times is particularly important for the acquisition of star-forming gas, which would explain the lack of subsequent star formation in these predictions. In the TNG simulations, low-mass haloes are particularly gas-rich[248] and so the star formation at early times is likely to be sensitive to the lack of accumulation of low mass progenitors.

Comparing the network's performance on TNG-Dark with TNG-Hydro predictions, we

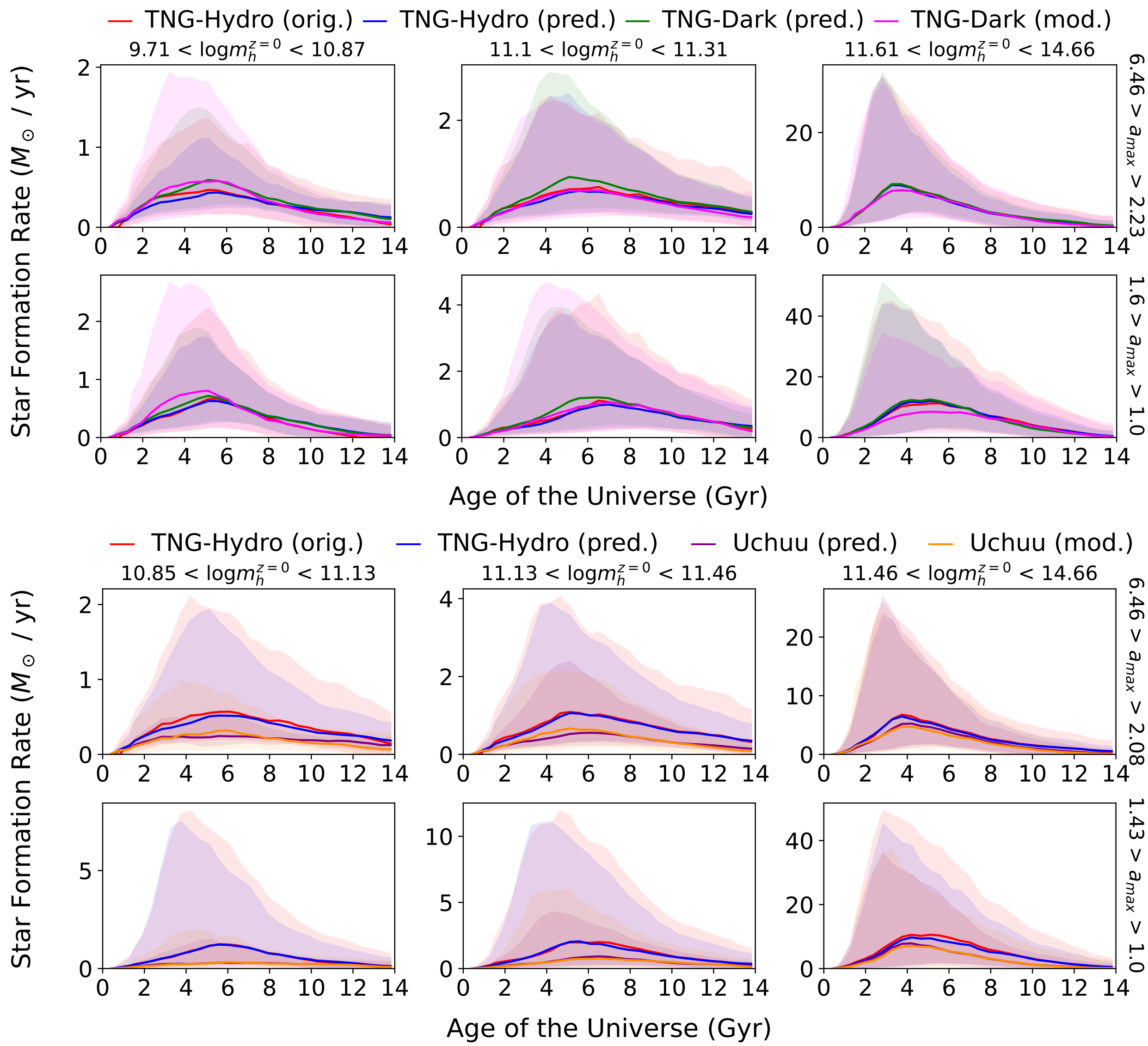

Figure 5.4.2: Star formation histories of satellite galaxies, binned in the same manner as in Figure 5.4.1. This figure shows apparently poor predictions of star formation histories of low mass galaxies in Uchuu, however many of these bins are underpopulated by the Uchuu data and contain low quality haloes.

Input Variable	Stellar Mass		Metallicity	
	Central	Satellite	Central	Satellite
Circular Velocity	0.783	0.637	0.693	0.361
Half-Mass Radius	-0.566	-0.573	-0.602	-0.314
Overdensity	0.176	0.298	0.174	0.142

Table 5.4.1: Spearman correlation coefficients of multiple halo properties with stellar mass and metallicity, in narrow, low halo mass bins. All quantities shown here are considered at $z = 0$ in the Uchuu simulation. We evaluate these within halo mass ranges of $11.77 < \log_{10} M_h^{z=0} < 11.97$ for centrals, and $10.9 < \log_{10} m_h^{z=0} < 11.1$ for satellites.

see that these star formation histories are enhanced rather than suppressed, which correlates with similarly excessive mass accretion rates. Sorini et al.[249] show that gas accretion and stellar feedback processes have their greatest influence on the size and shape of haloes and large scale structure at higher redshifts, and that stellar feedback is the principal cause of suppression of the star formation of low mass objects. The lack of stellar feedback in dark simulations will have limited the effects which result in halo mass loss, which are particularly prominent effects for low mass objects. The excess mass accretion in TNG-Dark will have resulted in excess star formation being predicted.

The stochastic modification to our star formation and metallicity histories, explained in detail in BTC24, introduces fluctuations to individual galaxies by drawing random phases from high frequency modes, which are scaled in amplitude according to a predicted Fourier Transform of the sample's SFH and ZH. By training the neural network introduced in CT23 to predict the Fourier amplitude of each SFH and ZH, while changing none of the input variables or aspects of the network's design, we predict Fourier amplitudes which resemble those of the original TNG data; unlike the fiducial predictions of the original network, which lose much of this information at high frequencies. A comparison of the Fourier Transforms derived from central star formation histories in an intermediate mass bin is shown in Figure 5.4.3.

With the stochastic modification, the additional fluctuations in the star formation histories have recovered enough information that the median and variance of the predictions in both N-body simulations are improved, with the dataset more akin to the target TNG data. However, the correction is less effective for low mass centrals and satellites, and does not resolve the severe offsets seen in Uchuu galaxies.

As star formation histories are poorly predicted for under-resolved Uchuu galaxies, so are their Fourier Transforms, making mass resolution crucial to the stochastic modification as much as the machine learning model itself. We illustrate this in Figure 5.4.3, comparing the predicted Fourier amplitudes of each simulation with the desired result from the raw TNG data. For this bin in β and M_h, we can see that the Uchuu Fourier amplitudes are reduced across all frequencies. With the stochastic method sampling high frequency modes with suppressed amplitudes, the same variable star formation history will be lacking in under-resolved Uchuu data, regardless of the efficacy of the stochastic method.

In BTC24, several post-processing filters were applied to mitigate unphysical results in the modified data. For Uchuu and TNG-Dark, an additional threshold is necessary due to unphysical distortions arising from less accurate predictions of the Fourier Transform amplitudes. To address this, we impose a threshold that limits the relative difference between the corrections and the predictions to less than a factor of 1.5, which is typical of the modified TNG-Hydro data. Any larger deviations are replaced by the fiducial predictions to avoid further distortions.

For the TNG-Dark data in this mass/gradient range, the Fourier Transforms resemble the TNG-Hydro predictions, suggesting that the same stochastic data as in BTC24 may appear in TNG-Dark, yet as discussed below, the modifications in TNG-Dark are also subject to certain resolution effects.

Numerically integrating the SFHs of each simulation result in self-consistent and accurate SHMRs, shown in Figure 5.4.4. Between predictions in the TNG-Hydro and TNG-Dark simulations, there is little noticeable difference between the relations, other than slightly reduced scatter in the dark predictions. For Uchuu, the difference in stellar and halo mass distributions is apparent, and this is most profoundly so at low masses. From intermediate to high masses, there is a slight underprediction of mass and scatter, but the Uchuu SHMR remains very well matched to the predictions in TNG-Hydro and TNG-Dark.

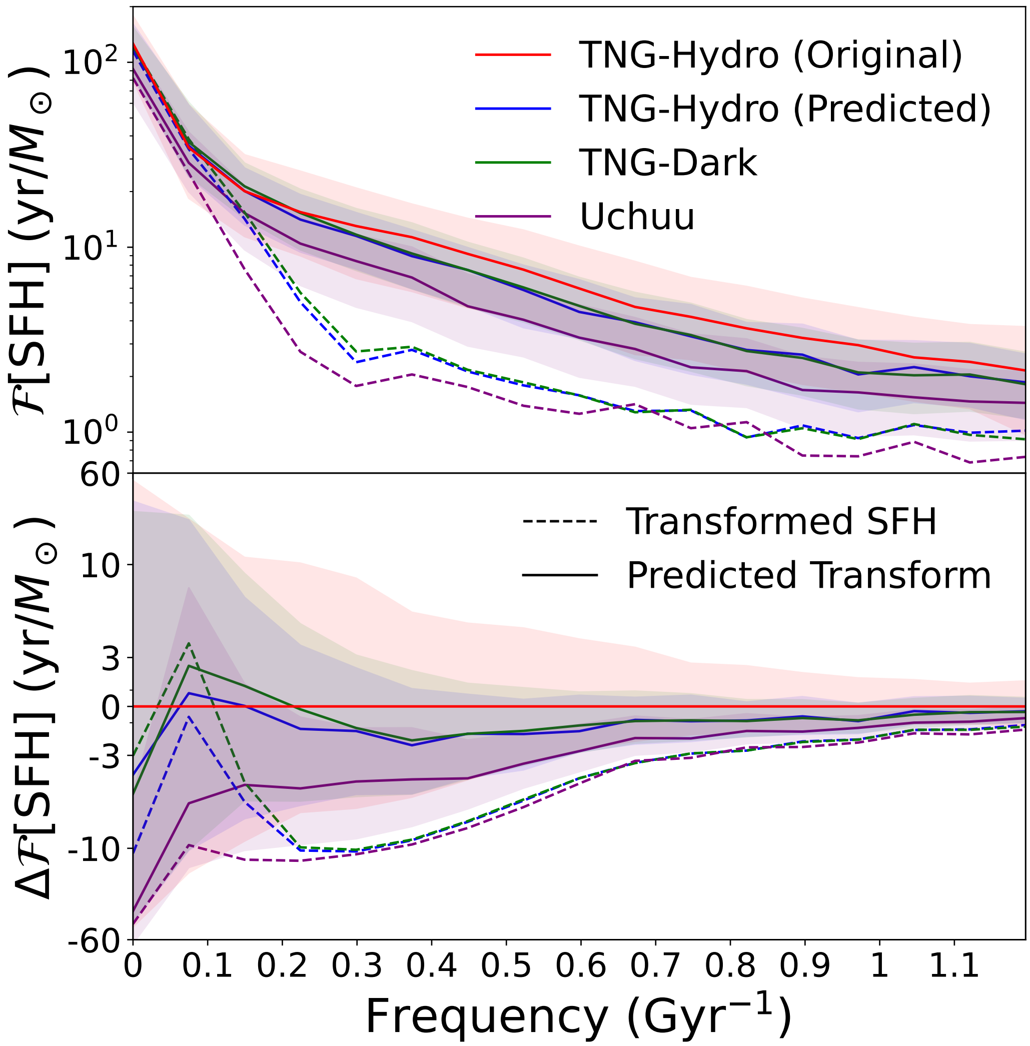

Figure 5.4.3: A figure comparing the median and interquartile ranges of predicted Fourier Transforms of star formation histories, for central galaxies where $11.97 < \log_{10} M_h^{z=0} < 12.35$ and $-1.45 < \beta < -0.7$, corresponding to the centre-bottom panel in Figure 5.4.1. The top panel of this figure shows the Fourier Transforms themselves, while the bottom panel shows the difference in the Fourier amplitude with respect to the median target amplitude. Where applicable, dashed and solid lines indicate the Fourier Transforms of the star formation histories predicted by the neural network, and the predicted Fourier Transforms when the network is trained to predict these, respectively. This illustrates that for all datasets, the median Fourier Transforms are all of similar shape to the target TNG-Hydro data (red) when predicted by the neural network. The TNG-Dark result (green) is similar enough to the TNG-Hydro predictions (blue) that the modified data in TNG-Dark will be just as accurate as in our companion paper. On the contrary, the Uchuu data (purple) has clearly suppressed amplitudes at all frequencies, and does not align with the target distribution. This suggests that star formation modes which influence our results will be absent from some of our Uchuu predictions, regardless of the power of the stochastic amendment.

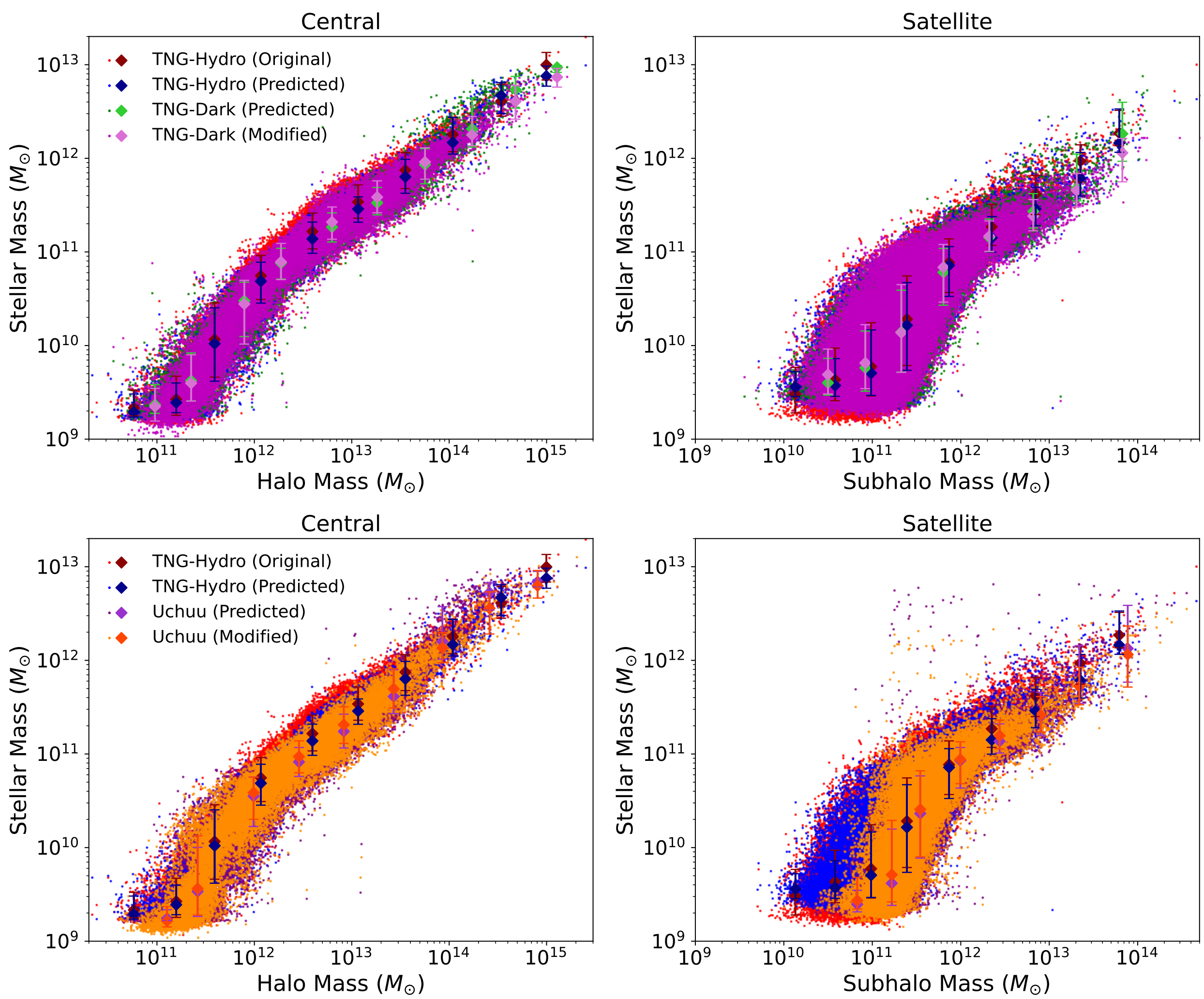

Figure 5.4.4: The graph displays the numerical SHMR of central galaxies (left) and satellite galaxies (right) based on their star formation rates. Each galaxy is represented by a data point, while the error bars indicate the median and 15^{th} and 85^{th} percentiles of stellar mass within a specific halo mass range. The comparable shapes of these relationships suggest an accurate prediction of the star formation histories.

In both the central and satellite SHMRs, particularly at intermediate masses, the stochastic modification makes subtle improvements to the average height of the relation, as well as marginally increasing the scatter. The fluctuations added by the modification add a component of stellar mass which may be attributed to events such as galactic winds or minor mergers; while the cycling of gas through the CGM acts on sub-Gyr timescales for high-mass galaxies in dense environments[250], stellar feedback is more likely to disrupt the shallow potentials of low mass galaxies and contribute to greater mass loss on short timescales[251;252]. Despite the variety of processes acting on different timescales in different mass regimes, the fluctuations implemented by the modification are rarer for the most massive galaxies, and smaller relative to the total mass, making little change to the modified stellar mass.

At low masses, we see that the distribution of stellar masses in Uchuu is significantly biased towards low masses, owing to the compromised star formation rates. Low mass satellite haloes have already been discarded by our cuts, resulting in a clearly distinct halo mass distribution, yet there are still a large number of objects below $\sim 2 \times 10^{11} M_\odot$ with severely miscalculated stellar mass. Satellite objects within this mass range clearly cannot be trusted.

Low mass satellites are additionally made impractical due to a limitation of the resolution correction presented in CT23. The correction consists of a $z = 0$ halo mass dependent ratio between star formation and metallicity histories in TNG100-1 and the lower resolution run TNG100-2, with a mass resolution consistent with TNG300-1; to adjust TNG300-1 galaxies to match TNG100-1 data prior to their simultaneous use as training data. For satellite galaxies, however, the correction ceases to be accurate when satellite galaxies of distinct evolutionary geometries (e.g. early-forming, quenched and late-forming, star-forming galaxies), are combined as one correction for all low mass satellites; as well as when the TNG100-2 galaxies on which the correction is based are affected by low sample size and resolution.

The correction was based on a method used by Pillepich et al.[238], who find that the radius enclosing star particles influenced their corrected data. As Nelson et al.[214] find higher gas densities in TNG300-1 than in TNG100-2, it is possible that differences in the density of gas surrounding satellites results in differences in ram pressure stripping and thus stellar mass profiles. These are two effects which reduce the validity of the correction for satellites, particularly at low mass. It is such a substantial difference for low mass satellites that these samples are ignored when applying the stochastic correction.

In high mass bins, the majority of stellar mass is formed earlier than in TNG-Hydro, for both dark simulation suites. This is illustrated by Figure 5.4.5, which shows mass weighted ages of galaxies in bins of stellar mass, and shows that high mass galaxies are biased towards older ages. In part, this can be seen by a sharp increase in both halo and stellar mass at earlier times, but in Figures 5.4.1 and 5.4.2 the following decline in star formation rate happens sooner. The stochastic modification corrects for excess early star formation, indicating that the missing stellar feedback effect is encoded in the SFH power spectra; yet the modification does not influence the quenching tail, which is a smoother feature and so potentially unseen by the modification due to frequency-dependent phase selection, while unphysical noise in the quenching tail is eliminated by our post-processing filters (see Section 4.3.3 of BTC24 or Chapter 4). In Uchuu, the mass-weighted ages of intermediate mass galaxies are shifted by a lesser degree than in TNG-Dark, potentially due to differences in the accuracy of Fourier amplitudes. For TNG-Dark, star formation histories are initially aligned with TNG-Hydro, but these star formation rates decline and line up with the lower SFH in Uchuu. The quenching of these galaxies is therefore more efficient than their hydrodynamical equivalents.

Davies et al.[248] show that the expulsion of the circumgalactic medium in TNG and the Eagle simulations[253] is correlated strongly with the central black hole mass of the galaxy, which influence the subsequent specific star formation rate, and act on timescales of multiple gigayears[254–256]. Bluck et al.[257] show that central velocity dispersion, effectively measuring the AGN mass, is a critical parameter for galaxy quenching in MANGA observations, which correlates with the circular velocity and half-mass radius of the halo. In TNG, these internal feedback quenching mechanisms dominating central galaxy quenching, alongside environmental effects dominating satellite quenching, are shown to qualitatively agree with other hydrodynamical and semi-analytic models, and with low-redshift SDSS data[258;259].

We have shown in Figure 5.3.3 that the half-mass radius of high mass haloes is underestimated in TNG-Dark, leading to overestimated halo concentration. This velocity dispersion measures the dynamics of the halo centre on sub-kiloparsec scales, and therefore the halo con-

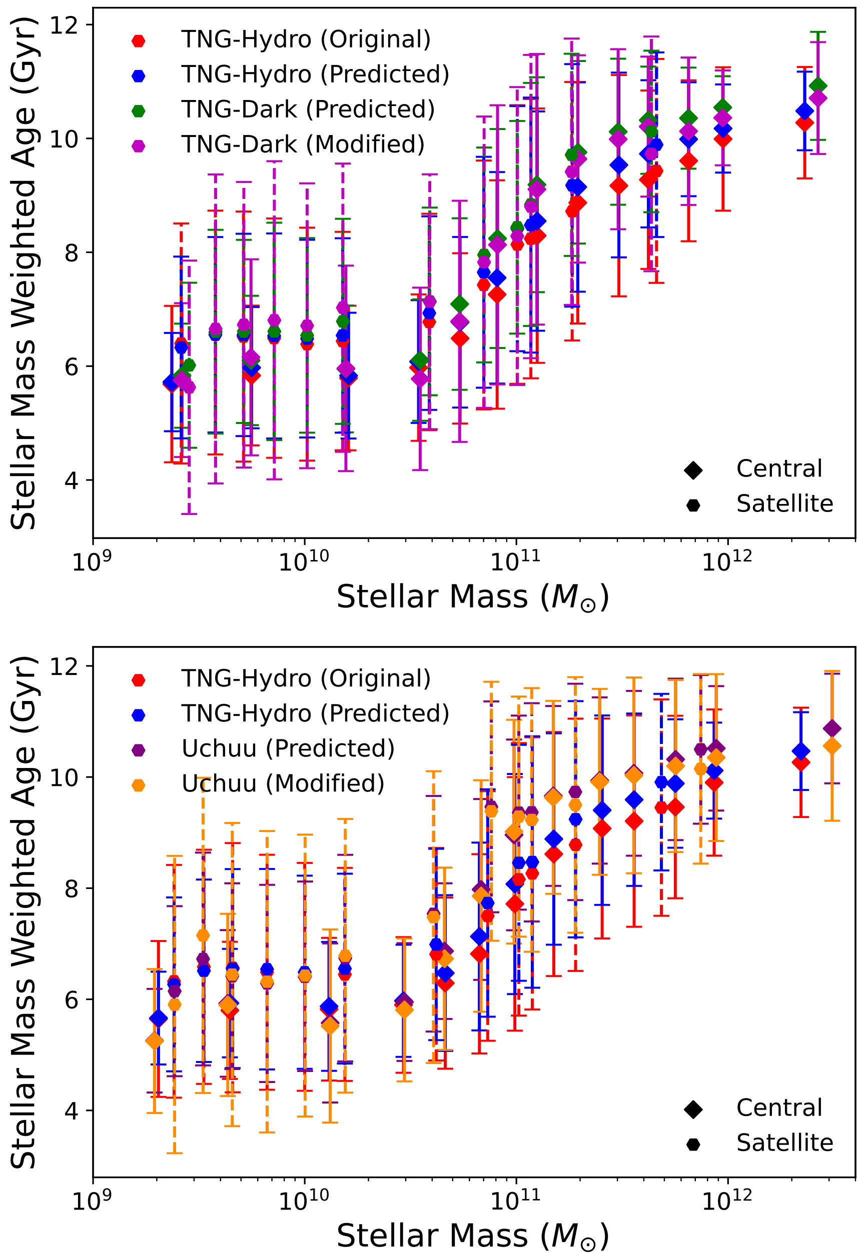

Figure 5.4.5: The figure depicts the mass-weighted ages of central and satellite galaxies, showing TNG-Dark results in the top panel and Uchuu results in the bottom panel. The ages are plotted against the predicted stellar mass, with the median and interquartile range of ages shown in various mass bins. The plot reveals a precise recovery of the age-mass trend in the dark simulations. There is a bias towards higher ages in the dark simulations which grows with mass, partially corrected by the stochastic modification.

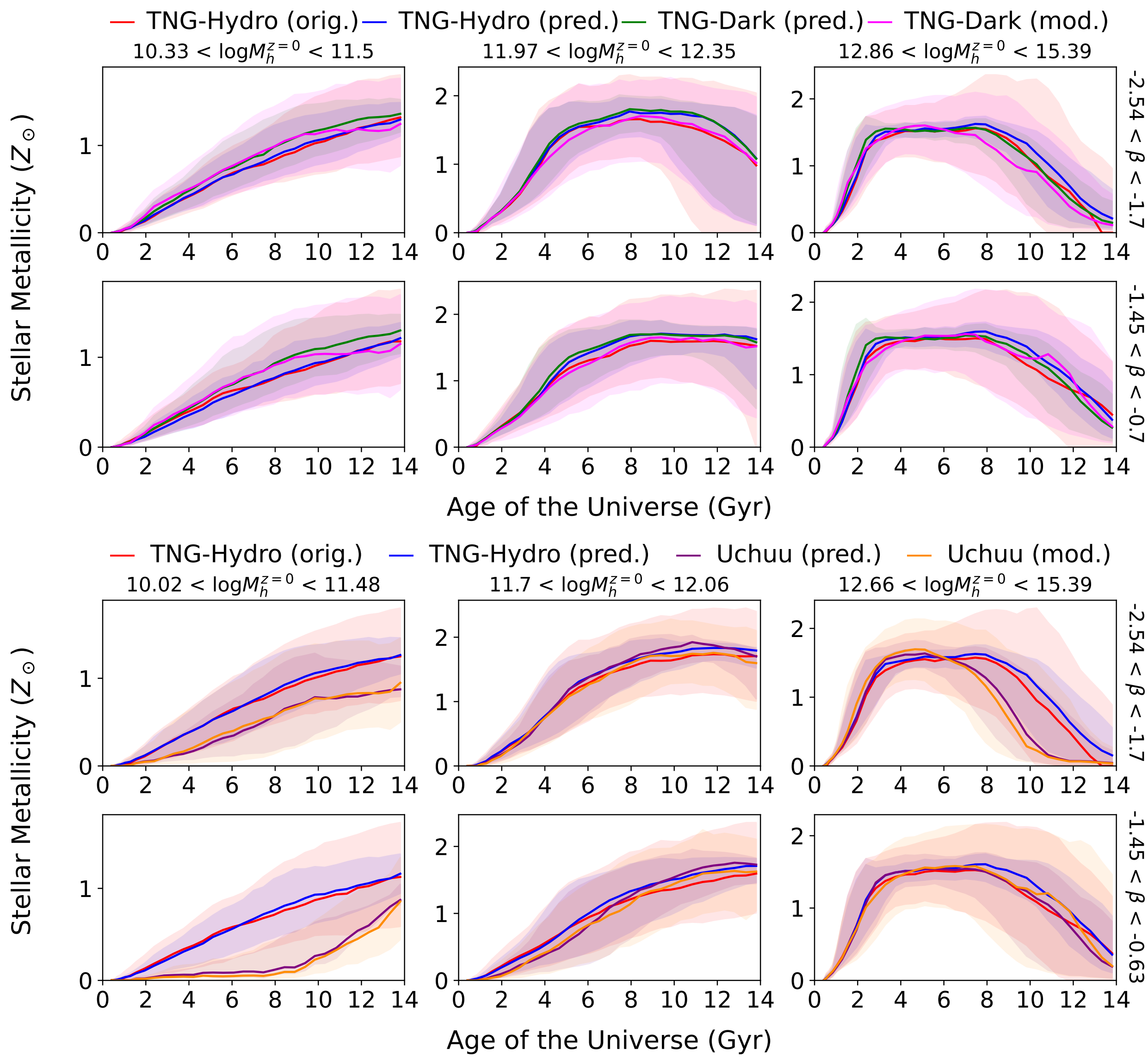

Figure 5.4.6: Stellar metallicity histories of central galaxies in bins of halo mass and specific mass accretion gradient.

centration relates to the density of this region, which in galaxies of this mass are most likely to be dominated by an AGN. Overestimating the concentration can cause overprediction of AGN feedback, thereby quenching the galaxies sooner.

5.4.2 Metallicity History

Figures 5.4.6 and 5.4.7 show the median and $15^{\text{th}} - 85^{\text{th}}$ percentile ranges of predicted metallicity histories. These show a clear failure of the model to predict the metallicity histories of low mass objects in Uchuu, as well as underpredicted results in most satellite galaxies. There is nonetheless a good agreement between Uchuu and TNG with most intermediate to high mass central haloes.

The stochastic amendment introduces a significant improvement to the median and variance of metallicity histories. Like the star formation histories, it is likely that the network-predicted Fourier Transforms contain information on the frequency of chemical enrichment events on short timescales, such as merger-driven star formation. The simultaneous recovery of the variability in star formation and metallicity histories results in a more realistic metallicity distribution with respect to the fiducial network predictions, as shown in BTC24.

We see similar characteristics when comparing metallicity histories seen in TNG-Hydro, TNG-Dark and Uchuu to what we showed for star formation histories in Section 5.4.1. The suppression of metal synthesis in low mass galaxies can be reconciled with the lack of early accretion, particularly as the gas and metal content of these objects' progenitors play an important role in early metal synthesis. The enhanced metal synthesis in low mass TNG-Dark galaxies may also be explained by the abscence of stellar feedback, where stars retain more of their mass and thus produce metals more efficiently.

In narrow mass bins, we find similar but weaker correlations between stellar metallicity and structural and density quantities. The results from CT23, however, suggest that environmental history influences chemical enrichment over time. We show in Figure 5.3.5 that calculated overdensities are marginally larger in Uchuu, which can influence metallicities by predicting an overabundance of mergers and flybys which redistribute the metals into high mass galaxies, as well as contribute to quenching. The anisotropic nature of these interactions are traced by radial skews, which we show in CT23 to have a profound effect on the metallicity history. Though these skews are not well correlated with other halo properties, being difficult to compare between simulations, the low number density of haloes around low mass objects can fail to produce the extreme skews experienced during close interactions which contribute significantly to chemical enrichment.

Evaluating the mass-metallicity relations from these star formation and metallicity histories in Figure 5.4.8, we see that the underpredicted metallicity histories distort both relations

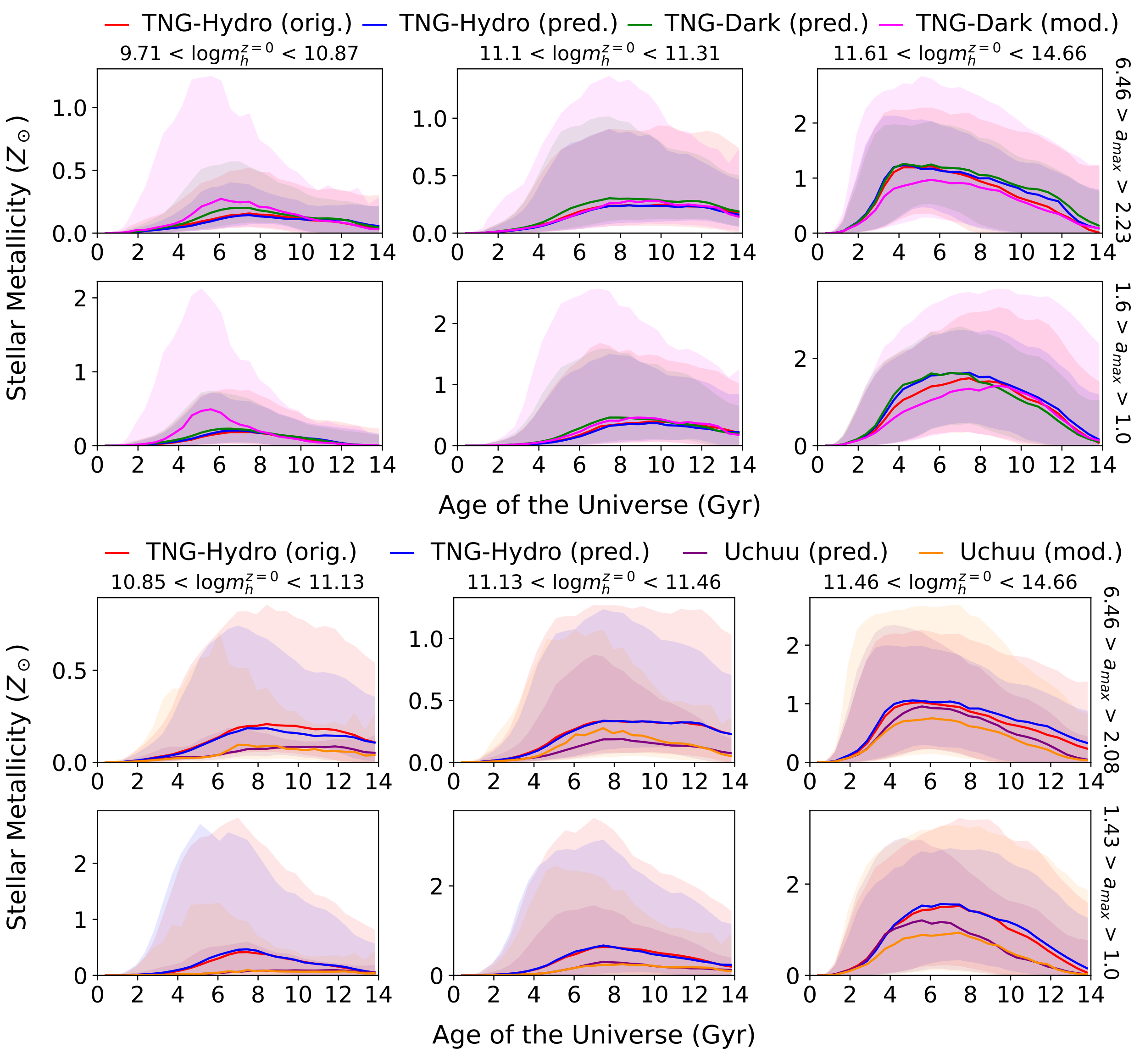

Figure 5.4.7: Metallicity histories of satellite galaxies, binned in the same manner as in Figure 5.4.6.

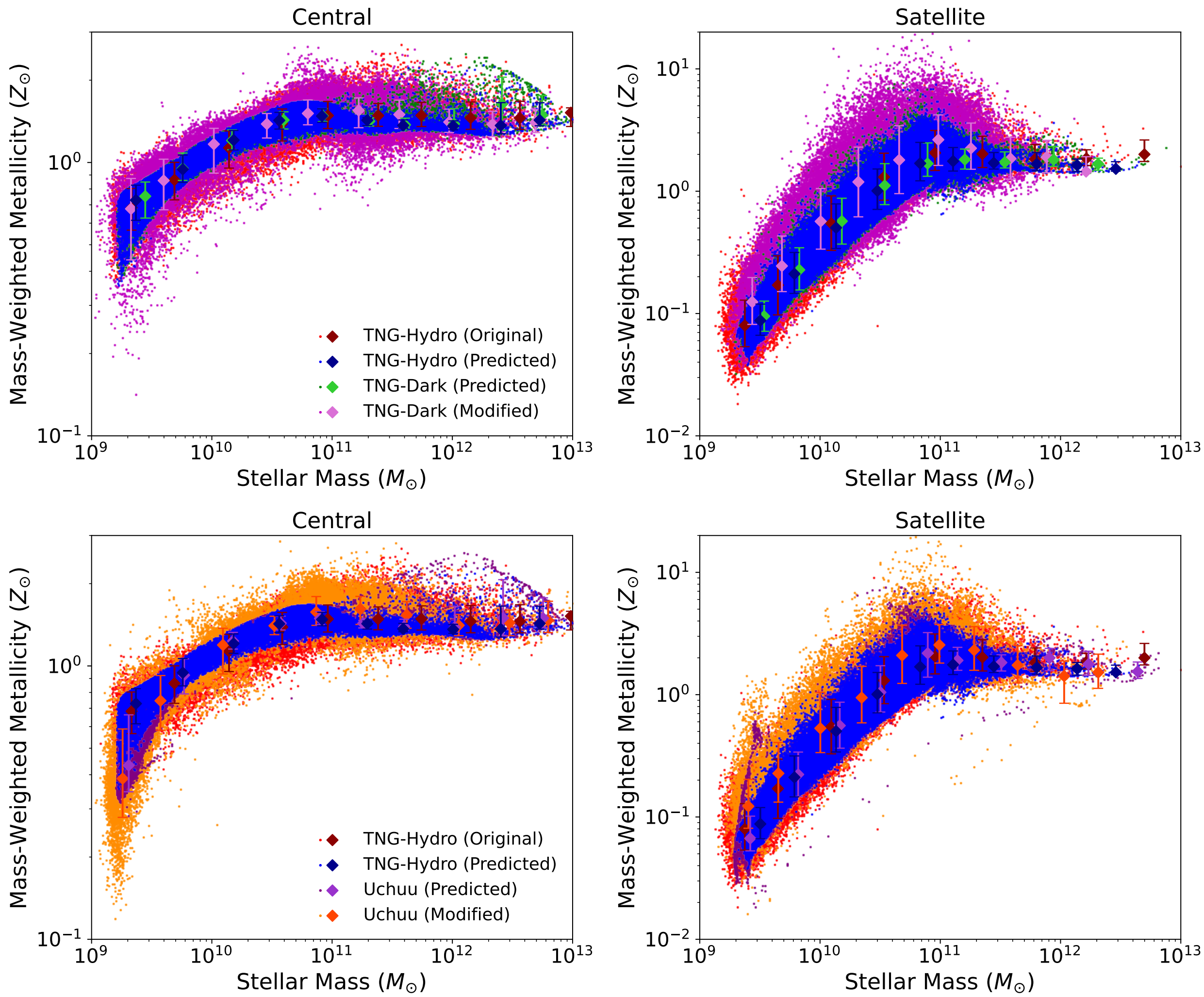

Figure 5.4.8: Numerical MZR of central galaxies (left) and satellite galaxies (right), with numerically evaluated total stellar mass and mass-weighted metallicity.

at the low mass end. Nevertheless, the shape and scatter of the relations in the dark simulation are very similar to the network's original TNG-Hydro predictions; particularly following stochastic modification. For central galaxies, however, there are a handful of overpredicted metallicities in the dark simulations at high mass. A Spearman coefficient of 0.683 between age and metallicity for Uchuu galaxies above a stellar mass of $10^{11.5}M_{\odot}$ shows that these are the same galaxies whose mass-weighted ages are overpredicted. Therefore, there is a greater contribution of metallicity histories at early times to these results, and negligible contribution from when the galaxies begin to lose their star formation. This effect is subsequently mitigated by the stochastic amendment, reducing the coefficient to 0.135; similar to the TNG-Hydro target value of 0.217. As previously established, the modification shifts the early star formation towards the TNG target; but additionally adds fluctuations which enhance the chemical evolution across all times. This improvement, not only to the stellar formation time, but to the characteristic time of metal synthesis, is of great value for refining the predicted spectroscopy which would be used in N-body mocks. We discuss these results in the following section.

5.5 Mock Observables

This section concerns how the previously described similarities and discrepancies between halo and galaxy properties affects the derived observational results, and thus is an assessment of the quality of hypothetical mock survey statistics.

5.5.1 Galaxy Spectra

We show in Figures 5.5.1 and 5.5.2 the spectral energy distributions of the four simulation datasets in bins of stellar mass, for central and satellite galaxies. In CT23 we relate the smaller variance of the predicted SEDs to the lack of variability in star formation histories, and unconstrained implicit features such as merger-driven starbursts and quenching timescales, which are more common to central galaxies.

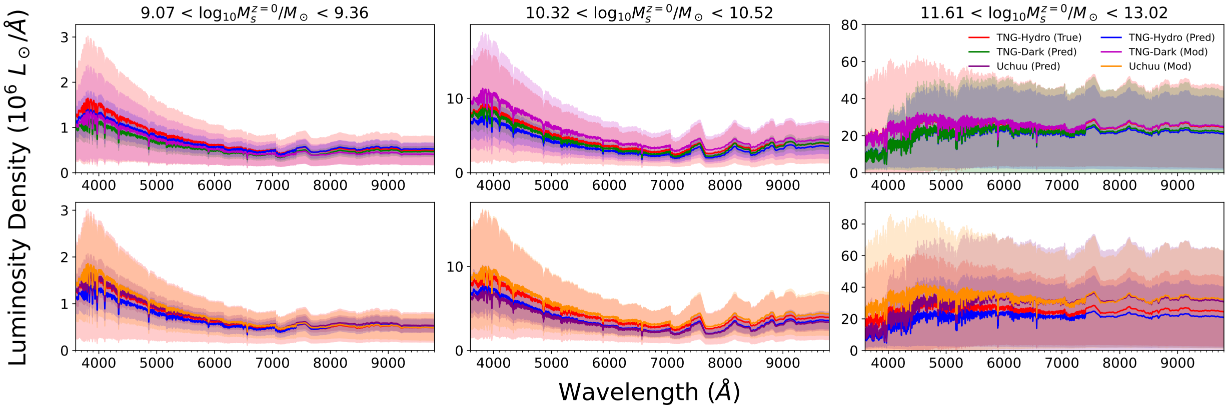

Figure 5.5.1: For central galaxies, this figure shows the evaluated SEDs in TNG-Dark (upper panels) and Uchuu (lower panels) in contrast with the TNG-Hydro data. For each dark simulation, the spectral luminosities are on average under-predicted slightly, but this is improved by the stochastic modification, which also serves to restore some of the variance in these spectra. Emission lines are not shown in these spectra, for the sake of clearly showing the mean continuum from each simulation.

The mean amplitudes of TNG-Dark spectra are subtly smaller for low and intermediate mass central galaxies, and larger for satellites. The satellite discrepancy can be attributed to higher peaks in star formation histories at such masses. While this affects both networks, the effect is particularly prominent for satellites, which in addition to the abscence of stellar feedback driven mass loss[249], may be likened to the lack of environmental harrassment serving to strip the satellite halo following infall[260], as discussed in Section 5.4.1. For central galaxies, the cause of this offset is not clear, but given differences in star formation histories and mass-weighted ages, the offset may owe to the underpredicted star formation histories seen in Uchuu predictions.

The stochastic modification systematically increases these galaxy luminosities as it does with stellar masses, therefore it improves the spectral amplitudes for central galaxies, and not satellites, which additionally exhibit greater variance than the target data when compared with central galaxies. While central galaxies tend to quench as a result of mergers and AGN growth, satellite quenching can be driven by the environment imposed by its host galaxy, and can act on short and on long timescales, depending on infall trajectory, satellite-host mass ratio and redshift[261;262]; potentially explaining this inconsistency between the spectral modifications of central and satellite galaxies as a matter of the physical processes which act

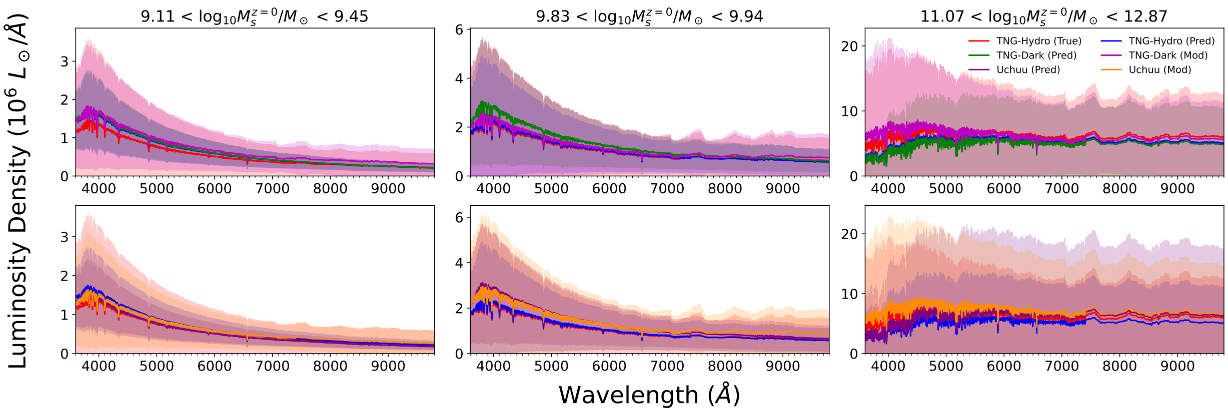

Figure 5.5.2: For satellite galaxies, this figure shows the evaluated SEDs in TNG-Dark (upper panels) and Uchuu (lower panels) in contrast with the TNG-Hydro data. The two dark simulations are well-matched to the predictions in the hydrodynamic TNG simulation, showing similar means and variances in the spectra. This causes the average luminosity to be brighter than needed following modification, but like the central galaxies, the modification recovers the variance originally seen in spectral luminosities. As with Figure 5.5.1, emission lines are omitted from the figure for clarity.

in the effective frequency range of the stochastic modification.

In Uchuu, the spectra are generally lower in amplitude than TNG-Dark spectra, resulting from weaker star formation histories. One exception is low mass central galaxies, which may be a result of poorly predicted metallicity. In high mass bins, however, the variance in Uchuu spectra is larger than that of any TNG data, corresponding to larger variance in mass accretion histories. The large spread of Uchuu overdensities in high mass bins will additionally have misled the network, either by enhancing star formation by associating densities with merger rates, or conversely, suppressing it by association with quenching. In fact, the Spearman coefficient between zero-redshift overdensity and each band magnitude is on average -0.43 for galaxies above $10^{11.5} M_{\odot}$, compared with -0.09 overall. The stochastic amendment may reshape these large Uchuu galaxy spectra to more closely resemble that of TNG galaxies, but of course cannot reduce the variance as needed, only reducing this average Spearman coefficient to -0.38.

5.5.2 Galaxy Photometry

In Figure 5.5.3, we show the colour-mass diagrams evaluated from the four simulation datasets, showing for both central and satellite galaxies, the dependence of colours calculated using neighbouring SDSS wavebands on their stellar mass. As in CT23, these results show the bimodal colour distributions of the galaxy populations, and the tendency for high mass galaxies to be redder in colour, to be a feature of all predictions of the neural network. This work showed that the unconstrained emission at UV frequencies has resulted in underpredicted colours for u and g bands, which shows in the dark simulations as well. While the stochastic modification recovers the widths of the bimodal colour peaks, owing to improved variance in star formation and spectral amplitude, it fails to correct the offset peak in $u-g$ colour, suggesting the need to accurately predict recent star formation to accurately constrain key spectral features such as the 4000 break.

In the predictions of the dark simulations, there are some notable differences with respect to the predictions of the hydrodynamical simulation. The ages of high mass galaxies are biased towards high values in both dark simulations, as shown in Figure 5.4.5; and correspondingly, a higher fraction of galaxies reside in the "red" peak in these datasets, particularly for $u-g$ and $g-r$ colours. This overabundance of red galaxies is potentially a result of a greater rate of galaxy quenching, which was discussed in Section 5.4.1 in relation to the differing morphology of haloes in dark simulations; or the denser environments in dark simulations which serve to quench interacting or infalling galaxies. The stochastic modification partially corrects this by reshaping the star formation histories, making some red galaxies bluer by eliminating excess early star formation, but does not recover quenching tails which would also shift these colours.

Another aspect of the dark predictions is the greater abundance of galaxies in the "green valley": the transition phase between star-forming and quiescent galaxies. This feature applies to both dark simulations, but is particularly prominent in Uchuu. Unlike the colour offset arising from the galaxy age bias, galaxies in this mass range typically have a broader range of calculated magnitudes, which, like the high variance in spectra, may be attributed

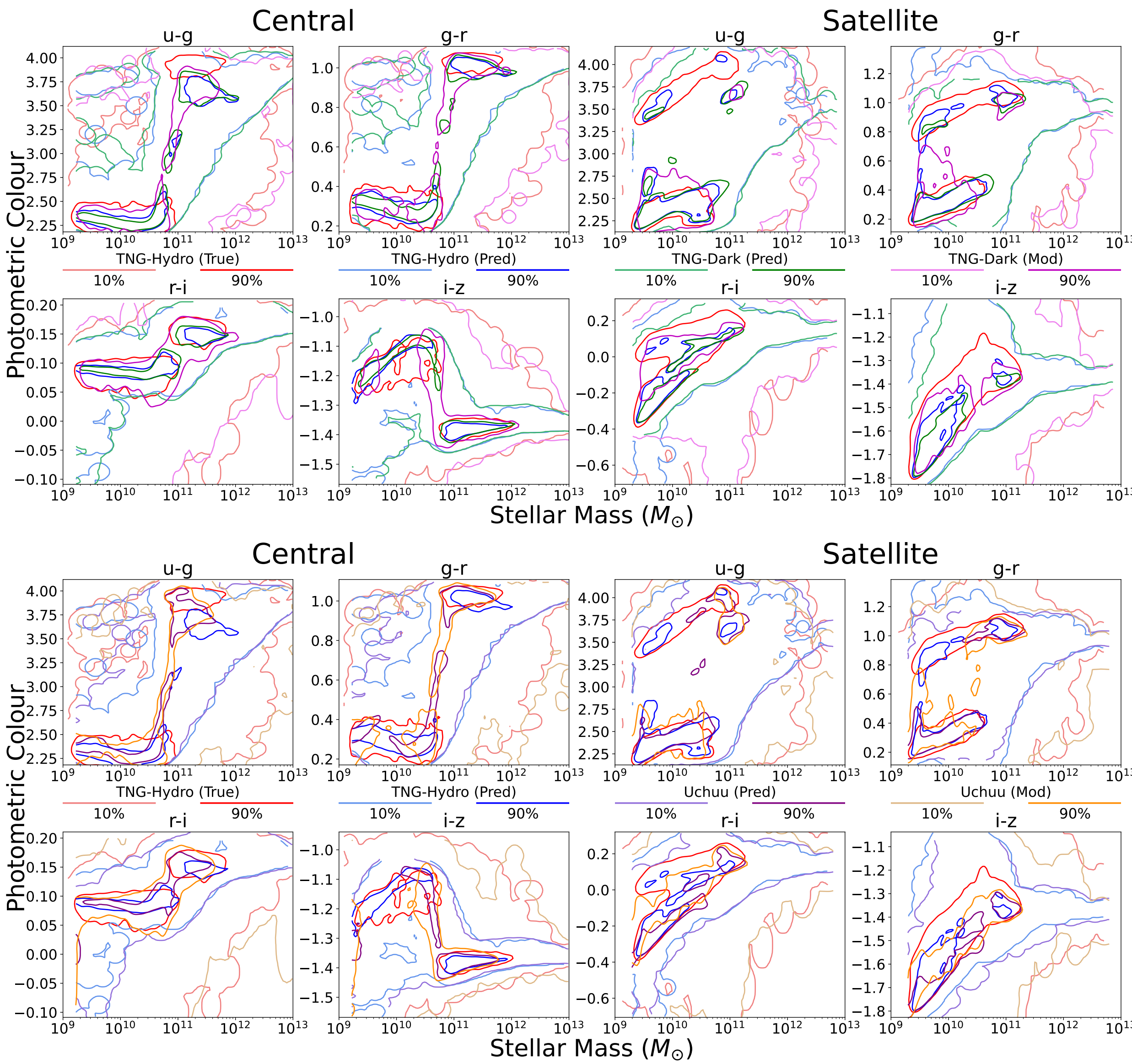

Figure 5.5.3: Colour-Mass diagrams from each of the four simulation datasets, shown for central galaxies (left panels) and satellite galaxies (right panels), and for four colours taken as the difference between successive SDSS band magnitudes. Contour lines show the 10$^{\text{th}}$ percentiles (light-coloured lines) and the 90$^{\text{th}}$ percentiles (dark-coloured lines) of the 2D histograms of the data. The predicted photometric colour bimodality and its association with mass are shown by all datasets, however the dark simulations show an excess of samples in between the peaks of the colour distributions. The stochastic modification redistributes the colour distributions, recovering the full range of colours seen in the original data, but does not rectify the offset peaks of the distribution, particularly for high frquency bands.

to variations in the halo mass accretion history and internal dynamics. Furthermore, an effect of the stochastic modification is the enhancement of the green valley population; as the modification is sensitive to a frequency range which captures a specific set of star formation features, the correction it implements may only constitute a partial shift in the photometric colour.

We have shown in CT23 that the variability in star formation history, particularly at late times, is an important factor in modelling photometry. Even in cases where the Fourier Transforms of Uchuu galaxies are well predicted, the star formation histories in Uchuu may have lost additional high frequency information from being based on temporal predictors which were interpolated over a sparser time domain in Uchuu, which could not be rectified by a TNG-based stochastic modification. Uchuu galaxies in this mass range are also slightly redder on average than TNG-Dark, which corresponds to declining star formation histories in this mass range. In fact, this effect is seen at the approximate mass at which biased galaxy ages begin to appear, suggesting that over-quenching is indeed a prevalent issue. This emphasises the importance of constraining late star formation to accurately predict observable data.

This modification is additionally only able to correct for differences in network predictions which translate to intermediate frequency modes in the star formation history, including supernova feedback but not necessarily slow quenching. Concerning high mass galaxies, Iyer et al.[254] show that kinetic AGN feedback introduces a variable star formation component, which in TNG contributes significant energy injection only above a threshold black hole mass of $\sim 10^8 M_\odot$; unlike the thermal feedback mode, whose energy yield is tightly correlated with this AGN mass[255]. These distinct modes of AGN feedback may explain why the stochastic modification can only partially amend the geometry of quenched star formation histories; the kinetic feedback resides in the frequency domain of the modification, while the long-timescale thermal feedback is closely related to AGN mass. While AGN mass is not predicted directly by this model, it is closely related to star formation history, and so the effects of thermal AGN feedback will be influenced by the established shortcomings of the fiducial neural network predictions, in both the hydrodynamic and the dark simulations.

5.6 Discussion

The machine learning model used to predict the star formation and stellar metallicity histories of TNG galaxies has performed modestly at reproducing the same results in pure dark matter simulations, having obtained similar quantitative relations between galaxies and haloes, and relating physical galaxy properties to observational quantities as in the hydrodynamical TNG simulation. Despite this, a number of discrepancies exist between the different predictions, resulting from differences in the growth and interactions of haloes in N-body simulations compared with the full physics, differences in the calculation of halo properties in the simulation data, and the degree of improvements made by stochastic modification. In this discussion, we identify the important differences between these simulations and how the model may be modified in future to suit high volume N-body simulations while maintaining an accurate characterisation of the galaxy-halo connection.

5.6.1 Resolution Effects

We have shown that the lower resolution of the Uchuu simulation compared with the TNG simulations has worsened the quality of predictions of low mass and slowly accreting galaxies, with considerable errors in the masses and metallicities of low mass galaxies; particularly satellite galaxies, whose sampling was already compromised by our quality cuts. One such effect is that the truncated power spectrum suppresses the coalescence of simulation particles into low mass haloes, such that low mass haloes take longer to germinate and grow. Low mass samples in Uchuu are evidently unreliable for causal galaxy-halo modelling. The SFH and ZH Fourier Transforms predicted by this model are poorly predicted for these haloes as well, indicating that resolution differences additionally effect our stochastic methodology.

The limits of computational resources required to generate a high-volume, high-resolution simulation has been a long-standing issue in this field. Li et al. [263] demonstrate a machine learning model which can enhance the matter power spectrum in N-body simulations and generate accurate snapshot data from low-resolution simulation images, which accurately replicate halo substructures and correlation functions below the limit of said low-resolution simulations[264]. This may be a practical technique to enhance the Uchuu data in this work, provided that the complete merger trees can be constructed from the model's outputs. Given the importance of the properties of progenitor subhaloes, such as the mass and metallicity of incoming gas and stars, it may be necessary to develop a similar enhancement model of the merger trees themselves; however, as the matter power spectra, distributions and velocity fields can be accurately enhanced this way, these properties may be inferred from the predicted environments.

Our results show that haloes in Uchuu of similar mass to the lowest mass haloes in TNG are substantially affected by the difference in resolution, and thus, it may be worthwhile in a future study to apply the neural network to Shin-Uchuu: a smaller but higher resolution run of the Uchuu model, to restore low mass galaxies; or to enhance the Uchuu snapshots and merger trees based on the superior Shin-Uchuu resolution and the methods of Li et al.[263]. As the growth of large scale structure will influence the distribution of galaxies and the properties of their environments, the latter may be a more suitable method for creating self-consistent mock catalogues on both gigaparsec and sub-kiloparsec scales.

5.6.2 Alternative Models

Uchuu-UniverseMachine

Aung et al.[265] use the UniverseMachine (UM) model[266] to compute galaxy formation histories in the Uchuu simulation, which is used to recover statistics such as stellar mass functions and number density profiles which show reasonable agreement with observational stellar mass and luminosity functions[267]. UM is an empirical model, relying on MCMC optimisation of star formation rates using prior relations of star formation rates and quenched fractions to the halo rotation curve. The model was able to qualitatively reproduce the environmental dependence of star formation without explicit implementation of the environment, yet this implies that the star formation in dense environments is only supported by halo mass accretion. UM does not model metallicity histories, which we have shown in CT23 to be more dependent on environmental quantities, and thus our model may be more suitable for modelling the dependence of chemical enrichment in high fidelity mocks. On the contrary, as quenched fractions are a direct parameter of the UM model, the colour bimodality in these mocks is likely to be more accurate, and so UM may be more suitable for observational statistics.

A predictive, self-consistent model such as ours may prove complementary to empirical models such as UM when it comes to producing high-fidelity mocks, as it can be used to causally model the growth of galaxies over time based upon the halo and environmental quantities driving the physics of the galaxy-halo connection, and may be modified to predict galaxy properties, for instance, spiral-bar structures. Yet, our machine learning model and stochastic modification have shown to be sensitive to effects which come from translation to a pure dark matter simulation, such as the biased growth of internal dynamics of the halo, resolution effects such as the delayed collapse and virialisation of haloes, and potentially, nuances in the calculation of key variables such as halo mass and radius due to the use of different halo finder algorithms; which may, for instance, identify different structures in the halo centre, or different boundaries enclosing the total halo mass[247]. Thus, it may prove that an N-body simulation with resolution enhancements and consistent halo definitions with the training data will be necessary for accurate self-consistent mocks.

Gaussian Process Stochasticity

Our stochastic modification to the star formation and metallicity histories has proven valuable in recovering the short-timescale features which were absent from the fiducial neural network model, corresponding to relevant phenomena such as starbursts and baryon cycling. It is nonetheless difficult to disentangle the physical drivers of these stochastic features, which have different degrees of influence depending on a galaxy's mass, age and environment. To analyse the relative contribution of these dynamical processes to observable features such as the H-α flux and D_n4000 index, one might incorporate a Gaussian process formulism as seen in Iyer et al.[268], where the stochastic modes of star formation are parameterised by three analytic power spectra with controlled effective timescales.

Our stochastic phase selection method presented in BTC24 considers the cross-correlation of modes of halo and galaxy growth, and so the evolution of the host halo may provide some insight into constraining these timescales for individual galaxies, if these two stochastic analyses can be reconciled. This may require more historical information than what was presented in our companion paper, considering that the gas inflow and cycling studied by Iyer et al.[268] are potentially sensitive to environments; but if the stochastic parameters can be infefrred solely from properties of the halo and environment, they can contribute to self-consistent predictions in N-body simulations.

5.6.3 Alternative Variables

The variables used in the design of our neural network were meant to characterise the growth, structure and environment of haloes in a manner which was as immune to the effects of baryons and different resolutions and halo finders as possible. However, the results have shown that these differences have nonetheless produced small to substantial differences in the statistics of galaxies predicted using pure dark matter simulations with respect to the hydrodynamical TNG model. Thus, the inclusion of as-yet unused variables in a future rendition of the neural network model may constitute more suitable measures of the galaxy-halo connection for use in a pure dark matter model.

As mentioned in Section 5.6.2, the use of a consistent halo finder in the hydrodynamical and N-body simulations under consideration may be necessary to avoid biasing predictions due to the different identification of halo mass and substructure. Onions et al.[247] state that while many halo finder algorithms are alike in their capability of identifying halo structures, Rockstar is the only algorithm in this study which accurately resolves substructures in highly dense regions such as the halo centre, which can influence the structural variables $R_{\frac{1}{2}}$ and $\tilde{v}_{\rm vir}$; shown in CT23 to influence the SHMRs of both central and satellite galaxies. The planned TNG halo catalogue based on the Rockstar halo finder may be used to train an "Uchuu-friendly" neural network model in future research, effectively eliminating this potentially appreciable biasing factor.

Another quantity which differs between the simulations in this study is overdensity, which in Uchuu is computed using halo coordinates and masses in lieu of similar subhalo information as in TNG. Overdensities may be smoothed according to a Gaussian kernel to produce a continuous, position-dependent density field as in Chen et al.[269], which may be tuned for each simulation according to their resolution or their density tracers. This method was used similarly to create a position-dependent tidal field tensor, which the authors show to correlate strongly with halo assembly bias.

For satellite galaxies, the environment they experience in the vicinity of their host halo dominates their behaviour in the satellite phase, and therefore the properties of the host are additionally important. In CT23 we mention how the explicit inclusion of further satellite host properties may have been beneficial to the model; examples include formation time[270], distance from the halo centre[260;271] and angular momentum[272]. The location of the satellite with respect to the host, or of a central halo with respect to a local cluster, would constitute a tracer of the local environment, and thus could be a valuable measure of effects such as gas stripping and tidal disruption.

As the skew parameter, being mass-independent and calculated using halo positions, does not differ substantially between any of the simulations used in this work, a purely position or velocity dependent environmental metric may be a feasible solution to the bias introduced by basing environments on overdensities. However, as the skew is a measure of halo-halo interactions over time, something not characterised by smooth fields, it remains a valuable parameter of this model. Yet as mentioned in Section 5.4.2, an effectively random error may be introduced due to the lack of low-mass objects in a low resolution simulation, which would be particularly detrimental to the major interactions inferred from high or low skews.

As for the stochastic modification, while we use stellar mass to bin samples and recover separate phase distributions, we have tested the modification with an additional halo mass binning, though this introduced errors in cases of low sample size, such as with the most massive haloes. It should be stressed here that the stochastic modification does not perfectly reconstruct important summary statistics such as the MZR, and sampling phases with respect to environment may improve this result, provided that these are well sampled.

Otherwise, combining results from simulations on different scales, such as TNG50[273;274] for small scales and MillenniumTNG[275–277] for large scales, may provide enough data to justify multidimensional phase selection.

5.7 Conclusions

In this work, we have compared the quality of predictions of galaxy star formation and metallicity histories in pure dark matter simulations, based on the semi-recurrent neural network described in Chittenden and Tojeiro[209], and the stochastic star formation modelling of Behera, Tojeiro, and Chittenden[188]. We have compared our original predictions with cross-matched haloes from the dark equivalents of the TNG simulations on which the model was trained, evaluating the effects on predictions due to the abscence of baryonic processes; and we have applied the model to similar haloes from the Uchuu N-body simulation, to evaluate the effects of alternative halo definitions, and the lower mass resolution of the simulation.

Our findings can be summarised as follows:

- Important input properties such as the mass accretion history of a halo are similar between the simulations under consideration, for haloes of most masses and mass accretion gradients, as defined in Montero-Dorta et al.[231] and Shi et al.[229]. Nevertheless, we show in Section 5.3.1 that the mass accretion histories of TNG-Dark haloes are exaggerated, potentially as a result of the lack of stellar feedback shaping the haloes. This has noticable, similar influence on the star formation histories of low mass galaxies, as shown in Section 5.4.1. As this difference in mass accretion histories arises at early times, the consequential effect on the star formation histories applies for most of the time domain of the simulation, due to the recurrent design of the neural network.
- The network quantities relating to internal structure and dynamics: the half-mass radius and virial velocity of the halo, are similarly affected by the lack of baryons, but are more profoundly affected by the lower resolution of the Uchuu simulation, which becomes apparent in Section 5.3.2. This causes the germination and initial growth of

haloes to be delayed and for the mass accretion and concentration of Uchuu haloes to appear smaller than their TNG equivalents. For low mass and slowly growing haloes, these effects are dramatic, due to the sensitivity of these results to the simulation resolution.

- We show largely similar neural network predictions to the original, hydrodynamical predictions in Section 5.4, which recover similar shape and scatter to the hydrodynamical SHMR and MZR. Despite this success, the severely underpredicted growth of low mass haloes in Uchuu leads to poor predictions of the star formation and metallicity histories of low mass galaxies. Conversely, TNG-Dark results are overpredicted due to excess mass accretion at early times, which leads to an effective overabundance of star-forming gas. In each case, the lowest mass galaxies in any dark matter simulation are the least accurate predictions.

- With the added stochastic modification, we make significant improvements to key statistics of the dark simulations, recovering the scatter of the mass-metallicity relation and correcting the spectral luminosity and photometric colour distributions of predicted galaxies. While this amendment is shown to fruitfully reproduce the missing features in the hydrodynamic TNG simulations in our companion paper, it also rectifies systematic distortions in the geometry of the star formation and metallicity histories in these pure dark matter simulations, providing more accurate predictions for future N-body mocks. However, it does not amend the issues which largely apply to low mass, under-resolved objects, where the predicted Fourier Transforms are under-predicted, thereby failing to provide adequate information to compute realistic baryonic data. The modified data is therefore just as sensitive to resolution effects as the fiducial predictions.

- In both dark simulations, the number of quenched galaxies predicted by the neural network is larger than the hydrodynamical result. In TNG-Dark, this is a result of the difference in structural parameters, which may be attributed to AGN feedback. In

Uchuu, this is an effect of higher overdensities, which owe to the use of halo tracers rather than subhaloes. This excess quenching corresponds to a greater abundance of photometrically red galaxies, which we show in Section 5.5. In Uchuu, the abundance of red galaxies is further attributed to underpredicted Fourier amplitudes, and possibly interpolation over a coarser time domain, resulting in greater information loss regarding time variations in their star formation history. The stochastic modification serves to redistribute the photometric colours by reshaping the star formation histories, but does not amend differences in the shapes of quenching tails, leading to some "partly corrected" photometry, and an overabundance of green valley galaxies. The spectroscopic and photometric statistics of the dark matter simulations nonetheless show similar characteristics to the original TNG results, yet this could be improved by explicit constraints on star formation at the lowest redshifts.

This paper has shown that the neural network model we have developed is a highly practical tool for emulating galaxies into N-body simulations using the learned physics of the galaxy-halo connection. However, the shortcomings of the study illustrate the necessity for an N-body simulation of similar mass resolution and halo properties to compute accurate mock surveys on all mass scales. This may be achieved in future work by redesigning the predictive model to measure halo structure and cosmic environment using identical halo finders, metrics which are independent of the simulation, or by enhancement of existing low-resolution haloes and merger trees using machine learning or similar methods.

By obtaining results in the Uchuu simulation with similar characteristics to the TNG simulation suite on which the model was trained, we have shown that it is possible to produce a physical, self-consistent model of the galaxy-halo connection on present-day, gigaparsec-scale N-body simulations. This may be a pathway to a future of AI-based simulations which entail a physical explanation of galaxy evolution on cosmological and substructure levels simultaneously, and corresponding observational data to complement the most ambitious galaxy surveys to date, for a fraction of the computational cost of a hydrodynamical or semi-analytic model of the same level of detail.

Acknowledgements

The UKRI Science and Technology Facilities Council supported this study under grant ID ST/T506448/1, which the authors gratefully acknowledge. We appreciate the IllustrisTNG project allowing us access to their data and JupyterLab environment, and the Uchuu and YTree projects for access to Uchuu halo merger trees. We also thank Katarina Kraljic for assistance with modelling the cosmic web in Uchuu, and we thank Daniele Sorini and Kartheik Iyer for useful discussion on the role of gas-rich progenitors on developing halo shapes, and the influence of different modes of baryon cycling and AGN feedback in galaxy evolution models, respectively.

Data Availability

The principal author is currently willing to provide the information that backs up the research's conclusions upon reasonable request. The data will later be posted online for general access.

“ଅନ୍ଧକାର ମାଧ୍ୟମରେ ଆଲୋକ ନିର୍ମିତ — ଏହି ହେଉଛି ପ୍ରକୃତିର ବିପରୀତ ଦୈତ୍ୱା”

— Inspired by Vidyapati's dualism

“Through darkness is light constructed — this is nature's sacred duality.”

Chapter 6

Complementary Research: Extensions and Independent Projects

This chapter presents research efforts that complement the core themes developed in this dissertation, while also reflecting independent contributions within the Dark Energy Spectroscopic Instrument (DESI) collaboration. The work described here extends beyond the modeling of large-scale structure presented in earlier chapters, and includes both theoretical developments and observational analyses. Specifically, the chapter covers: (i) an extension of bispectrum BAO modeling to capture anisotropic information via angular multipoles, (ii) the mitigation of imaging systematics in large-scale structure clustering catalogs, and (iii) the spectroscopic classification of AGNs using emission-line diagnostics across multiple DESI target classes.

These projects represent a broadening of scope from simulation-driven modeling to data-oriented contributions, and demonstrate continued engagement with cosmological and astrophysical challenges in current and future spectroscopic surveys.

6.1 Anisotropic Clustering in the Bispectrum via Angular Multipoles

Chapters 1 and 2 established a framework for detecting the Baryon Acoustic Oscillation (BAO) signal in the bispectrum monopole. While the monopole captures angle-averaged clustering, it discards directional information induced by redshift-space distortions (RSDs), thereby limiting its sensitivity to anisotropic cosmic expansion. The ongoing work described here extends the formalism to include anisotropic components of the bispectrum by incorporating its angular multipole structure, enabling joint constraints on transverse and line-of-sight distances.

6.1.1 Scientific Motivation

Redshift-space distortions imprint anisotropies in the observed galaxy distribution due to coherent peculiar velocities and the Alcock–Paczyński effect. These anisotropies manifest in the angular dependence of the bispectrum, which is sensitive to the geometry and dynamics of structure growth. Traditional bispectrum analyses focus on the monopole ($\ell = 0$), which integrates over angular orientations and thus loses this directional information.

To recover it, the bispectrum is decomposed using a spherical harmonic or Legendre multipole expansion. This approach, based on the formalism introduced by Sugiyama et al.[7],[119], projects the bispectrum onto angular basis functions labeled by ℓ, capturing the anisotropic signal induced by both cosmology and RSDs. Each multipole moment is differently sensitive to distortions in radial ($\alpha_{\parallel}$) and transverse ($\alpha_{\perp}$) directions.

This decomposition allows the measurement of the anisotropic BAO scale by separating contributions from different triangle orientations. As shown in Figure 6.1.1, different combinations of $(\alpha_{\parallel}, \alpha_{\perp})$ leave distinct signatures in the bispectrum multipoles, and their combination breaks degeneracies present in angle-averaged statistics.

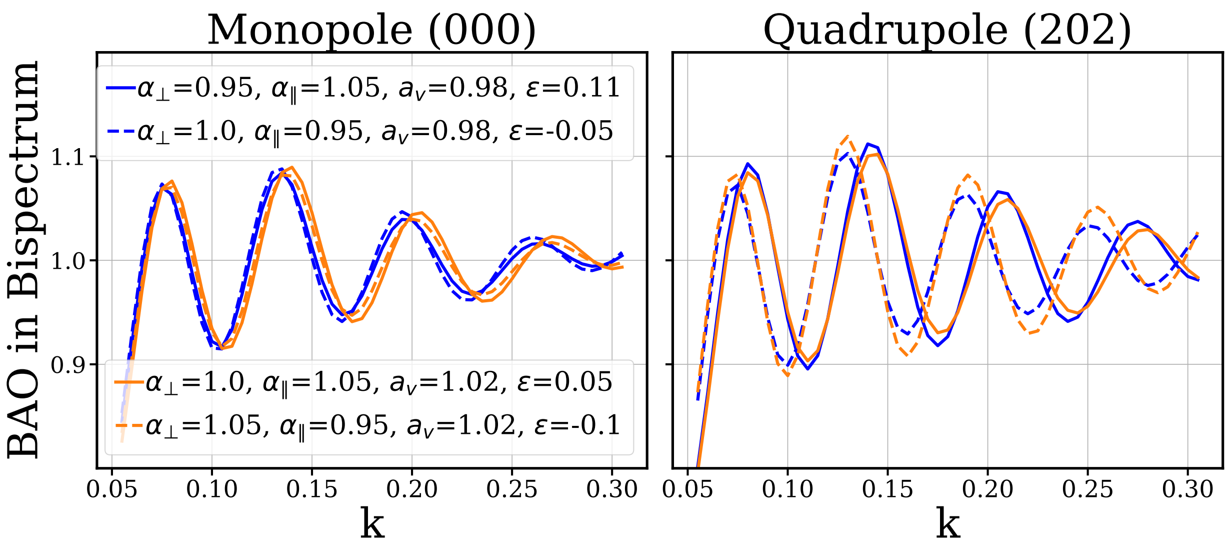

Figure 6.1.1: BAOs from Diagonal Bispectrum for different pairs of $\alpha_\parallel$ & $\alpha_\perp$, computed theoretically using Sugiyama et al.[7]. The plot shows how combining monopole(left) and quadrupole/s (right) can break the degeneracy of α and ϵ.

6.1.2 Methodology and Implementation

The analysis uses diagonal and near-diagonal triangle configurations measured from dark matter halo catalogs in the ABACUS simulation suite, as well as GLAM mocks for volume and noise control. Bispectrum multipoles are extracted using the `Triumvirate`[48;119] analysis package, and modeled using perturbation theory templates that incorporate RSDs and BAO oscillations.

By comparing theoretical predictions for the monopole and quadrupole across a grid of $(\alpha_\parallel, \alpha_\perp)$ values, sensitivity to anisotropic scaling can be mapped. Preliminary results demonstrate that:

- The monopole alone provides constraints on an average BAO scale, but remains degenerate in $\alpha_\parallel$ vs $\alpha_\perp$.
- The quadrupole introduces shape-dependence that varies distinctly with $\alpha_\parallel$ and $\alpha_\perp$, enabling their separate recovery.
- Combining multipoles can improve statistical precision and breaks degeneracies, especially at lower redshifts where nonlinear RSD effects are stronger.

6.1.3 Challenges in Bispectrum-Based Anisotropic Analysis

While promising, angular bispectrum analyses face several technical and conceptual challenges:

- **Angular basis complexity:** Spherical harmonic decomposition introduces mode-mixing and requires careful treatment of angular resolution and sampling completeness.

- **Computational cost:** Measuring and modeling bispectrum multipoles across all triangle shapes and orientations is resource-intensive.

- **Nonlinearities and modeling systematics:** The impact of nonlinear growth, fingers-of-God effects, and bias complicates the interpretation of higher multipoles beyond $\ell = 0$.

- **Survey realism:** The bispectrum is sensitive to window functions, masking, and fiber collisions. Incorporating these into anisotropic measurements is non-trivial and is the subject of ongoing methodological development.

Initial measurements of the monopole and quadrupole from simulation boxes show that multipole combinations can recover input dilation parameters with high accuracy. The next phase involves adapting the analysis to DESI data, starting with Year-1 ELG and LRG samples. This will require accounting for observational systematics, integrating the survey window function, and validating the theoretical templates against realistic mocks. The ultimate goal is to deliver joint BAO constraints from the two- and three-point statistics, with the bispectrum multipoles contributing unique leverage on anisotropic scaling.

6.2 Mitigation of Imaging Systematics in DESI Clustering Analysis

6.2.1 Background and Motivation

Large-scale structure surveys rely on photometric data to pre-select galaxy targets for spectroscopy. However, these imaging datasets are often affected by observational systematics such as seeing, stellar contamination, Galactic extinction, sky brightness, and depth variations. These systematic effects can correlate with observed galaxy densities and mimic or obscure cosmological clustering signals.

In cosmological analyses, particularly in angular or three-dimensional correlation functions, this leads to:

- Biases in the inferred clustering amplitude.
- Spurious large-scale modes introduced by survey systematics.
- Underestimated uncertainties in covariance matrices.

To mitigate these biases, regression-based methods are employed to model and remove correlations between target density and imaging systematics. DESI's Year-1 analysis relied on linear methods such as `IMLIN`, `mode subtraction` and machine learning (e.g., `SYSNET`, `REGRESSIS`), but newer approaches involve including additional maps and dimensionality reduction techniques to improve robustness and interpretability.

6.2.2 Methodology and Implementation

This work aims to construct and evaluate systematic templates for DESI Year-3 (Y3) analysis. The main components of this project are:

- Compiling over the imaging maps including:
  - Stellar density maps from *GAIA* DR2.

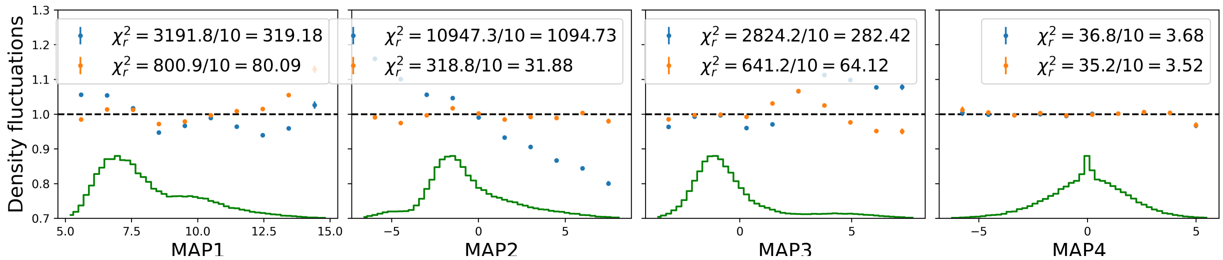

Figure 6.2.1: Preliminary plots from regression using 4 transformed and reduced maps (using PCA) on Y1 ELG $z = 0.8$-1.1. Bottom panel shows the density profile of each map and top panel shows Chi-sq before correction in blue and after in orange.

  - Emission line flux maps from DESI MWS.
  - Atmospheric seeing and sky brightness metrics.

- Angular cross-correlation analyses between these maps and galaxy tracer densities to quantify the impact of systematic contamination of each map on clustering.
- Applying principal component analysis (PCA) for dimensionality reduction and orthogonal template generation.
- Testing corrections using three regression frameworks: `IMLIN`, `REGRESSIS`, and `SYSNET`.
- Applying corrections to realistic Y3 mock catalogs for BGS, LRGs, ELGs, and QSOs.

As part of the correction workflow, the impact of each systematic on galaxy density is quantified, and regression models are trained to remove their influence. Chi-squared values before and after correction serve as validation metrics (Figure 6.2.1).

Preliminary analysis on Y1 data shows that using a subset of PCA-compressed maps achieves comparable or better performance than using the full template set, with lower risk of overfitting. These methods will be deployed in DESI's Y3 clustering pipeline and cosmological analyses.

6.3 Spectroscopic Classification of AGN and QSO Tracers in DESI

6.3.1 Background and Scientific Context

Active galactic nuclei (AGN) are powered by accretion onto supermassive black holes and represent a key phase of galaxy evolution. Spectroscopic classification of AGNs in surveys like DESI enables studies of:

- The co-evolution of galaxies and their central black holes.
- Feedback mechanisms that regulate or quench star formation.
- The role of large-scale environment in triggering AGN activity.

Due to their diversity in physical properties and observability, AGNs are identified using multiple diagnostic techniques. Emission-line ratios remain a cornerstone of optical AGN classification and are used to distinguish between Star-forming galaxies, Composite or transition objects, and AGN-dominated sources such as Seyferts and LINERs.
Common diagnostic diagrams include:

- **BPT diagram:** Uses [OIII]/Hβ vs [NII]/Hα or [SII]/Hα.
- **WHAN diagram:** Plots Hα equivalent width vs [NII]/Hα to include weak-lined AGN.
- **MEx (Mass–Excitation):** Combines stellar mass and [OIII]/Hβ to classify AGNs at higher redshift where Hα is inaccessible.
- **KEx (Kinematics–Excitation):** Uses velocity dispersion and line ratios to extend classification beyond classical boundaries.

With its wide redshift coverage and diverse spectroscopic target classes (BGS, LRG, ELG, QSO), DESI provides an unprecedented dataset for systematic AGN classification and demographic analysis.

6.3.2 Contributions to the DESI AGN Classification Effort

This work contributes to the construction of a diagnostic-based classification framework for AGNs and QSOs within the DESI Data Release 1 (DR1). Specific efforts have focused on:

- Constructing and applying BPT, WHAN, MEx, and KEx diagrams across multiple DESI tracers to distinguish AGN activity from star formation.
- Performing a tracer-by-tracer analysis of redshift distributions and emission-line completeness relevant to AGN identification.
- Interpreting how selection biases affect AGN demographics across the main DESI target classes, including emission-line galaxies (ELGs), luminous red galaxies (LRGs), and broad-line quasars.
- Investigating the consistency and complementarity of multiple classification schemes in overlapping redshift ranges.
- Identifying potential target selection effects and evaluating the need for revised or new demarcation lines in diagnostic diagrams across DESI's spectral and redshift coverage.

Preliminary WHAN diagrams constructed from a subset of DESI DR1 data (Figure 6.3.1) demonstrate clear classification separability across target classes. Each tracer exhibits characteristic distributions in the WHAN plane: ELGs primarily occupy the star-forming region; LRGs are concentrated in the passive and retired zones; and BGS galaxies span the full range of activity types, including strong and weak AGN.

Notably, two unexpected patterns emerge. First, QSOs appear underrepresented in the AGN-classified regions of WHAN space, with a significant fraction instead falling into the star-forming or intermediate zones. This trend likely arises from the dominance of broad-line emission in many quasars, which can obscure or dilute narrow-line features such as Hα and [NII], making standard WHAN diagnostics less effective. Additionally, some high-redshift QSOs may lack coverage or sufficient signal-to-noise in these lines. Second, the MWS sample—nominally targeting Milky Way stars—shows a non-negligible population

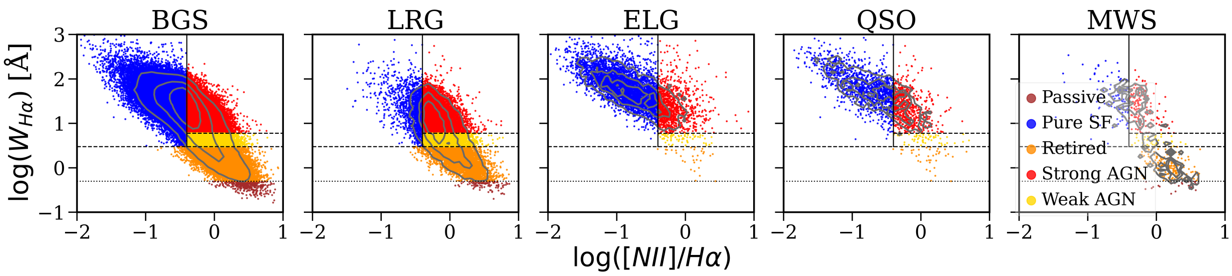

Figure 6.3.1: WHAN diagrams (Equivalent width of the Hα vs log([NII]/Hα)) of different DESI targets showing their classification based on the star forming and AGN activities and the degree of target selection bias.

of sources classified as AGNs in the WHAN framework. These likely represent compact extragalactic contaminants or low-redshift emission-line galaxies misidentified as stars during target selection, particularly near star-galaxy separation boundaries.

These results highlight both the power and limitations of emission-line diagnostics when applied across heterogeneous tracer samples. They also motivate the need for tracer-aware adjustments to classification thresholds and reinforce the value of multi-dimensional diagnostics in refining target selection and understanding classification systematics in spectroscopic surveys like DESI.

While detailed results and catalogs are forthcoming, this analysis is expected to enable a range of studies connecting AGN properties to their host galaxies, redshift evolution, and large-scale structure. This effort also connects with topics explored in earlier chapters of this thesis, particularly in understanding quenched galaxy populations and emission line characteristics within synthetic galaxy models.

6.4 Summary

The projects presented in this chapter highlight complementary extensions to the core themes of this thesis. The analysis of anisotropic clustering through the bispectrum's angular multipoles builds directly on the theoretical framework developed in earlier chapters, advancing the use of three-point statistics for precision cosmology. By leveraging redshift-space anisotropies, this work aims to improve constraints on cosmic expansion through a richer statistical description of large-scale structure.

The studies of imaging systematics and AGN classification represent independent yet scientifically aligned contributions to the DESI collaboration. The imaging analysis addresses a critical source of uncertainty in galaxy clustering measurements, contributing to more robust cosmological inferences. Meanwhile, the AGN and QSO diagnostic work establishes a physically motivated classification framework that enhances the interpretability of DESI spectroscopic data across a wide redshift range.

Together, these efforts illustrate a research trajectory that integrates theoretical modeling, observational analysis, and survey calibration. They reflect a commitment to developing data-driven methodologies that will support not only DESI but also future missions such as *Euclid*, *LSST*, and the *Roman Space Telescope*, where systematics control, multi-tracer cosmology, and physical source classification will be essential.

“ବିଭିନ୍ନ ପଥ ଏକ ଲକ୍ଷ୍ୟର ଦିଗରେ — ତଥାପି ପ୍ରତିଟିର ଅନୁଭବ ଅଲଗା।”

— Inspired by the Panchasakha school of Odisha

“Different paths point to one goal — but each experience is distinct.”

Chapter 7

Conclusion

This dissertation explored three complementary approaches to understanding and modeling the large-scale structure of the Universe: the precise extraction of the Baryon Acoustic Oscillation (BAO) signal from the bispectrum; the development of improved machine learning frameworks to predict galaxy formation histories from dark matter simulations; and a set of independent, supporting studies addressing observational systematics and tracer classification in ongoing surveys.

In the first part of this work, we investigated the BAO feature in the three-point correlation function, or bispectrum, using a "wiggle-only" modeling framework designed to isolate the oscillatory signal from the broadband shape. Through extensive analyses on large suites of GLAM simulations, we developed an aligned-template fitting method that robustly recovered the BAO dilation parameter with percent-level precision. This study demonstrated the potential of the bispectrum as a powerful, complementary probe to traditional two-point statistics, particularly for future large-volume galaxy surveys aiming for sub-percent cosmological distance measurements.

The second part of the dissertation addressed a different but equally crucial challenge: improving the fidelity of galaxy formation history predictions in machine learning and semi-analytical models trained on dark matter simulations. Recognizing the architectural biases of standard regressions, we proposed a semi-stochastic correction framework that restores missing short-timescale variability in star formation and metallicity histories. By decomposing predicted histories in Fourier space and reinjecting statistically consistent fluctuations, the corrected models achieved better agreement with observables such as the stellar-halo mass relation, mass-metallicity relation, and color distributions. We further extended this framework to dark-matter-only simulations like TNG-Dark and Uchuu, demonstrating its generalizability and highlighting the role of halo assembly history and environment in accurate galaxy property inference.

In addition to these two major threads, the final chapter presented several complementary studies motivated by ongoing survey efforts. These include an angular multipole expansion of the bispectrum to probe anisotropic clustering, techniques to mitigate imaging systematics in DESI clustering measurements, and a spectroscopic classification framework for AGN and QSO tracers. While independent in implementation, these projects collectively reinforce the broader methodology of integrating theory, simulation, and data-driven analysis for robust cosmological inference.

Together, these studies highlight the importance of multiscale, simulation-informed modeling in large-scale structure research. The bispectrum analysis enhances the precision of cosmological measurements, the semi-stochastic framework improves the realism and utility of mock galaxy catalogs, and the complementary projects address practical challenges in interpreting and modeling real survey data. By connecting physical modeling, numerical simulation, and data science, this work contributes to advancing both the theoretical and observational frontiers of cosmology and galaxy evolution — supporting the goals of current and next-generation surveys.

“ଯାତ୍ରା ଶେଷ ନୁହେଁ, ବ୍ରହ୍ମାଣ୍ଡ ଏବଂ ଚିତ୍ତ ମଧ୍ୟରେ ଏକ ନିରନ୍ତର ସଂବାଦ।”

— Contemporary Odia philosophical perspective

“The journey is not over — it is a dialogue between the cosmos and the mind.”

Bibliography

[1] Lauren Anderson, Éric Aubourg, Stephen Bailey, Florian Beutler, Vaishali Bhardwaj, Michael Blanton, Adam S. Bolton, J. Brinkmann, Joel R. Brownstein, Angela Burden, Chia-Hsun Chuang, Antonio J. Cuesta, Kyle S. Dawson, Daniel J. Eisenstein, Stephanie Escoffier, James E. Gunn, Hong Guo, Shirley Ho, Klaus Honscheid, Cullan Howlett, David Kirkby, Robert H. Lupton, Marc Manera, Claudia Maraston, Cameron K. McBride, Olga Mena, Francesco Montesano, Robert C. Nichol, Sebastián E. Nuza, Matthew D. Olmstead, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, John Parejko, Will J. Percival, Patrick Petitjean, Francisco Prada, Adrian M. Price-Whelan, Beth Reid, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Cristiano G. Sabiu, Shun Saito, Lado Samushia, Ariel G. Sánchez, David J. Schlegel, Donald P. Schneider, Claudia G. Scoccola, Hee-Jong Seo, Ramin A. Skibba, Michael A. Strauss, Molly E. C. Swanson, Daniel Thomas, Jeremy L. Tinker, Rita Tojeiro, Mariana Vargas Magaña, Licia Verde, 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 Releases 10 and 11 Galaxy samples. , 441(1):24–62, June 2014. doi: 10.1093/mnras/stu523.

[2] Max Tegmark, Michael R. Blanton, Michael A. Strauss, Fiona Hoyle, David Schlegel, Roman Scoccimarro, Michael S. Vogeley, David H. Weinberg, Idit Zehavi, Andreas Berlind, Tamás Budavari, Andrew Connolly, Daniel J. Eisenstein, Douglas Finkbeiner, Joshua A. Frieman, James E. Gunn, Andrew J. S. Hamilton, Lam Hui, Bhuvnesh Jain, David Johnston, Stephen Kent, Huan Lin, Reiko Nakajima, Robert C. Nichol, Jeremiah P. Ostriker, Adrian Pope, Ryan Scranton, Uroš Seljak, Ravi K. Sheth, Albert Stebbins, Alexander S. Szalay, István Szapudi, Licia Verde, Yongzhong Xu, James

Annis, Neta A. Bahcall, J. Brinkmann, Scott Burles, Francisco J. Castander, Istvan Csabai, Jon Loveday, Mamoru Doi, Masataka Fukugita, J. Richard Gott, III, Greg Hennessy, David W. Hogg, Željko Ivezić, Gillian R. Knapp, Don Q. Lamb, Brian C. Lee, Robert H. Lupton, Timothy A. McKay, Peter Kunszt, Jeffrey A. Munn, Liam O'Connell, John Peoples, Jeffrey R. Pier, Michael Richmond, Constance Rockosi, Donald P. Schneider, Christopher Stoughton, Douglas L. Tucker, Daniel E. Vanden Berk, Brian Yanny, Donald G. York, and SDSS Collaboration. The Three-Dimensional Power Spectrum of Galaxies from the Sloan Digital Sky Survey. , 606(2):702–740, May 2004. doi: 10.1086/382125.

[3] N. W. Boggess, J. C. Mather, R. Weiss, C. L. Bennett, E. S. Cheng, E. Dwek, S. Gulkis, M. G. Hauser, M. A. Janssen, T. Kelsall, S. S. Meyer, S. H. Moseley, T. L. Murdock, R. A. Shafer, R. F. Silverberg, G. F. Smoot, D. T. Wilkinson, and E. L. Wright. The COBE Mission: Its Design and Performance Two Years after Launch. , 397:420, October 1992. doi: 10.1086/171797.

[4] C. L. Bennett, D. Larson, J. L. Weiland, N. Jarosik, G. Hinshaw, N. Odegard, K. M. Smith, R. S. Hill, B. Gold, M. Halpern, E. Komatsu, M. R. Nolta, L. Page, D. N. Spergel, E. Wollack, J. Dunkley, A. Kogut, M. Limon, S. S. Meyer, G. S. Tucker, and E. L. Wright. Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Final Maps and Results. , 208(2):20, October 2013. doi: 10.1088/0067-0049/208/2/20.

[5] Planck Collaboration, P. A. R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett, N. Bartolo, E. Battaner, R. Battye, K. Benabed, A. Benoît, A. Benoit-Lévy, J. P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, A. Bonaldi, L. Bonavera, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J. F. Cardoso, A. Catalano, A. Challinor, A. Chamballu, R. R. Chary, H. C. Chiang, J. Chluba, P. R. Christensen, S. Church, D. L. Clements, S. Colombi, L. P. L. Colombo, C. Com-

bet, A. Coulais, B. P. Crill, A. Curto, F. Cuttaia, L. Danese, R. D. Davies, R. J. Davis, P. de Bernardis, A. de Rosa, G. de Zotti, J. Delabrouille, F. X. Désert, E. Di Valentino, C. Dickinson, J. M. Diego, K. Dolag, H. Dole, S. Donzelli, O. Doré, M. Douspis, A. Ducout, J. Dunkley, X. Dupac, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, M. Farhang, J. Fergusson, F. Finelli, O. Forni, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frejsel, S. Galeotta, S. Galli, K. Ganga, C. Gauthier, M. Gerbino, T. Ghosh, M. Giard, Y. Giraud-Héraud, E. Giusarma, E. Gjerløw, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gregorio, A. Gruppuso, J. E. Gudmundsson, J. Hamann, F. K. Hansen, D. Hanson, D. L. Harrison, G. Helou, S. Henrot-Versillé, C. Hernández-Monteagudo, D. Herranz, S. R. Hildebrandt, E. Hivon, M. Hobson, W. A. Holmes, A. Hornstrup, W. Hovest, Z. Huang, K. M. Huffenberger, G. Hurier, A. H. Jaffe, T. R. Jaffe, W. C. Jones, M. Juvela, E. Keihänen, R. Keskitalo, T. S. Kisner, R. Kneissl, J. Knoche, L. Knox, M. Kunz, H. Kurki-Suonio, G. Lagache, A. Lähteenmäki, J. M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, J. P. Leahy, R. Leonardi, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Linden-Vørnle, M. López-Caniego, P. M. Lubin, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marchini, M. Maris, P. G. Martin, M. Martinelli, E. Martínez-González, S. Masi, S. Matarrese, P. McGehee, P. R. Meinhold, A. Melchiorri, J. B. Melin, L. Mendes, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M. A. Miville-Deschênes, A. Moneti, L. Montier, G. Morgante, D. Mortlock, A. Moss, D. Munshi, J. A. Murphy, P. Naselsky, F. Nati, P. Natoli, C. B. Netterfield, H. U. Nørgaard-Nielsen, F. Noviello, D. Novikov, I. Novikov, C. A. Oxborrow, F. Paci, L. Pagano, F. Pajot, R. Paladini, D. Paoletti, B. Partridge, F. Pasian, G. Patanchon, T. J. Pearson, O. Perdereau, L. Perotto, F. Perrotta, V. Pettorino, F. Piacentini, M. Piat, E. Pierpaoli, D. Pietrobon, S. Plaszczynski, E. Pointecouteau, G. Polenta, L. Popa, G. W. Pratt, and G. Prézeau. Planck 2015 results. XIII. Cosmological parameters. , 594:A13, September 2016. doi: 10.1051/0004-6361/201525830.

[6] Harry George Chittenden and Rita Tojeiro. Modelling the galaxy-halo connection

with semi-recurrent neural networks. , 518(4):5670–5692, January 2023. doi: 10.1093/mnras/stac3498.

[7] Naonori S. Sugiyama, Shun Saito, Florian Beutler, and Hee-Jong Seo. Towards a self-consistent analysis of the anisotropic galaxy two- and three-point correlation functions on large scales: application to mock galaxy catalogues. , 501(2):2862–2896, February 2021. doi: 10.1093/mnras/staa3725.

[8] P. J. E. Peebles and M. G. Hauser. Statistical Analysis of Catalogs of Extragalactic Objects. III. The Shane-Wirtanen and Zwicky Catalogs. , 28:19, November 1974. doi: 10.1086/190308.

[9] M. Davis and P. J. E. Peebles. A survey of galaxy redshifts. V. The two-point position and velocity correlations. , 267:465–482, April 1983. doi: 10.1086/160884.

[10] A. J. S. Hamilton. Toward Better Ways to Measure the Galaxy Correlation Function. , 417:19, November 1993. doi: 10.1086/173288.

[11] Stephen D. Landy and Alexander S. Szalay. Bias and Variance of Angular Correlation Functions. , 412:64, July 1993. doi: 10.1086/172900.

[12] Nick Kaiser. Clustering in real space and in redshift space. , 227:1–21, July 1987. doi: 10.1093/mnras/227.1.1.

[13] A. J. S. Hamilton. Linear Redshift Distortions: a Review. In Donald Hamilton, editor, *The Evolving Universe*, volume 231 of *Astrophysics and Space Science Library*, page 185, January 1998. doi: 10.1007/978-94-011-4960-0_17.

[14] Román Scoccimarro, H. M. P. Couchman, and Joshua A. Frieman. The Bispectrum as a Signature of Gravitational Instability in Redshift Space. , 517(2):531–540, June 1999. doi: 10.1086/307220.

[15] Emiliano Sefusatti and Eiichiro Komatsu. Bispectrum of galaxies from high-redshift

galaxy surveys: Primordial non-Gaussianity and nonlinear galaxy bias. , 76(8):083004, October 2007. doi: 10.1103/PhysRevD.76.083004.

[16] Nico Hamaus, Benjamin D. Wandelt, P. M. Sutter, Guilhem Lavaux, and Michael S. Warren. Cosmology with Void-Galaxy Correlations. , 112(4):041304, January 2014. doi: 10.1103/PhysRevLett.112.041304.

[17] P. M. Sutter, Guilhem Lavaux, Benjamin D. Wandelt, and David H. Weinberg. A First Application of the Alcock-Paczynski Test to Stacked Cosmic Voids. , 761(2):187, December 2012. doi: 10.1088/0004-637X/761/2/187.

[18] Ravi K. Sheth. The halo-model description of marked statistics. , 364(3):796–806, December 2005. doi: 10.1111/j.1365-2966.2005.09609.x.

[19] Martin White. A marked correlation function for constraining modified gravity models. , 2016(11):057, November 2016. doi: 10.1088/1475-7516/2016/11/057.

[20] Abel Yang and William C. Saslaw. The Galaxy Counts-in-cells Distribution from the Sloan Digital Sky Survey. , 729(2):123, March 2011. doi: 10.1088/0004-637X/729/2/123.

[21] Chiaki Hikage, Jens Schmalzing, Thomas Buchert, Yasushi Suto, Issha Kayo, Atsushi Taruya, Michael S. Vogeley, Fiona Hoyle, J. Richard Gott, III, and Jon Brinkmann. Minkowski Functionals of SDSS Galaxies I : Analysis of Excursion Sets. , 55:911–931, October 2003. doi: 10.1093/pasj/55.5.911.

[22] Geraint Pratten and Dipak Munshi. Non-Gaussianity in large-scale structure and Minkowski functionals. , 423(4):3209–3226, July 2012. doi: 10.1111/j.1365-2966.2012.21103.x.

[23] Pratyush Pranav, Herbert Edelsbrunner, Rien van de Weygaert, Gert Vegter, Michael Kerber, Bernard J. T. Jones, and Mathijs Wintraecken. The topology of the cosmic web in terms of persistent Betti numbers. , 465(4):4281–4310, March 2017. doi: 10.1093/mnras/stw2862.

[24] Antony Lewis and Sarah Bridle. Cosmological parameters from CMB and other data: A Monte Carlo approach. , 66:103511, 2002. doi: 10.1103/PhysRevD.66.103511.

[25] Diego Blas, Julien Lesgourgues, and Thomas Tram. The Cosmic Linear Anisotropy Solving System (CLASS). Part II: Approximation schemes. , 2011(7):034, July 2011. doi: 10.1088/1475-7516/2011/07/034.

[26] Florian Beutler, Hee-Jong Seo, Ashley J. Ross, Patrick McDonald, Shun Saito, Adam S. Bolton, Joel R. Brownstein, Chia-Hsun Chuang, Antonio J. Cuesta, Daniel J. Eisenstein, Andreu Font-Ribera, Jan Niklas Grieb, Nick Hand, Francisco-Shu Kitaura, Chirag Modi, Robert C. Nichol, Will J. Percival, Francisco Prada, Sergio Rodriguez-Torres, Natalie A. Roe, Nicholas P. Ross, Salvador Salazar-Albornoz, Ariel G. Sánchez, Donald P. Schneider, Anže Slosar, Jeremy Tinker, Rita Tojeiro, Mariana Vargas-Magaña, and Jose A. Vazquez. The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Fourier space. , 464(3):3409–3430, January 2017. doi: 10.1093/mnras/stw2373.

[27] Will J. Percival. Large Scale Structure Observations. *arXiv e-prints*, art. arXiv:1312.5490, December 2013. doi: 10.48550/arXiv.1312.5490.

[28] Alison L. Coil. The Large-Scale Structure of the Universe. In Terry D. Oswalt and William C. Keel, editors, *Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology*, volume 6, page 387. 2013. doi: 10.1007/978-94-007-5609-0_8.

[29] Shadab Alam, Marie Aubert, Santiago Avila, Christophe Balland, Julian E. Bautista, Matthew A. Bershady, Dmitry Bizyaev, Michael R. Blanton, Adam S. Bolton, Jo Bovy, Jonathan Brinkmann, Joel R. Brownstein, Etienne Burtin, Solène Chabanier, Michael J. Chapman, Peter Doohyun Choi, Chia-Hsun Chuang, Johan Comparat, Marie-Claude Cousinou, Andrei Cuceu, Kyle S. Dawson, Sylvain de la Torre, Arnaud de Mattia, Victoria de Sainte Agathe, Hélion du Mas des Bourboux, Stephanie Escoffier, Thomas Etourneau, James Farr, Andreu Font-Ribera, Peter M. Frinchaboy, Se-

bastien Fromenteau, Héctor Gil-Marín, Jean-Marc Le Goff, Alma X. Gonzalez-Morales, Violeta Gonzalez-Perez, Kathleen Grabowski, Julien Guy, Adam J. Hawken, Jiamin Hou, Hui Kong, James Parker, Mark Klaene, Jean-Paul Kneib, Sicheng Lin, Daniel Long, Brad W. Lyke, Axel de la Macorra, Paul Martini, Karen Masters, Faizan G. Mohammad, Jeongin Moon, Eva-Maria Mueller, Andrea Muñoz-Gutiérrez, Adam D. Myers, Seshadri Nadathur, Richard Neveux, Jeffrey A. Newman, Pasquier Noterdaeme, Audrey Oravetz, Daniel Oravetz, Nathalie Palanque-Delabrouille, Kaike Pan, Romain Paviot, Will J. Percival, Ignasi Pérez-Ràfols, Patrick Petitjean, Matthew M. Pieri, Abhishek Prakash, Anand Raichoor, Corentin Ravoux, Mehdi Rezaie, James Rich, Ashley J. Ross, Graziano Rossi, Rossana Ruggeri, Vanina Ruhlmann-Kleider, Ariel G. Sánchez, F. Javier Sánchez, José R. Sánchez-Gallego, Conor Sayres, Donald P. Schneider, Hee-Jong Seo, Arman Shafieloo, Anže Slosar, Alex Smith, Julianna Stermer, Amelie Tamone, Jeremy L. Tinker, Rita Tojeiro, Mariana Vargas-Magaña, Andrei Variu, Yuting Wang, Benjamin A. Weaver, Anne-Marie Weijmans, Christophe Yèche, Pauline Zarrouk, Cheng Zhao, Gong-Bo Zhao, and Zheng Zheng. Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Cosmological implications from two decades of spectroscopic surveys at the Apache Point Observatory. , 103(8):083533, April 2021. doi: 10.1103/PhysRevD.103.083533.

[30] Cheng Zhao, Andrei Variu, Mengfan He, Daniel Forero-Sánchez, Amélie Tamone, Chia-Hsun Chuang, Francisco-Shu Kitaura, Charling Tao, Jiaxi Yu, Jean-Paul Kneib, Will J. Percival, Huanyuan Shan, Gong-Bo Zhao, Etienne Burtin, Kyle S. Dawson, Graziano Rossi, Donald P. Schneider, and Axel de la Macorra. The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: cosmological implications from multitracer BAO analysis with galaxies and voids. , 511(4):5492–5524, April 2022. doi: 10.1093/mnras/stac390.

[31] Donald G. York, J. Adelman, John E. Anderson, Jr., Scott F. Anderson, James Annis, Neta A. Bahcall, J. A. Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, William N. Boroski, Steve Bracker, Charlie Briegel, John W. Briggs, J. Brinkmann,

Robert Brunner, Scott Burles, Larry Carey, Michael A. Carr, Francisco J. Castander, Bing Chen, Patrick L. Colestock, A. J. Connolly, J. H. Crocker, István Csabai, Paul C. Czarapata, John Eric Davis, Mamoru Doi, Tom Dombeck, Daniel Eisenstein, Nancy Ellman, Brian R. Elms, Michael L. Evans, Xiaohui Fan, Glenn R. Federwitz, Larry Fiscelli, Scott Friedman, Joshua A. Frieman, Masataka Fukugita, Bruce Gillespie, James E. Gunn, Vijay K. Gurbani, Ernst de Haas, Merle Haldeman, Frederick H. Harris, J. Hayes, Timothy M. Heckman, G. S. Hennessy, Robert B. Hindsley, Scott Holm, Donald J. Holmgren, Chi-hao Huang, Charles Hull, Don Husby, Shin-Ichi Ichikawa, Takashi Ichikawa, Željko Ivezić, Stephen Kent, Rita S. J. Kim, E. Kinney, Mark Klaene, A. N. Kleinman, S. Kleinman, G. R. Knapp, John Korienek, Richard G. Kron, Peter Z. Kunszt, D. Q. Lamb, B. Lee, R. French Leger, Siriluk Limmongkol, Carl Lindenmeyer, Daniel C. Long, Craig Loomis, Jon Loveday, Rich Lucinio, Robert H. Lupton, Bryan MacKinnon, Edward J. Mannery, P. M. Mantsch, Bruce Margon, Peregrine McGehee, Timothy A. McKay, Avery Meiksin, Aronne Merelli, David G. Monet, Jeffrey A. Munn, Vijay K. Narayanan, Thomas Nash, Eric Neilsen, Rich Neswold, Heidi Jo Newberg, R. C. Nichol, Tom Nicinski, Mario Nonino, Norio Okada, Sadanori Okamura, Jeremiah P. Ostriker, Russell Owen, A. George Pauls, John Peoples, R. L. Peterson, Donald Petravick, Jeffrey R. Pier, Adrian Pope, Ruth Pordes, Angela Prosapio, Ron Rechenmacher, Thomas R. Quinn, Gordon T. Richards, Michael W. Richmond, Claudio H. Rivetta, Constance M. Rockosi, Kurt Ruthmansdorfer, Dale Sandford, David J. Schlegel, Donald P. Schneider, Maki Sekiguchi, Gary Sergey, Kazuhiro Shimasaku, Walter A. Siegmund, Stephen Smee, J. Allyn Smith, S. Snedden, R. Stone, Chris Stoughton, Michael A. Strauss, Christopher Stubbs, Mark SubbaRao, Alexander S. Szalay, Istvan Szapudi, Gyula P. Szokoly, Anirudda R. Thakar, Christy Tremonti, Douglas L. Tucker, Alan Uomoto, Dan Vanden Berk, Michael S. Vogeley, Patrick Waddell, Shu-i. Wang, Masaru Watanabe, David H. Weinberg, Brian Yanny, Naoki Yasuda, and SDSS Collaboration. The Sloan Digital Sky Survey: Technical Summary. , 120(3):1579–1587, September 2000. doi: 10.1086/301513.

[32] DESI Collaboration, M. Abdul-Karim, J. Aguilar, S. Ahlen, S. Alam, L. Allen, C. Allende Prieto, O. Alves, A. Anand, U. Andrade, E. Armengaud, A. Aviles, S. Bailey, C. Baltay, P. Bansal, A. Bault, J. Behera, S. BenZvi, D. Bianchi, C. Blake, S. Brieden, A. Brodzeller, D. Brooks, E. Buckley-Geer, E. Burtin, R. Calderon, R. Canning, A. Carnero Rosell, P. Carrilho, L. Casas, F. J. Castander, R. Cereskaite, M. Charles, E. Chaussidon, J. Chaves-Montero, D. Chebat, X. Chen, T. Claybaugh, S. Cole, A. P. Cooper, A. Cuceu, K. S. Dawson, A. de la Macorra, A. de Mattia, N. Deiosso, J. Della Costa, R. Demina, A. Dey, B. Dey, Z. Ding, P. Doel, J. Edelstein, D. J. Eisenstein, W. Elbers, P. Fagrelius, K. Fanning, E. Fern\'andez-Garc\'ia, S. Ferraro, A. Font-Ribera, J. E. Forero-Romero, C. S. Frenk, C. Garcia-Quintero, L. H. Garrison, E. Gazta\, H. Gil-Mar\'in, S. Gontcho A Gontcho, D. Gonzalez, A. X. Gonzalez-Morales, C. Gordon, D. Green, G. Gutierrez, J. Guy, B. Hadzhiyska, C. Hahn, S. He, M. Herbold, H. K. Herrera-Alcantar, M. Ho, K. Honscheid, C. Howlett, D. Huterer, M. Ishak, S. Juneau, N. V. Kamble, N. G. Kara{c}ayl{}, R. Kehoe, S. Kent, A. G. Kim, D. Kirkby, T. Kisner, S. E. Koposov, A. Kremin, A. Krolewski, O. Lahav, C. Lamman, M. Landriau, D. Lang, J. Lasker, J. M. Le Goff, L. Le Guillou, A. Leauthaud, M. E. Levi, Q. Li, T. S. Li, K. Lodha, M. Lokken, F. Lozano-Rodr\'iguez, C. Magneville, M. Manera, P. Martini, W. L. Matthewson, A. Meisner, J. Mena-Fern\'andez, A. Menegas, T. Mergulh\, R. Miquel, J. Moustakas, A. Mu\-Guti\'errez, D. Mu\-Santos, A. D. Myers, S. Nadathur, K. Naidoo, L. Napolitano, J. A. Newman, G. Niz, H. E. Noriega, E. Paillas, N. Palanque-Delabrouille, J. Pan, J. Peacock, Marcos Pellejero Ibanez, W. J. Percival, A. P\'erez-Fern\'andez, I. P\'erez-R\`afols, M. M. Pieri, C. Poppett, F. Prada, D. Rabinowitz, A. Raichoor, C. Ram\'irez-P\'erez, M. Rashkovetskyi, C. Ravoux, J. Rich, A. Rocher, C. Rockosi, J. Rohlf, J. O. Rom\'an-Herrera, A. J. Ross, G. Rossi, R. Ruggeri, V. Ruhlmann-Kleider, L. Samushia, E. Sanchez, N. Sanders, D. Schlegel, M. Schubnell, H. Seo, A. Shafieloo, R. Sharples, J. Silber, F. Sinigaglia, D. Sprayberry, T. Tan, G. Tarl\'e, P. Taylor, W. Turner, L. A. Ure\-L\'opez, R. Vaisakh, F. Valdes, G. Valogiannis, M. Vargas-Maga\, L. Verde, M. Walther, B. A. Weaver, D. H. Weinberg, M. White, M. Wolfson,

C. Y\`eche, J. Yu, E. A. Zaborowski, P. Zarrouk, Z. Zhai, H. Zhang, C. Zhao, G. B. Zhao, R. Zhou, and H. Zou. DESI DR2 Results II: Measurements of Baryon Acoustic Oscillations and Cosmological Constraints. *arXiv e-prints*, art. arXiv:2503.14738, March 2025. doi: 10.48550/arXiv.2503.14738.

[33] Daniel J. Eisenstein, Hee-Jong Seo, and Martin White. On the Robustness of the Acoustic Scale in the Low-Redshift Clustering of Matter. , 664(2):660–674, August 2007. doi: 10.1086/518755.

[34] A. Burden, W. J. Percival, M. Manera, Antonio J. Cuesta, Mariana Vargas Magana, and Shirley Ho. Efficient reconstruction of linear baryon acoustic oscillations in galaxy surveys. , 445(3):3152–3168, December 2014. doi: 10.1093/mnras/stu1965.

[35] Ashley J. Ross, Florian Beutler, Chia-Hsun Chuang, Marcos Pellejero-Ibanez, Hee-Jong Seo, Mariana Vargas-Magaña, Antonio J. Cuesta, Will J. Percival, Angela Burden, Ariel G. Sánchez, Jan Niklas Grieb, Beth Reid, Joel R. Brownstein, Kyle S. Dawson, Daniel J. Eisenstein, Shirley Ho, Francisco-Shu Kitaura, Robert C. Nichol, Matthew D. Olmstead, Francisco Prada, Sergio A. Rodríguez-Torres, Shun Saito, Salvador Salazar-Albornoz, Donald P. Schneider, Daniel Thomas, Jeremy Tinker, Rita Tojeiro, Yuting Wang, Martin White, and Gong-bo Zhao. The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: observational systematics and baryon acoustic oscillations in the correlation function. , 464(1):1168–1191, January 2017. doi: 10.1093/mnras/stw2372.

[36] Mariana Vargas-Magaña, Shirley Ho, Antonio J. Cuesta, Ross O'Connell, Ashley J. Ross, Daniel J. Eisenstein, Will J. Percival, Jan Niklas Grieb, Ariel G. Sánchez, Jeremy L. Tinker, Rita Tojeiro, Florian Beutler, Chia-Hsun Chuang, Francisco-Shu Kitaura, Francisco Prada, Sergio A. Rodríguez-Torres, Graziano Rossi, Hee-Jong Seo, Joel R. Brownstein, Matthew Olmstead, and Daniel Thomas. The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: theoretical

systematics and Baryon Acoustic Oscillations in the galaxy correlation function. , 477 (1):1153–1188, June 2018. doi: 10.1093/mnras/sty571.

[37] Eiichiro Komatsu and David N. Spergel. Acoustic signatures in the primary microwave background bispectrum. , 63(6):063002, March 2001. doi: 10.1103/PhysRevD.63.063002.

[38] Román Scoccimarro. Redshift-space distortions, pairwise velocities, and nonlinearities. , 70(8):083007, October 2004. doi: 10.1103/PhysRevD.70.083007.

[39] Héctor Gil-Marín, Will J. Percival, Licia Verde, Joel R. Brownstein, Chia-Hsun Chuang, Francisco-Shu Kitaura, Sergio A. Rodríguez-Torres, and Matthew D. Olmstead. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: RSD measurement from the power spectrum and bispectrum of the DR12 BOSS galaxies. , 465(2):1757–1788, February 2017. doi: 10.1093/mnras/stw2679.

[40] Federico Rizzo, Chiara Moretti, Kevin Pardede, Alexander Eggemeier, Andrea Oddo, Emiliano Sefusatti, Cristiano Porciani, and Pierluigi Monaco. The halo bispectrum multipoles in redshift space. , 2023(1):031, January 2023. doi: 10.1088/1475-7516/2023/01/031.

[41] Emiliano Sefusatti, Martín Crocce, Sebastián Pueblas, and Román Scoccimarro. Cosmology and the bispectrum. , 74(2):023522, July 2006. doi: 10.1103/PhysRevD.74.023522.

[42] David W. Pearson and Lado Samushia. A Detection of the Baryon Acoustic Oscillation features in the SDSS BOSS DR12 Galaxy Bispectrum. , 478(4):4500–4512, August 2018. doi: 10.1093/mnras/sty1266.

[43] Román Scoccimarro. The Bispectrum: From Theory to Observations. , 544(2):597–615, December 2000. doi: 10.1086/317248.

[44] Donghui Jeong and Eiichiro Komatsu. Primordial Non-Gaussianity, Scale-dependent

Bias, and the Bispectrum of Galaxies. , 703(2):1230–1248, October 2009. doi: 10.1088/0004-637X/703/2/1230.

[45] Tobias Baldauf, Lorenzo Mercolli, Mehrdad Mirbabayi, and Enrico Pajer. The bispectrum in the Effective Field Theory of Large Scale Structure. , 2015(5):007–007, May 2015. doi: 10.1088/1475-7516/2015/05/007.

[46] Mikhail M. Ivanov, Oliver H. E. Philcox, Takahiro Nishimichi, Marko Simonović, Masahiro Takada, and Matias Zaldarriaga. Precision analysis of the redshift-space galaxy bispectrum. , 105(6):063512, March 2022. doi: 10.1103/PhysRevD.105.063512.

[47] Davide Gualdi, Héctor Gil-Marín, and Licia Verde. Joint analysis of anisotropic power spectrum, bispectrum and trispectrum: application to N-body simulations. , 2021(7):008, July 2021. doi: 10.1088/1475-7516/2021/07/008.

[48] Mike Wang, Florian Beutler, and Naonori Sugiyama. Triumvirate: A Python/C++ package for three-point clustering measurements. *The Journal of Open Source Software*, 8(91):5571, November 2023. doi: 10.21105/joss.05571.

[49] Oliver H. E. Philcox and Mikhail M. Ivanov. BOSS DR12 full-shape cosmology: Λ CDM constraints from the large-scale galaxy power spectrum and bispectrum monopole. , 105(4):043517, February 2022. doi: 10.1103/PhysRevD.105.043517.

[50] Naonori S. Sugiyama, Shun Saito, Florian Beutler, and Hee-Jong Seo. Perturbation theory approach to predict the covariance matrices of the galaxy power spectrum and bispectrum in redshift space. , 497(2):1684–1711, September 2020. doi: 10.1093/mnras/staa1940.

[51] Anatoly Klypin and Francisco Prada. Dark matter statistics for large galaxy catalogues: power spectra and covariance matrices. , 478(4):4602–4621, August 2018. doi: 10.1093/mnras/sty1340.

[52] C. M. Baugh, Violeta Gonzalez-Perez, Claudia D. P. Lagos, Cedric G. Lacey, John C. Helly, Adrian Jenkins, Carlos S. Frenk, Andrew J. Benson, Richard G. Bower, and

Shaun Cole. Galaxy formation in the Planck Millennium: the atomic hydrogen content of dark matter haloes. , 483(4):4922–4937, March 2019. doi: 10.1093/mnras/sty3427.

[53] Jayashree Behera, Mehdi Rezaie, Lado Samushia, and Julia Ereza. Modelling the BAO feature in bispectrum. , 531(3):3326–3335, July 2024. doi: 10.1093/mnras/stae1161.

[54] Tobias Baldauf, Mehrdad Mirbabayi, Marko Simonović, and Matias Zaldarriaga. Equivalence principle and the baryon acoustic peak. , 92(4):043514, August 2015. doi: 10.1103/PhysRevD.92.043514.

[55] Zvonimir Vlah, Martin White, and Alejandro Aviles. A Lagrangian effective field theory. , 2015(9):014–014, September 2015. doi: 10.1088/1475-7516/2015/09/014.

[56] Lucía Fonseca de la Bella, Donough Regan, David Seery, and Shaun Hotchkiss. The matter power spectrum in redshift space using effective field theory. , 2017(11):039, November 2017. doi: 10.1088/1475-7516/2017/11/039.

[57] Hidenori Nomura, Kazuhiro Yamamoto, and Takahiro Nishimichi. Damping of the baryon acoustic oscillations in the matter power spectrum as a probe of the growth factor. , 2008(10):031, October 2008. doi: 10.1088/1475-7516/2008/10/031.

[58] Martín. Crocce, Román. Scoccimarro, and Francis Bernardeau. MPTBREEZE: a fast renormalized perturbative scheme. , 427(3):2537–2551, December 2012. doi: 10.1111/j.1365-2966.2012.22127.x.

[59] Francisco Prada, Claudia G. Scóccola, Chia-Hsun Chuang, Gustavo Yepes, Anatoly A. Klypin, Francisco-Shu Kitaura, Stefan Gottlöber, and Cheng Zhao. Hunting down systematics in baryon acoustic oscillations after cosmic high noon. , 458(1):613–623, May 2016. doi: 10.1093/mnras/stw312.

[60] Hee-Jong Seo, Florian Beutler, Ashley J. Ross, and Shun Saito. Modeling the reconstructed BAO in Fourier space. , 460(3):2453–2471, August 2016. doi: 10.1093/mnras/stw1138.

[61] K. Hoffmann, E. Gaztañaga, R. Scoccimarro, and M. Crocce. Testing the consistency of three-point halo clustering in Fourier and configuration space. , 476(1):814–829, May 2018. doi: 10.1093/mnras/sty187.

[62] Shi-Fan Chen, Zvonimir Vlah, and Martin White. The reconstructed power spectrum in the Zeldovich approximation. , 2019(9):017, September 2019. doi: 10.1088/1475-7516/2019/09/017.

[63] Samuel R. Hinton, Cullan Howlett, and Tamara M. Davis. BARRY and the BAO model comparison. , 493(3):4078–4093, April 2020. doi: 10.1093/mnras/staa361.

[64] Patrick Valageas and Takahiro Nishimichi. Lagrangian-space Gaussian ansatz for the matter redshift-space power spectrum and correlation function. , 102(4):043530, August 2020. doi: 10.1103/PhysRevD.102.043530.

[65] Ivana Babić, Fabian Schmidt, and Beatriz Tucci. BAO scale inference from biased tracers using the EFT likelihood. , 2022(8):007, August 2022. doi: 10.1088/1475-7516/2022/08/007.

[66] Metin Ata, Falk Baumgarten, Julian Bautista, Florian Beutler, Dmitry Bizyaev, Michael R. Blanton, Jonathan A. Blazek, Adam S. Bolton, Jonathan Brinkmann, Joel R. Brownstein, Etienne Burtin, Chia-Hsun Chuang, Johan Comparat, Kyle S. Dawson, Axel de la Macorra, Wei Du, Hélion du Mas des Bourboux, Daniel J. Eisenstein, Héctor Gil-Marín, Katie Grabowski, Julien Guy, Nick Hand, Shirley Ho, Timothy A. Hutchinson, Mikhail M. Ivanov, Francisco-Shu Kitaura, Jean-Paul Kneib, Pierre Laurent, Jean-Marc Le Goff, Joseph E. McEwen, Eva-Maria Mueller, Adam D. Myers, Jeffrey A. Newman, Nathalie Palanque-Delabrouille, Kaike Pan, Isabelle Pâris, Marcos Pellejero-Ibanez, Will J. Percival, Patrick Petitjean, Francisco Prada, Abhishek Prakash, Sergio A. Rodríguez-Torres, Ashley J. Ross, Graziano Rossi, Rossana Ruggeri, Ariel G. Sánchez, Siddharth Satpathy, David J. Schlegel, Donald P. Schneider, Hee-Jong Seo, Anže Slosar, Alina Streblyanska, Jeremy L. Tinker, Rita Tojeiro, Mariana Vargas Magaña, M. Vivek, Yuting Wang, Christophe Yèche, Liang Yu, Pauline

Zarrouk, Cheng Zhao, Gong-Bo Zhao, and Fangzhou Zhu. The clustering of the SDSS-IV extended Baryon Oscillation Spectroscopic Survey DR14 quasar sample: first measurement of baryon acoustic oscillations between redshift 0.8 and 2.2. , 473(4):4773–4794, February 2018. doi: 10.1093/mnras/stx2630.

[67] Jiamin Hou, Ariel G. Sánchez, Ashley J. Ross, Alex Smith, Richard Neveux, Julian Bautista, Etienne Burtin, Cheng Zhao, Román Scoccimarro, Kyle S. Dawson, Arnaud de Mattia, Axel de la Macorra, Hélion du Mas des Bourboux, Daniel J. Eisenstein, Héctor Gil-Marín, Brad W. Lyke, Faizan G. Mohammad, Eva-Maria Mueller, Will J. Percival, Graziano Rossi, Mariana Vargas Magaña, Pauline Zarrouk, Gong-Bo Zhao, Jonathan Brinkmann, Joel R. Brownstein, Chia-Hsun Chuang, Adam D. Myers, Jeffrey A. Newman, Donald P. Schneider, and M. Vivek. The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: BAO and RSD measurements from anisotropic clustering analysis of the quasar sample in configuration space between redshift 0.8 and 2.2. , 500(1):1201–1221, January 2021. doi: 10.1093/mnras/staa3234.

[68] Anand Raichoor, Arnaud de Mattia, Ashley J. Ross, Cheng Zhao, Shadab Alam, Santiago Avila, Julian Bautista, Jonathan Brinkmann, Joel R. Brownstein, Etienne Burtin, Michael J. Chapman, Chia-Hsun Chuang, Johan Comparat, Kyle S. Dawson, Arjun Dey, Hélion du Mas des Bourboux, Jack Elvin-Poole, Violeta Gonzalez-Perez, Claudio Gorgoni, Jean-Paul Kneib, Hui Kong, Dustin Lang, John Moustakas, Adam D. Myers, Eva-Maria Müller, Seshadri Nadathur, Jeffrey A. Newman, Will J. Percival, Mehdi Rezaie, Graziano Rossi, Vanina Ruhlmann-Kleider, David J. Schlegel, Donald P. Schneider, Hee-Jong Seo, Amélie Tamone, Jeremy L. Tinker, Rita Tojeiro, M. Vivek, Christophe Yèche, and Gong-Bo Zhao. The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: large-scale structure catalogues and measurement of the isotropic BAO between redshift 0.6 and 1.1 for the Emission Line Galaxy Sample. , 500(3):3254–3274, January 2021. doi: 10.1093/mnras/staa3336.

[69] Nikhil Padmanabhan and Martin White. Constraining anisotropic baryon oscillations.

, 77(12):123540, June 2008. doi: 10.1103/PhysRevD.77.123540.

[70] Beth A. Reid, Lado Samushia, Martin White, Will J. Percival, Marc Manera, Nikhil Padmanabhan, Ashley J. Ross, Ariel G. Sánchez, Stephen Bailey, Dmitry Bizyaev, Adam S. Bolton, Howard Brewington, J. Brinkmann, Joel R. Brownstein, Antonio J. Cuesta, Daniel J. Eisenstein, James E. Gunn, Klaus Honscheid, Elena Malanushenko, Viktor Malanushenko, Claudia Maraston, Cameron K. McBride, Demitri Muna, Robert C. Nichol, Daniel Oravetz, Kaike Pan, Roland de Putter, N. A. Roe, Nicholas P. Ross, David J. Schlegel, Donald P. Schneider, Hee-Jong Seo, Alaina Shelden, Erin S. Sheldon, Audrey Simmons, Ramin A. Skibba, Stephanie Snedden, Molly E. C. Swanson, Daniel Thomas, Jeremy Tinker, Rita Tojeiro, Licia Verde, David A. Wake, Benjamin A. Weaver, David H. Weinberg, Idit Zehavi, and Gong-Bo Zhao. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: measurements of the growth of structure and expansion rate at z = 0.57 from anisotropic clustering. , 426(4):2719–2737, November 2012. doi: 10.1111/j.1365-2966.2012.21779.x.

[71] Jonathan Blazek, Uroš Seljak, Zvonimir Vlah, and Teppei Okumura. Geometric and dynamic distortions in anisotropic galaxy clustering. , 2014(4):001, April 2014. doi: 10.1088/1475-7516/2014/04/001.

[72] Ashley J. Ross, Will J. Percival, and Marc Manera. The information content of anisotropic Baryon Acoustic Oscillation scale measurements. , 451(2):1331–1340, August 2015. doi: 10.1093/mnras/stv966.

[73] Florian Beutler, Hee-Jong Seo, Shun Saito, Chia-Hsun Chuang, Antonio J. Cuesta, Daniel J. Eisenstein, Héctor Gil-Marín, Jan Niklas Grieb, Nick Hand, Francisco-Shu Kitaura, Chirag Modi, Robert C. Nichol, Matthew D. Olmstead, Will J. Percival, Francisco Prada, Ariel G. Sánchez, Sergio Rodriguez-Torres, Ashley J. Ross, Nicholas P. Ross, Donald P. Schneider, Jeremy Tinker, Rita Tojeiro, and Mariana Vargas-Magaña. The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic

Survey: anisotropic galaxy clustering in Fourier space. , 466(2):2242–2260, April 2017. doi: 10.1093/mnras/stw3298.

[74] Fangzhou Zhu, Nikhil Padmanabhan, Ashley J. Ross, Martin White, Will J. Percival, Rossana Ruggeri, Gong-bo Zhao, Dandan Wang, Eva-Maria Mueller, Etienne Burtin, Héctor Gil-Marín, Julian Bautista, Florian Beutler, Jonathan Brinkmann, Joel R. Brownstein, Kyle Dawson, Axel de la Macorra, Graziano Rossi, Donald P. Schneider, Rita Tojeiro, and Yuting Wang. The clustering of theSDSS-IV extended Baryon Oscillation Spectroscopic Survey DR14 quasar sample: measuring the anisotropic baryon acoustic oscillations with redshift weights. , 480(1):1096–1105, October 2018. doi: 10.1093/mnras/sty1955.

[75] Jiamin Hou, Ariel G. Sánchez, Román Scoccimarro, Salvador Salazar-Albornoz, Etienne Burtin, Héctor Gil-Marín, Will J. Percival, Rossana Ruggeri, Pauline Zarrouk, Gong-Bo Zhao, Julian Bautista, Jonathan Brinkmann, Joel R. Brownstein, Kyle S. Dawson, N. Chandrachani Devi, Adam D. Myers, Salman Habib, Katrin Heitmann, Rita Tojeiro, Graziano Rossi, Donald P. Schneider, Hee-Jong Seo, and Yuting Wang. The clustering of the SDSS-IV extended Baryon Oscillation Spectroscopic Survey DR14 quasar sample: anisotropic clustering analysis in configuration space. , 480(2):2521–2534, October 2018. doi: 10.1093/mnras/sty1984.

[76] Mariana Vargas Magaña, Shirley Ho, Xiaoying Xu, Ariel G. Sánchez, Ross O'Connell, Daniel J. Eisenstein, Antonio J. Cuesta, Will J. Percival, Ashley J. Ross, Eric Aubourg, Stéphanie Escoffier David Kirkby, Marc Manera, Donald P. Schneider, Jeremy L. Tinker, and Benjamin A. Weaver. SDSS-III Baryon Oscillation Spectroscopic Survey: Analysis of Potential Systematics in Fitting of Baryon Acoustic Feature. *arXiv e-prints*, art. arXiv:1312.4996, December 2013. doi: 10.48550/arXiv.1312.4996.

[77] Kiattisak Thepsuriya and Antony Lewis. Accuracy of cosmological parameters using the baryon acoustic scale. , 2015(1):034–034, January 2015. doi: 10.1088/1475-7516/2015/01/034.

[78] Zhejie Ding, Hee-Jong Seo, Zvonimir Vlah, Yu Feng, Marcel Schmittfull, and Florian Beutler. Theoretical systematics of Future Baryon Acoustic Oscillation Surveys. , 479 (1):1021–1054, September 2018. doi: 10.1093/mnras/sty1413.

[79] Mariana Vargas-Magaña, Shirley Ho, Xiaoying Xu, Ariel G. Sánchez, Ross O'Connell, Daniel J. Eisenstein, Antonio J. Cuesta, Will J. Percival, Ashley J. Ross, Eric Aubourg, Joel R. Brownstein, Stéphanie Escoffier, David Kirkby, Marc Manera, Donald P. Schneider, Jeremy L. Tinker, and Benjamin A. Weaver. The clustering of Galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: potential systematics in fitting of baryon acoustic feature. , 445(1):2–28, November 2014. doi: 10.1093/mnras/stu1681.

[80] Richard Neveux, Etienne Burtin, Arnaud de Mattia, Alex Smith, Ashley J. Ross, Jiamin Hou, Julian Bautista, Jonathan Brinkmann, Chia-Hsun Chuang, Kyle S. Dawson, Héctor Gil-Marín, Brad W. Lyke, Axel de la Macorra, Hélion du Mas des Bourboux, Faizan G. Mohammad, Eva-Maria Müller, Adam D. Myers, Jeffrey A. Newman, Will J. Percival, Graziano Rossi, Donald Schneider, M. Vivek, Pauline Zarrouk, Cheng Zhao, and Gong-Bo Zhao. The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: BAO and RSD measurements from the anisotropic power spectrum of the quasar sample between redshift 0.8 and 2.2. , 499(1):210–229, November 2020. doi: 10.1093/mnras/staa2780.

[81] Julian E. Bautista, Romain Paviot, Mariana Vargas Magaña, Sylvain de la Torre, Sebastien Fromenteau, Hector Gil-Marín, Ashley J. Ross, Etienne Burtin, Kyle S. Dawson, Jiamin Hou, Jean-Paul Kneib, Arnaud de Mattia, Will J. Percival, Graziano Rossi, Rita Tojeiro, Cheng Zhao, Gong-Bo Zhao, Shadab Alam, Joel Brownstein, Michael J. Chapman, Peter D. Choi, Chia-Hsun Chuang, Stéphanie Escoffier, Axel de la Macorra, Hélion du Mas des Bourboux, Faizan G. Mohammad, Jeongin Moon, Eva-Maria Müller, Seshadri Nadathur, Jeffrey A. Newman, Donald Schneider, Hee-Jong Seo, and Yuting Wang. The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: measurement of the BAO and growth rate of structure of the

luminous red galaxy sample from the anisotropic correlation function between redshifts 0.6 and 1. , 500(1):736–762, January 2021. doi: 10.1093/mnras/staa2800.

[82] Arnaud de Mattia, Vanina Ruhlmann-Kleider, Anand Raichoor, Ashley J. Ross, Amélie Tamone, Cheng Zhao, Shadab Alam, Santiago Avila, Etienne Burtin, Julian Bautista, Florian Beutler, Jonathan Brinkmann, Joel R. Brownstein, Michael J. Chapman, Chia-Hsun Chuang, Johan Comparat, Hélion du Mas des Bourboux, Kyle S. Dawson, Axel de la Macorra, Héctor Gil-Marín, Violeta Gonzalez-Perez, Claudio Gorgoni, Jiamin Hou, Hui Kong, Sicheng Lin, Seshadri Nadathur, Jeffrey A. Newman, Eva-Maria Mueller, Will J. Percival, Mehdi Rezaie, Graziano Rossi, Donald P. Schneider, Prabhakar Tiwari, M. Vivek, Yuting Wang, and Gong-Bo Zhao. The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: measurement of the BAO and growth rate of structure of the emission line galaxy sample from the anisotropic power spectrum between redshift 0.6 and 1.1. , 501(4):5616–5645, March 2021. doi: 10.1093/mnras/staa3891.

[83] Héctor Gil-Marín, Julián E. Bautista, Romain Paviot, Mariana Vargas-Magaña, Sylvain de la Torre, Sebastien Fromenteau, Shadab Alam, Santiago Ávila, Etienne Burtin, Chia-Hsun Chuang, Kyle S. Dawson, Jiamin Hou, Arnaud de Mattia, Faizan G. Mohammad, Eva-Maria Müller, Seshadri Nadathur, Richard Neveux, Will J. Percival, Anand Raichoor, Mehdi Rezaie, Ashley J. Ross, Graziano Rossi, Vanina Ruhlmann-Kleider, Alex Smith, Amélie Tamone, Jeremy L. Tinker, Rita Tojeiro, Yuting Wang, Gong-Bo Zhao, Cheng Zhao, Jonathan Brinkmann, Joel R. Brownstein, Peter D. Choi, Stephanie Escoffier, Axel de la Macorra, Jeongin Moon, Jeffrey A. Newman, Donald P. Schneider, Hee-Jong Seo, and Mariappan Vivek. The Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: measurement of the BAO and growth rate of structure of the luminous red galaxy sample from the anisotropic power spectrum between redshifts 0.6 and 1.0. , 498(2):2492–2531, October 2020. doi: 10.1093/mnras/staa2455.

[84] M. Betoule, R. Kessler, J. Guy, J. Mosher, D. Hardin, R. Biswas, P. Astier, P. El-Hage, M. Konig, S. Kuhlmann, J. Marriner, R. Pain, N. Regnault, C. Balland, B. A. Bassett, P. J. Brown, H. Campbell, R. G. Carlberg, F. Cellier-Holzem, D. Cinabro, A. Conley, C. B. D'Andrea, D. L. DePoy, M. Doi, R. S. Ellis, S. Fabbro, A. V. Filippenko, R. J. Foley, J. A. Frieman, D. Fouchez, L. Galbany, A. Goobar, R. R. Gupta, G. J. Hill, R. Hlozek, C. J. Hogan, I. M. Hook, D. A. Howell, S. W. Jha, L. Le Guillou, G. Leloudas, C. Lidman, J. L. Marshall, A. Möller, A. M. Mourão, J. Neveu, R. Nichol, M. D. Olmstead, N. Palanque-Delabrouille, S. Perlmutter, J. L. Prieto, C. J. Pritchet, M. Richmond, A. G. Riess, V. Ruhlmann-Kleider, M. Sako, K. Schahmaneche, D. P. Schneider, M. Smith, J. Sollerman, M. Sullivan, N. A. Walton, and C. J. Wheeler. Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples. , 568:A22, August 2014. doi: 10.1051/0004-6361/201423413.

[85] Éric Aubourg, Stephen Bailey, Julian E. Bautista, Florian Beutler, Vaishali Bhardwaj, Dmitry Bizyaev, Michael Blanton, Michael Blomqvist, Adam S. Bolton, Jo Bovy, Howard Brewington, J. Brinkmann, Joel R. Brownstein, Angela Burden, Nicolás G. Busca, William Carithers, Chia-Hsun Chuang, Johan Comparat, Rupert A. C. Croft, Antonio J. Cuesta, Kyle S. Dawson, Timothée Delubac, Daniel J. Eisenstein, Andreu Font-Ribera, Jian Ge, J. M. Le Goff, Satya Gontcho A. Gontcho, J. Richard Gott, James E. Gunn, Hong Guo, Julien Guy, Jean-Christophe Hamilton, Shirley Ho, Klaus Honscheid, Cullan Howlett, David Kirkby, Francisco S. Kitaura, Jean-Paul Kneib, Khee-Gan Lee, Dan Long, Robert H. Lupton, Mariana Vargas Magaña, Viktor Malanushenko, Elena Malanushenko, Marc Manera, Claudia Maraston, Daniel Margala, Cameron K. McBride, Jordi Miralda-Escudé, Adam D. Myers, Robert C. Nichol, Pasquier Noterdaeme, Sebastián E. Nuza, Matthew D. Olmstead, Daniel Oravetz, Isabelle Pâris, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Kaike Pan, Marcos Pellejero-Ibanez, Will J. Percival, Patrick Petitjean, Matthew M. Pieri, Francisco Prada, Beth Reid, James Rich, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Graziano Rossi, Jose Alberto Rubiño-Martín, Ariel G. Sánchez, Lado Samushia, Ri-

cardo Tanausú Génova-Santos, Claudia G. Scóccola, David J. Schlegel, Donald P. Schneider, Hee-Jong Seo, Erin Sheldon, Audrey Simmons, Ramin A. Skibba, Anže Slosar, Michael A. Strauss, Daniel Thomas, Jeremy L. Tinker, Rita Tojeiro, Jose Alberto Vazquez, Matteo Viel, David A. Wake, Benjamin A. Weaver, David H. Weinberg, W. M. Wood-Vasey, Christophe Yèche, Idit Zehavi, Gong-Bo Zhao, and BOSS Collaboration. Cosmological implications of baryon acoustic oscillation measurements. , 92(12):123516, December 2015. doi: 10.1103/PhysRevD.92.123516.

[86] Shadab Alam, Metin Ata, Stephen Bailey, Florian Beutler, Dmitry Bizyaev, Jonathan A. Blazek, Adam S. Bolton, Joel R. Brownstein, Angela Burden, Chia-Hsun Chuang, Johan Comparat, Antonio J. Cuesta, Kyle S. Dawson, Daniel J. Eisenstein, Stephanie Escoffier, Héctor Gil-Marín, Jan Niklas Grieb, Nick Hand, Shirley Ho, Karen Kinemuchi, David Kirkby, Francisco Kitaura, Elena Malanushenko, Viktor Malanushenko, Claudia Maraston, Cameron K. McBride, Robert C. Nichol, Matthew D. Olmstead, Daniel Oravetz, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Kaike Pan, Marcos Pellejero-Ibanez, Will J. Percival, Patrick Petitjean, Francisco Prada, Adrian M. Price-Whelan, Beth A. Reid, Sergio A. Rodríguez-Torres, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Graziano Rossi, Jose Alberto Rubiño-Martín, Shun Saito, Salvador Salazar-Albornoz, Lado Samushia, Ariel G. Sánchez, Siddharth Satpathy, David J. Schlegel, Donald P. Schneider, Claudia G. Scóccola, Hee-Jong Seo, Erin S. Sheldon, Audrey Simmons, Anže Slosar, Michael A. Strauss, Molly E. C. Swanson, Daniel Thomas, Jeremy L. Tinker, Rita Tojeiro, Mariana Vargas Magaña, Jose Alberto Vazquez, Licia Verde, David A. Wake, Yuting Wang, David H. Weinberg, Martin White, W. Michael Wood-Vasey, Christophe Yèche, Idit Zehavi, Zhongxu Zhai, and Gong-Bo Zhao. The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample. , 470(3):2617–2652, September 2017. doi: 10.1093/mnras/stx721.

[87] Jan Niklas Grieb, Ariel G. Sánchez, Salvador Salazar-Albornoz, Román Scoccimarro, Martín Crocce, Claudio Dalla Vecchia, Francesco Montesano, Héctor Gil-Marín, Ash-

ley J. Ross, Florian Beutler, Sergio Rodríguez-Torres, Chia-Hsun Chuang, Francisco Prada, Francisco-Shu Kitaura, Antonio J. Cuesta, Daniel J. Eisenstein, Will J. Percival, Mariana Vargas-Magaña, Jeremy L. Tinker, Rita Tojeiro, Joel R. Brownstein, Claudia Maraston, Robert C. Nichol, Matthew D. Olmstead, Lado Samushia, Hee-Jong Seo, Alina Streblyanska, and Gong-bo Zhao. The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: Cosmological implications of the Fourier space wedges of the final sample. , 467(2):2085–2112, May 2017. doi: 10.1093/mnras/stw3384.

[88] Ariel G. Sánchez, Román Scoccimarro, Martín Crocce, Jan Niklas Grieb, Salvador Salazar-Albornoz, Claudio Dalla Vecchia, Martha Lippich, Florian Beutler, Joel R. Brownstein, Chia-Hsun Chuang, Daniel J. Eisenstein, Francisco-Shu Kitaura, Matthew D. Olmstead, Will J. Percival, Francisco Prada, Sergio Rodríguez-Torres, Ashley J. Ross, Lado Samushia, Hee-Jong Seo, Jeremy Tinker, Rita Tojeiro, Mariana Vargas-Magaña, Yuting Wang, and Gong-Bo Zhao. The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: Cosmological implications of the configuration-space clustering wedges. , 464(2):1640–1658, January 2017. doi: 10.1093/mnras/stw2443.

[89] T. M. C. Abbott, F. B. Abdalla, A. Alarcon, J. Aleksić, S. Allam, S. Allen, A. Amara, J. Annis, J. Asorey, S. Avila, D. Bacon, E. Balbinot, M. Banerji, N. Banik, W. Barkhouse, M. Baumer, E. Baxter, K. Bechtol, M. R. Becker, A. Benoit-Lévy, B. A. Benson, G. M. Bernstein, E. Bertin, J. Blazek, S. L. Bridle, D. Brooks, D. Brout, E. Buckley-Geer, D. L. Burke, M. T. Busha, A. Campos, D. Capozzi, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, R. Cawthon, C. Chang, N. Chen, M. Childress, A. Choi, C. Conselice, R. Crittenden, M. Crocce, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, R. Das, T. M. Davis, C. Davis, J. De Vicente, D. L. DePoy, J. DeRose, S. Desai, H. T. Diehl, J. P. Dietrich, S. Dodelson, P. Doel, A. Drlica-Wagner, T. F. Eifler, A. E. Elliott, F. Elsner, J. Elvin-Poole, J. Estrada, A. E. Evrard, Y. Fang, E. Fernandez, A. Ferté, D. A. Finley, B. Flaugher, P. Fosalba, O. Friedrich, J. Frie-

man, J. García-Bellido, M. Garcia-Fernandez, M. Gatti, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, M. S. S. Gill, K. Glazebrook, D. A. Goldstein, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, S. Hamilton, W. G. Hartley, S. R. Hinton, K. Honscheid, B. Hoyle, D. Huterer, B. Jain, D. J. James, M. Jarvis, T. Jeltema, M. D. Johnson, M. W. G. Johnson, T. Kacprzak, S. Kent, A. G. Kim, A. King, D. Kirk, N. Kokron, A. Kovacs, E. Krause, C. Krawiec, A. Kremin, K. Kuehn, S. Kuhlmann, N. Kuropatkin, F. Lacasa, O. Lahav, T. S. Li, A. R. Liddle, C. Lidman, M. Lima, H. Lin, N. MacCrann, M. A. G. Maia, M. Makler, M. Manera, M. March, J. L. Marshall, P. Martini, R. G. McMahon, P. Melchior, F. Menanteau, R. Miquel, V. Miranda, D. Mudd, J. Muir, A. Möller, E. Neilsen, R. C. Nichol, B. Nord, P. Nugent, R. L. C. Ogando, A. Palmese, J. Peacock, H. V. Peiris, J. Peoples, W. J. Percival, D. Petravick, A. A. Plazas, A. Porredon, J. Prat, A. Pujol, M. M. Rau, A. Refregier, P. M. Ricker, N. Roe, R. P. Rollins, A. K. Romer, A. Roodman, R. Rosenfeld, A. J. Ross, E. Rozo, E. S. Rykoff, M. Sako, A. I. Salvador, S. Samuroff, C. Sánchez, E. Sanchez, B. Santiago, V. Scarpine, R. Schindler, D. Scolnic, L. F. Secco, S. Serrano, I. Sevilla-Noarbe, E. Sheldon, R. C. Smith, M. Smith, J. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, G. Tarle, D. Thomas, M. A. Troxel, D. L. Tucker, B. E. Tucker, S. A. Uddin, T. N. Varga, P. Vielzeuf, V. Vikram, A. K. Vivas, A. R. Walker, M. Wang, R. H. Wechsler, J. Weller, W. Wester, R. C. Wolf, B. Yanny, F. Yuan, A. Zenteno, B. Zhang, Y. Zhang, J. Zuntz, and Dark Energy Survey Collaboration. Dark Energy Survey year 1 results: Cosmological constraints from galaxy clustering and weak lensing. , 98(4): 043526, August 2018. doi: 10.1103/PhysRevD.98.043526.

[90] G. Alestas, L. Kazantzidis, and L. Perivolaropoulos. H_0 tension, phantom dark energy, and cosmological parameter degeneracies. , 101(12):123516, June 2020. doi: 10.1103/PhysRevD.101.123516.

[91] Guido d'Amico, Jérôme Gleyzes, Nickolas Kokron, Katarina Markovic, Leonardo Senatore, Pierre Zhang, Florian Beutler, and Héctor Gil-Marín. The cosmological analysis

of the SDSS/BOSS data from the Effective Field Theory of Large-Scale Structure. , 2020(5):005, May 2020. doi: 10.1088/1475-7516/2020/05/005.

[92] Eleonora Di Valentino, Alessandro Melchiorri, Olga Mena, and Sunny Vagnozzi. Non-minimal dark sector physics and cosmological tensions. , 101(6):063502, March 2020. doi: 10.1103/PhysRevD.101.063502.

[93] Mikhail M. Ivanov, Marko Simonović, and Matias Zaldarriaga. Cosmological parameters from the BOSS galaxy power spectrum. , 2020(5):042, May 2020. doi: 10.1088/1475-7516/2020/05/042.

[94] 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. , 641:A6, September 2020. doi: 10.1051/0004-6361/201833910.

[95] Tilman Tröster, Ariel. G. Sánchez, Marika Asgari, Chris Blake, Martín Crocce, Catherine Heymans, Hendrik Hildebrandt, Benjamin Joachimi, Shahab Joudaki, Arun Kannawadi, Chieh-An Lin, and Angus Wright. Cosmology from large-scale structure. Constraining ΛCDM with BOSS. , 633:L10, January 2020. doi: 10.1051/0004-6361/201936772.

[96] Luis A. Escamilla, William Giarè, Eleonora Di Valentino, Rafael C. Nunes, and Sunny Vagnozzi. The state of the dark energy equation of state circa 2023. *arXiv e-prints*, art. arXiv:2307.14802, July 2023. doi: 10.48550/arXiv.2307.14802.

[97] Joseph Silk. Cosmic Black-Body Radiation and Galaxy Formation. , 151:459, February 1968. doi: 10.1086/149449.

[98] P. J. E. Peebles and J. T. Yu. Primeval Adiabatic Perturbation in an Expanding Universe. , 162:815, December 1970. doi: 10.1086/150713.

[99] R. A. Sunyaev and Ya. B. Zeldovich. Small-Scale Fluctuations of Relic Radiation. , 7 (1):3–19, April 1970. doi: 10.1007/BF00653471.

[100] J. R. Bond and G. Efstathiou. Cosmic background radiation anisotropies in universes dominated by nonbaryonic dark matter. , 285:L45–L48, October 1984. doi: 10.1086/184362.

[101] J. R. Bond and G. Efstathiou. The statistics of cosmic background radiation fluctuations. , 226:655–687, June 1987. doi: 10.1093/mnras/226.3.655.

[102] Jon A. Holtzman. Microwave Background Anisotropies and Large-Scale Structure in Universes with Cold Dark Matter, Baryons, Radiation, and Massive and Massless Neutrinos. , 71:1, September 1989. doi: 10.1086/191362.

[103] A. Meiksin, Martin White, and J. A. Peacock. Baryonic signatures in large-scale structure. , 304(4):851–864, April 1999. doi: 10.1046/j.1365-8711.1999.02369.x.

[104] Volker Springel, Simon D. M. White, Adrian Jenkins, Carlos S. Frenk, Naoki Yoshida, Liang Gao, Julio Navarro, Robert Thacker, Darren Croton, John Helly, John A. Peacock, Shaun Cole, Peter Thomas, Hugh Couchman, August Evrard, Jörg Colberg, and Frazer Pearce. Simulations of the formation, evolution and clustering of galaxies and quasars. , 435(7042):629–636, June 2005. doi: 10.1038/nature03597.

[105] Martin White. Baryon oscillations. *Astroparticle Physics*, 24(4-5):334–344, December 2005. doi: 10.1016/j.astropartphys.2005.07.007.

[106] Hee-Jong Seo and Daniel J. Eisenstein. Baryonic Acoustic Oscillations in Simulated Galaxy Redshift Surveys. , 633(2):575–588, November 2005. doi: 10.1086/491599.

[107] Andrei Lazanu, Tommaso Giannantonio, Marcel Schmittfull, and E. P. S. Shellard. Matter bispectrum of large-scale structure: Three-dimensional comparison between theoretical models and numerical simulations. , 93(8):083517, April 2016. doi: 10.1103/PhysRevD.93.083517.

[108] Tobias Baldauf, Mathias Garny, Petter Taule, and Theo Steele. Two-loop bispectrum of large-scale structure. , 104(12):123551, December 2021. doi: 10.1103/PhysRevD.104.123551.

[109] Oliver H. E. Philcox, Mikhail M. Ivanov, Giovanni Cabass, Marko Simonović, Matias Zaldarriaga, and Takahiro Nishimichi. Cosmology with the redshift-space galaxy bispectrum monopole at one-loop order. , 106(4):043530, August 2022. doi: 10.1103/PhysRevD.106.043530.

[110] Davit Alkhanishvili, Cristiano Porciani, Emiliano Sefusatti, Matteo Biagetti, Andrei Lazanu, Andrea Oddo, and Victoria Yankelevich. The reach of next-to-leading-order perturbation theory for the matter bispectrum. , 512(4):4961–4981, June 2022. doi: 10.1093/mnras/stac567.

[111] Mandar Karandikar, Cristiano Porciani, and Oliver Hahn. Testing the assumptions of the Effective Field Theory of Large-Scale Structure. *arXiv e-prints*, art. arXiv:2307.11090, July 2023. doi: 10.48550/arXiv.2307.11090.

[112] M. Guidi, A. Veropalumbo, E. Branchini, A. Eggemeier, and C. Carbone. Modelling the next-to-leading order matter three-point correlation function using FFTLog. , 2023 (8):066, August 2023. doi: 10.1088/1475-7516/2023/08/066.

[113] Zachary Slepian and Daniel J. Eisenstein. Computing the three-point correlation function of galaxies in O(N2) time. , 454(4):4142–4158, December 2015. doi: 10.1093/mnras/stv2119.

[114] 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. , 468(1):1070–1083, June 2017. doi: 10.1093/mnras/stw3234.

[115] Joyce Byun, Nishant Agarwal, Rachel Bean, and Richard Holman. Looking for non-Gaussianity in all the right places: A new basis for nonseparable bispectra. , 91(12): 123518, June 2015. doi: 10.1103/PhysRevD.91.123518.

[116] Joyce Byun, Alexander Eggemeier, Donough Regan, David Seery, and Robert E. Smith. Towards optimal cosmological parameter recovery from compressed bispectrum statistics. , 471(2):1581–1618, October 2017. doi: 10.1093/mnras/stx1681.

[117] Héctor Gil-Marín, Julien Guy, Pauline Zarrouk, Etienne Burtin, Chia-Hsun Chuang, Will J. Percival, Ashley J. Ross, Rossana Ruggeri, Rita Tojerio, Gong-Bo Zhao, Yuting Wang, Julian Bautista, Jiamin Hou, Ariel G. Sánchez, Isabelle Pâris, Falk Baumgarten, Joel R. Brownstein, Kyle S. Dawson, Sarah Eftekharzadeh, Violeta González-Pérez, Salman Habib, Katrin Heitmann, Adam D. Myers, Graziano Rossi, Donald P. Schneider, Hee-Jong Seo, Jeremy L. Tinker, and Cheng Zhao. The clustering of the SDSS-IV extended Baryon Oscillation Spectroscopic Survey DR14 quasar sample: structure growth rate measurement from the anisotropic quasar power spectrum in the redshift range $0.8 < z < 2.2$. , 477(2):1604–1638, June 2018. doi: 10.1093/mnras/sty453.

[118] Davide Gualdi, Héctor Gil-Marín, Marc Manera, Benjamin Joachimi, and Ofer Lahav. Geometrical compression: a new method to enhance the BOSS galaxy bispectrum monopole constraints. , 484(1):L29–L34, March 2019. doi: 10.1093/mnrasl/sly242.

[119] Naonori S. Sugiyama, Shun Saito, Florian Beutler, and Hee-Jong Seo. A complete FFT-based decomposition formalism for the redshift-space bispectrum. , 484(1):364–384, March 2019. doi: 10.1093/mnras/sty3249.

[120] Naonori S. Sugiyama, Daisuke Yamauchi, Tsutomu Kobayashi, Tomohiro Fujita, Shun Arai, Shin'ichi Hirano, Shun Saito, Florian Beutler, and Hee-Jong Seo. New constraints on cosmological modified gravity theories from anisotropic three-point correlation functions of BOSS DR12 galaxies. , 523(2):3133–3191, August 2023. doi: 10.1093/mnras/stad1505.

[121] Naonori S. Sugiyama, Daisuke Yamauchi, Tsutomu Kobayashi, Tomohiro Fujita, Shun Arai, Shin'ichi Hirano, Shun Saito, Florian Beutler, and Hee-Jong Seo. First test of the consistency relation for the large-scale structure using the anisotropic three-point

correlation function of BOSS DR12 galaxies. , 524(2):1651–1667, September 2023. doi: 10.1093/mnras/stad1935.

[122] Andrea Oddo, Emiliano Sefusatti, Cristiano Porciani, Pierluigi Monaco, and Ariel G. Sánchez. Toward a robust inference method for the galaxy bispectrum: likelihood function and model selection. , 2020(3):056, March 2020. doi: 10.1088/1475-7516/2020/03/056.

[123] Manuel Colavincenzo, Emiliano Sefusatti, Pierluigi Monaco, Linda Blot, Martin Crocce, Martha Lippich, Ariel G. Sánchez, Marcelo A. Alvarez, Aniket Agrawal, Santiago Avila, Andrés Balaguera-Antolínez, Richard Bond, Sandrine Codis, Claudio Dalla Vecchia, Antonio Dorta, Pablo Fosalba, Albert Izard, Francisco-Shu Kitaura, Marcos Pellejero-Ibanez, George Stein, Mohammadjavad Vakili, and Gustavo Yepes. Comparing approximate methods for mock catalogues and covariance matrices - III: bispectrum. , 482(4):4883–4905, February 2019. doi: 10.1093/mnras/sty2964.

[124] Benjamin Bose and Atsushi Taruya. The one-loop matter bispectrum as a probe of gravity and dark energy. , 2018(10):019, October 2018. doi: 10.1088/1475-7516/2018/10/019.

[125] 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. , 427(4):3435–3467, December 2012. doi: 10.1111/j.1365-2966.2012.22066.x.

[126] Antonio J. Cuesta, Mariana Vargas-Magaña, Florian Beutler, Adam S. Bolton, Joel R. Brownstein, Daniel J. Eisenstein, Héctor Gil-Marín, Shirley Ho, Cameron K. McBride, Claudia Maraston, Nikhil Padmanabhan, Will J. Percival, Beth A. Reid, Ashley J. Ross, Nicholas P. Ross, Ariel G. Sánchez, David J. Schlegel, Donald P. Schneider, Daniel Thomas, Jeremy Tinker, Rita Tojeiro, Licia Verde, and Martin White. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the correlation function of LOWZ and CMASS galaxies in Data Release 12. , 457(2):1770–1785, April 2016. doi: 10.1093/mnras/stw066.

[127] Héctor Gil-Marín, Will J. Percival, Antonio J. Cuesta, Joel R. Brownstein, Chia-Hsun Chuang, Shirley Ho, Francisco-Shu Kitaura, Claudia Maraston, Francisco Prada, Sergio Rodríguez-Torres, Ashley J. Ross, David J. Schlegel, Donald P. Schneider, Daniel Thomas, Jeremy L. Tinker, Rita Tojeiro, Mariana Vargas Magaña, and Gong-Bo Zhao. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: BAO measurement from the LOS-dependent power spectrum of DR12 BOSS galaxies. , 460(4):4210–4219, August 2016. doi: 10.1093/mnras/stw1264.

[128] Yookyung Noh, Martin White, and Nikhil Padmanabhan. Reconstructing baryon oscillations. , 80(12):123501, December 2009. doi: 10.1103/PhysRevD.80.123501.

[129] Lado Samushia, Zachary Slepian, and Francisco Villaescusa-Navarro. Information con-

tent of higher order galaxy correlation functions. , 505(1):628–641, July 2021. doi: 10.1093/mnras/stab1199.

[130] Cé sar Hernández-Aguayo, Francisco Prada, Carlton M Baugh, and Anatoly Klypin. Building a digital twin of a luminous red galaxy spectroscopic survey: galaxy properties and clustering covariance. *Monthly Notices of the Royal Astronomical Society*, 503(2): 2318–2339, feb 2021. doi: 10.1093/mnras/stab434. URL https://doi.org/10.1093%2Fmnras%2Fstab434.

[131] Planck Collaboration, P. A. R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett, E. Battaner, K. Benabed, A. Benoît, A. Benoit-Lévy, J. P. Bernard, M. Bersanelli, P. Bielewicz, J. Bobin, J. J. Bock, A. Bonaldi, J. R. Bond, J. Borrill, F. R. Bouchet, M. Bridges, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, B. Cappellini, J. F. Cardoso, A. Catalano, A. Challinor, A. Chamballu, R. R. Chary, X. Chen, H. C. Chiang, L. Y. Chiang, P. R. Christensen, S. Church, D. L. Clements, S. Colombi, L. P. L. Colombo, F. Couchot, A. Coulais, B. P. Crill, A. Curto, F. Cuttaia, L. Danese, R. D. Davies, R. J. Davis, P. de Bernardis, A. de Rosa, G. de Zotti, J. Delabrouille, J. M. Delouis, F. X. Désert, C. Dickinson, J. M. Diego, K. Dolag, H. Dole, S. Donzelli, O. Doré, M. Douspis, J. Dunkley, X. Dupac, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, F. Finelli, O. Forni, M. Frailis, A. A. Fraisse, E. Franceschi, T. C. Gaier, S. Galeotta, S. Galli, K. Ganga, M. Giard, G. Giardino, Y. Giraud-Héraud, E. Gjerløw, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gregorio, A. Gruppuso, J. E. Gudmundsson, J. Haissinski, J. Hamann, F. K. Hansen, D. Hanson, D. Harrison, S. Henrot-Versillé, C. Hernández-Monteagudo, D. Herranz, S. R. Hildebrandt, E. Hivon, M. Hobson, W. A. Holmes, A. Hornstrup, Z. Hou, W. Hovest, K. M. Huffenberger, A. H. Jaffe, T. R. Jaffe, J. Jewell, W. C. Jones, M. Juvela, E. Keihänen, R. Keskitalo, T. S. Kisner, R. Kneissl, J. Knoche, L. Knox, M. Kunz, H. Kurki-Suonio, G. Lagache, A. Lähteenmäki, J. M. Lamarre, A. Lasenby, M. Lattanzi, R. J. Laureijs, C. R. Lawrence, S. Leach, J. P. Leahy,

R. Leonardi, J. León-Tavares, J. Lesgourgues, A. Lewis, M. Liguori, P. B. Lilje, M. Linden-Vørnle, M. López-Caniego, P. M. Lubin, J. F. Macías-Pérez, B. Maffei, D. Maino, N. Mandolesi, M. Maris, D. J. Marshall, P. G. Martin, E. Martínez-González, S. Masi, M. Massardi, S. Matarrese, F. Matthai, P. Mazzotta, P. R. Meinhold, A. Melchiorri, J. B. Melin, L. Mendes, E. Menegoni, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M. A. Miville-Deschênes, A. Moneti, L. Montier, G. Morgante, D. Mortlock, A. Moss, D. Munshi, J. A. Murphy, P. Naselsky, F. Nati, P. Natoli, C. B. Netterfield, H. U. Nørgaard-Nielsen, F. Noviello, D. Novikov, I. Novikov, I. J. O'Dwyer, S. Osborne, C. A. Oxborrow, F. Paci, L. Pagano, F. Pajot, R. Paladini, D. Paoletti, B. Partridge, F. Pasian, G. Patanchon, D. Pearson, T. J. Pearson, H. V. Peiris, O. Perdereau, L. Perotto, F. Perrotta, V. Pettorino, F. Piacentini, M. Piat, E. Pierpaoli, D. Pietrobon, S. Plaszczynski, P. Platania, E. Pointecouteau, G. Polenta, N. Ponthieu, L. Popa, T. Poutanen, G. W. Pratt, G. Prézeau, S. Prunet, J. L. Puget, J. P. Rachen, W. T. Reach, R. Rebolo, M. Reinecke, M. Remazeilles, C. Renault, S. Ricciardi, T. Riller, I. Ristorcelli, G. Rocha, C. Rosset, G. Roudier, M. Rowan-Robinson, J. A. Rubiño-Martín, B. Rusholme, M. Sandri, D. Santos, M. Savelainen, G. Savini, D. Scott, M. D. Seiffert, E. P. S. Shellard, L. D. Spencer, J. L. Starck, V. Stolyarov, R. Stompor, R. Sudiwala, R. Sunyaev, F. Sureau, D. Sutton, A. S. Suur-Uski, J. F. Sygnet, J. A. Tauber, D. Tavagnacco, L. Terenzi, L. Toffolatti, M. Tomasi, M. Tristram, M. Tucci, J. Tuovinen, M. Türler, G. Umana, L. Valenziano, J. Valiviita, B. Van Tent, P. Vielva, F. Villa, N. Vittorio, L. A. Wade, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Wilkinson, D. Yvon, A. Zacchei, and A. Zonca. Planck 2013 results. XVI. Cosmological parameters. , 571:A16, November 2014. doi: 10.1051/0004-6361/201321591.

[132] Benjamin Wallisch. *Cosmological probes of light relics*. PhD thesis, University of Cambridge, UK, January 2018.

[133] Jan Hamann, Steen Hannestad, Julien Lesgourgues, Cornelius Rampf, and Yvonne

Y. Y. Wong. Cosmological parameters from large scale structure - geometric versus shape information. , 2010(7):022, July 2010. doi: 10.1088/1475-7516/2010/07/022.

[134] Praful Gagrani and Lado Samushia. Information Content of the Angular Multipoles of Redshift-Space Galaxy Bispectrum. , 467(1):928–935, May 2017. doi: 10.1093/mnras/stx135.

[135] Giorgi Khomeriki and Lado Samushia. Mixing Bispectrum Multipoles Under Geometric Distortions. *arXiv e-prints*, art. arXiv:2307.16498, July 2023. doi: 10.48550/arXiv.2307.16498.

[136] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. emcee: The MCMC Hammer. , 125(925):306, March 2013. doi: 10.1086/670067.

[137] J. Hartlap, P. Simon, and P. Schneider. Why your model parameter confidences might be too optimistic. Unbiased estimation of the inverse covariance matrix. , 464(1): 399–404, March 2007. doi: 10.1051/0004-6361:20066170.

[138] E. Sefusatti, M. Crocce, R. Scoccimarro, and H. M. P. Couchman. Accurate estimators of correlation functions in Fourier space. , 460(4):3624–3636, August 2016. doi: 10.1093/mnras/stw1229.

[139] Nezih C. Geçkinli and Davras Yavuz. *Discrete Fourier transformation and its applications to power spectra estimation*. 1983.

[140] Robert W. Ramirez. *The FFT. Fundamentals and concepts*. 1985.

[141] E. Oran Brigham. *The fast Fourier transform and its applications.* 1988.

[142] S. D. M. White and M. J. Rees. Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering. , 183:341–358, May 1978. doi: 10.1093/mnras/183.3.341.

[143] 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. doi: 10.1186/s40668-019-0028-x.

[144] Volker Springel, Rüdiger Pakmor, Annalisa Pillepich, Rainer Weinberger, Dylan Nelson, Lars Hernquist, Mark Vogelsberger, Shy Genel, Paul Torrey, Federico Marinacci, and Jill Naiman. First results from the IllustrisTNG simulations: matter and galaxy clustering. , 475(1):676–698, March 2018. doi: 10.1093/mnras/stx3304.

[145] Annalisa Pillepich, Volker Springel, Dylan Nelson, Shy Genel, Jill Naiman, Rüdiger Pakmor, Lars Hernquist, Paul Torrey, Mark Vogelsberger, Rainer Weinberger, and Federico Marinacci. Simulating galaxy formation with the IllustrisTNG model. , 473 (3):4077–4106, January 2018. doi: 10.1093/mnras/stx2656.

[146] Federico Marinacci, Mark Vogelsberger, Rüdiger Pakmor, Paul Torrey, Volker Springel, Lars Hernquist, Dylan Nelson, Rainer Weinberger, Annalisa Pillepich, Jill Naiman, and Shy Genel. First results from the IllustrisTNG simulations: radio haloes and magnetic fields. , 480(4):5113–5139, November 2018. doi: 10.1093/mnras/sty2206.

[147] Jill P. Naiman, Annalisa Pillepich, Volker Springel, Enrico Ramirez-Ruiz, Paul Torrey, Mark Vogelsberger, Rüdiger Pakmor, Dylan Nelson, Federico Marinacci, Lars Hernquist, Rainer Weinberger, and Shy Genel. First results from the IllustrisTNG simulations: a tale of two elements - chemical evolution of magnesium and europium. , 477(1):1206–1224, June 2018. doi: 10.1093/mnras/sty618.

[148] 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. , 506(3): 4210–4231, September 2021. doi: 10.1093/mnras/stab1755.

[149] 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. , 446(1):521–554, January 2015. doi: 10.1093/mnras/stu2058.

[150] 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. , 450(2):1937–1961, June 2015. doi: 10.1093/mnras/stv725.

[151] Uroš Seljak. Analytic model for galaxy and dark matter clustering. , 318(1):203–213, October 2000. doi: 10.1046/j.1365-8711.2000.03715.x.

[152] J. A. Peacock and R. E. Smith. Halo occupation numbers and galaxy bias. , 318(4): 1144–1156, November 2000. doi: 10.1046/j.1365-8711.2000.03779.x.

[153] A. J. Benson, C. S. Frenk, C. M. Baugh, S. Cole, and C. G. Lacey. The clustering evolution of the galaxy distribution. , 327(4):1041–1056, November 2001. doi: 10.1046/j.1365-8711.2001.04824.x.

[154] Andreas A. Berlind and David H. Weinberg. The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass. , 575 (2):587–616, August 2002. doi: 10.1086/341469.

[155] Asantha Cooray and Ravi Sheth. Halo models of large scale structure. , 372(1):1–129, December 2002. doi: 10.1016/S0370-1573(02)00276-4.

[156] Zheng Zheng, Andreas A. Berlind, David H. Weinberg, Andrew J. Benson, Carlton M. Baugh, Shaun Cole, Romeel Davé, Carlos S. Frenk, Neal Katz, and Cedric G. Lacey.

Theoretical Models of the Halo Occupation Distribution: Separating Central and Satellite Galaxies. , 633(2):791–809, November 2005. doi: 10.1086/466510.

[157] Zheng Zheng, Alison L. Coil, and Idit Zehavi. Galaxy Evolution from Halo Occupation Distribution Modeling of DEEP2 and SDSS Galaxy Clustering. , 667(2):760–779, October 2007. doi: 10.1086/521074.

[158] Zheng Zheng, Idit Zehavi, Daniel J. Eisenstein, David H. Weinberg, and Y. P. Jing. Halo Occupation Distribution Modeling of Clustering of Luminous Red Galaxies. , 707 (1):554–572, December 2009. doi: 10.1088/0004-637X/707/1/554.

[159] 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. , 609(1):35–49, July 2004. doi: 10.1086/420959.

[160] Argyro Tasitsiomi, Andrey V. Kravtsov, Risa H. Wechsler, and Joel R. Primack. Modeling Galaxy-Mass Correlations in Dissipationless Simulations. , 614(2):533–546, October 2004. doi: 10.1086/423784.

[161] Charlie Conroy, Risa H. Wechsler, and Andrey V. Kravtsov. Modeling Luminosity-dependent Galaxy Clustering through Cosmic Time. , 647(1):201–214, August 2006. doi: 10.1086/503602.

[162] Peter S. Behroozi, Charlie Conroy, and Risa H. Wechsler. A Comprehensive Analysis of Uncertainties Affecting the Stellar Mass-Halo Mass Relation for $0 < z < 4$. , 717 (1):379–403, July 2010. doi: 10.1088/0004-637X/717/1/379.

[163] Simon D. M. White and Carlos S. Frenk. Galaxy Formation through Hierarchical Clustering. , 379:52, September 1991. doi: 10.1086/170483.

[164] G. Kauffmann, S. D. M. White, and B. Guiderdoni. The formation and evolution of galaxies within merging dark matter haloes. , 264:201–218, September 1993. doi: 10.1093/mnras/264.1.201.

[165] Rachel S. Somerville and Joel R. Primack. Semi-analytic modelling of galaxy formation: the local Universe. , 310(4):1087–1110, December 1999. doi: 10.1046/j.1365-8711.1999.03032.x.

[166] Shaun Cole, Cedric G. Lacey, Carlton M. Baugh, and Carlos S. Frenk. Hierarchical galaxy formation. , 319(1):168–204, November 2000. doi: 10.1046/j.1365-8711.2000.03879.x.

[167] R. G. Bower, A. J. Benson, R. Malbon, J. C. Helly, C. S. Frenk, C. M. Baugh, S. Cole, and C. G. Lacey. Breaking the hierarchy of galaxy formation. , 370(2):645–655, August 2006. doi: 10.1111/j.1365-2966.2006.10519.x.

[168] Quan Guo, Shaun Cole, Vincent Eke, Carlos Frenk, and John Helly. Spatial and luminosity distributions of galactic satellites. , 434(3):1838–1848, September 2013. doi: 10.1093/mnras/stt903.

[169] Mark Vogelsberger, Shy Genel, Volker Springel, Paul Torrey, Debora Sijacki, Dandan Xu, Greg Snyder, Dylan Nelson, and Lars Hernquist. Introducing the Illustris Project: simulating the coevolution of dark and visible matter in the Universe. , 444(2):1518–1547, October 2014. doi: 10.1093/mnras/stu1536.

[170] Romeel Davé, Daniel Anglés-Alcázar, Desika Narayanan, Qi Li, Mika H. Rafieferantsoa, and Sarah Appleby. SIMBA: Cosmological simulations with black hole growth and feedback. , 486(2):2827–2849, June 2019. doi: 10.1093/mnras/stz937.

[171] Martin White, Lars Hernquist, and Volker Springel. The Halo Model and Numerical Simulations. , 550(2):L129–L132, April 2001. doi: 10.1086/319644.

[172] F. R. Pearce, A. Jenkins, C. S. Frenk, S. D. M. White, P. A. Thomas, H. M. P. Couchman, J. A. Peacock, and G. Efstathiou. Simulations of galaxy formation in a cosmological volume. , 326(2):649–666, September 2001. doi: 10.1046/j.1365-8711.2001.04616.x.

[173] Andreas A. Berlind, David H. Weinberg, Andrew J. Benson, Carlton M. Baugh, Shaun Cole, Romeel Davé, Carlos S. Frenk, Adrian Jenkins, Neal Katz, and Cedric G. Lacey. The Halo Occupation Distribution and the Physics of Galaxy Formation. , 593(1): 1–25, August 2003. doi: 10.1086/376517.

[174] Sownak Bose, Daniel J. Eisenstein, Lars Hernquist, Annalisa Pillepich, Dylan Nelson, Federico Marinacci, Volker Springel, and Mark Vogelsberger. Revealing the galaxy-halo connection in IllustrisTNG. , 490(4):5693–5711, December 2019. doi: 10.1093/mnras/stz2546.

[175] Boryana Hadzhiyska, Sownak Bose, Daniel Eisenstein, Lars Hernquist, and David N. Spergel. Limitations to the 'basic' HOD model and beyond. , 493(4):5506–5519, April 2020. doi: 10.1093/mnras/staa623.

[176] Boryana Hadzhiyska, Sownak Bose, Daniel Eisenstein, and Lars Hernquist. Extensions to models of the galaxy-halo connection. , 501(2):1603–1620, February 2021. doi: 10.1093/mnras/staa3776.

[177] Boryana Hadzhiyska, Sandro Tacchella, Sownak Bose, and Daniel J. Eisenstein. The galaxy-halo connection of emission-line galaxies in IllustrisTNG. , 502(3):3599–3617, April 2021. doi: 10.1093/mnras/stab243.

[178] A. J. Benson, F. R. Pearce, C. S. Frenk, C. M. Baugh, and A. Jenkins. A comparison of semi-analytic and smoothed particle hydrodynamics galaxy formation. , 320(2): 261–280, January 2001. doi: 10.1046/j.1365-8711.2001.03966.x.

[179] Vimal Simha, David H. Weinberg, Romeel Davé, Mark Fardal, Neal Katz, and Benjamin D. Oppenheimer. Testing subhalo abundance matching in cosmological smoothed particle hydrodynamics simulations. , 423(4):3458–3473, July 2012. doi: 10.1111/j.1365-2966.2012.21142.x.

[180] Jonás Chaves-Montero, Raul E. Angulo, Joop Schaye, Matthieu Schaller, Robert A. Crain, Michelle Furlong, and Tom Theuns. Subhalo abundance matching and assembly

bias in the EAGLE simulation. , 460(3):3100–3118, August 2016. doi: 10.1093/mnras/stw1225.

[181] Harshil M. Kamdar, Matthew J. Turk, and Robert J. Brunner. Machine learning and cosmological simulations - II. Hydrodynamical simulations. , 457(2):1162–1179, April 2016. doi: 10.1093/mnras/stv2981.

[182] Shankar Agarwal, Romeel Davé, and Bruce A. Bassett. Painting galaxies into dark matter haloes using machine learning. , 478(3):3410–3422, August 2018. doi: 10.1093/mnras/sty1169.

[183] Xiaoju Xu, Saurabh Kumar, Idit Zehavi, and Sergio Contreras. Predicting halo occupation and galaxy assembly bias with machine learning. , 507(4):4879–4899, November 2021. doi: 10.1093/mnras/stab2464.

[184] Ana Maria Delgado, Digvijay Wadekar, Boryana Hadzhiyska, Sownak Bose, Lars Hernquist, and Shirley Ho. Modelling the galaxy-halo connection with machine learning. , 515(2):2733–2746, September 2022. doi: 10.1093/mnras/stac1951.

[185] A. Jana and L. Samushia. Constraining Galaxy-Halo connection using machine learning. *Astronomy and Computing*, 49:100883, October 2024. doi: 10.1016/j.ascom.2024.100883.

[186] Charlie Conroy, James E. Gunn, and Martin White. The Propagation of Uncertainties in Stellar Population Synthesis Modeling. I. The Relevance of Uncertain Aspects of Stellar Evolution and the Initial Mass Function to the Derived Physical Properties of Galaxies. , 699(1):486–506, July 2009. doi: 10.1088/0004-637X/699/1/486.

[187] Kartheik G. Iyer, Camilla Pacifici, Gabriela Calistro-Rivera, and Christopher C. Lovell. The Spectral Energy Distributions of Galaxies. *arXiv e-prints*, art. arXiv:2502.17680, February 2025. doi: 10.48550/arXiv.2502.17680.

[188] Jayashree Behera, Rita Tojeiro, and Harry George Chittenden. Optimised neural

network predictions of galaxy formation histories using semi-stochastic corrections. , June 2025. doi: 10.1093/mnras/staf920.

[189] Alex Alarcon, Andrew P. Hearin, Matthew R. Becker, and Jonás Chaves-Montero. Diffstar: a fully parametric physical model for galaxy assembly history. , 518(1): 562–584, January 2023. doi: 10.1093/mnras/stac3118.

[190] Weiguang Cui, Romeel Davé, John A. Peacock, Daniel Anglés-Alcázar, and Xiaohu Yang. The origin of galaxy colour bimodality in the scatter of the stellar-to-halo mass relation. *Nature Astronomy*, 5:1069–1076, October 2021. doi: 10.1038/s41550-021-01404-1.

[191] Antonio D. Montero-Dorta, Jonás Chaves-Montero, M. Celeste Artale, and Ginevra Favole. On the influence of halo mass accretion history on galaxy properties and assembly bias. , 508(1):940–949, November 2021. doi: 10.1093/mnras/stab2556.

[192] Jonathan J. Davies, Robert A. Crain, and Andrew Pontzen. Quenching and morphological evolution due to circumgalactic gas expulsion in a simulated galaxy with a controlled assembly history. , 501(1):236–253, January 2021. doi: 10.1093/mnras/staa3643.

[193] Yongseok Jo and Ji-hoon Kim. Machine-assisted semi-simulation model (MSSM): estimating galactic baryonic properties from their dark matter using a machine trained on hydrodynamic simulations. , 489(3):3565–3581, November 2019. doi: 10.1093/mnras/stz2304.

[194] Jacky H. T. Yip, Xinyue Zhang, Yanfang Wang, Wei Zhang, Yueqiu Sun, Gabriella Contardo, Francisco Villaescusa-Navarro, Siyu He, Shy Genel, and Shirley Ho. From Dark Matter to Galaxies with Convolutional Neural Networks. *arXiv e-prints*, art. arXiv:1910.07813, October 2019. doi: 10.48550/arXiv.1910.07813.

[195] Digvijay Wadekar, Francisco Villaescusa-Navarro, Shirley Ho, and Laurence Perreault-Levasseur. Modeling assembly bias with machine learning and symbolic regression.

arXiv e-prints, art. arXiv:2012.00111, November 2020. doi: 10.48550/arXiv.2012.00111.

[196] Dylan Nelson, Annalisa Pillepich, Volker Springel, Rainer Weinberger, Lars Hernquist, Rüdiger Pakmor, Shy Genel, Paul Torrey, Mark Vogelsberger, Guinevere Kauffmann, Federico Marinacci, and Jill Naiman. First results from the IllustrisTNG simulations: the galaxy colour bimodality. , 475(1):624–647, March 2018. doi: 10.1093/mnras/stx3040.

[197] Harry George Chittenden, Jayashree Behera, and Rita Tojeiro. Evaluating the galaxy formation histories predicted by a neural network in pure dark matter simulations. *arXiv e-prints*, art. arXiv:2409.16079, September 2024. doi: 10.48550/arXiv.2409.16079.

[198] Martin Sparre, Christopher C. Hayward, Volker Springel, Mark Vogelsberger, Shy Genel, Paul Torrey, Dylan Nelson, Debora Sijacki, and Lars Hernquist. The star formation main sequence and stellar mass assembly of galaxies in the Illustris simulation. , 447(4):3548–3563, March 2015. doi: 10.1093/mnras/stu2713.

[199] Kartheik G. Iyer, Sandro Tacchella, Shy Genel, Christopher C. Hayward, Lars Hernquist, Alyson M. Brooks, Neven Caplar, Romeel Davé, Benedikt Diemer, John C. Forbes, Eric Gawiser, Rachel S. Somerville, and Tjitske K. Starkenburg. The diversity and variability of star formation histories in models of galaxy evolution. , 498(1):430–463, October 2020. doi: 10.1093/mnras/staa2150.

[200] Neven Caplar and Sandro Tacchella. Stochastic modelling of star-formation histories I: the scatter of the star-forming main sequence. , 487(3):3845–3869, August 2019. doi: 10.1093/mnras/stz1449.

[201] Sandro Tacchella, John C. Forbes, and Neven Caplar. Stochastic modelling of star-formation histories II: star-formation variability from molecular clouds and gas inflow. , 497(1):698–725, September 2020. doi: 10.1093/mnras/staa1838.

[202] Kartheik G. Iyer, Joshua S. Speagle, Neven Caplar, John C. Forbes, Eric Gawiser, Joel Leja, and Sandro Tacchella. Stochastic Modeling of Star Formation Histories. III. Constraints from Physically Motivated Gaussian Processes. , 961(1):53, January 2024. doi: 10.3847/1538-4357/acff64.

[203] Simon J. Lilly, C. Marcella Carollo, Antonio Pipino, Alvio Renzini, and Yingjie Peng. Gas Regulation of Galaxies: The Evolution of the Cosmic Specific Star Formation Rate, the Metallicity-Mass-Star-formation Rate Relation, and the Stellar Content of Halos. , 772(2):119, August 2013. doi: 10.1088/0004-637X/772/2/119.

[204] Simon Birrer, Simon Lilly, Adam Amara, Aseem Paranjape, and Alexandre Refregier. A Simple Model Linking Galaxy and Dark Matter Evolution. , 793(1):12, September 2014. doi: 10.1088/0004-637X/793/1/12.

[205] Charlie Conroy and James E. Gunn. The Propagation of Uncertainties in Stellar Population Synthesis Modeling. III. Model Calibration, Comparison, and Evaluation. , 712(2):833–857, April 2010. doi: 10.1088/0004-637X/712/2/833.

[206] Charlie Conroy and James E. Gunn. FSPS: Flexible Stellar Population Synthesis. Astrophysics Source Code Library, record ascl:1010.043, October 2010.

[207] Sorour Shamshiri, Peter A. Thomas, Bruno M. Henriques, Rita Tojeiro, Gerard Lemson, Seb J. Oliver, and Stephen Wilkins. Galaxy formation in the Planck cosmology - II. Star-formation histories and post-processing magnitude reconstruction. , 451(3): 2681–2691, August 2015. doi: 10.1093/mnras/stv883.

[208] Bruno M. B. Henriques, Simon D. M. White, Peter A. Thomas, Raul Angulo, Qi Guo, Gerard Lemson, Volker Springel, and Roderik Overzier. Galaxy formation in the Planck cosmology - I. Matching the observed evolution of star formation rates, colours and stellar masses. , 451(3):2663–2680, August 2015. doi: 10.1093/mnras/stv705.

[209] Harry George Chittenden and Rita Tojeiro. Modelling the galaxy-halo connection with

semi-recurrent neural networks. *MNRAS*, 518(4):5670, Feb 2023. doi: 10.1093/mnras/stac3498. URL https://dx.doi.org/10.1093/mnras/stac3498.

[210] Rachel S. Somerville and Romeel Davé. Physical models of galaxy formation in a cosmological framework. *ARA&A*, 53(1):51, Aug 2015. ISSN 1545-4282. doi: 10.1146/annurev-astro-082812-140951. URL http://dx.doi.org/10.1146/annurev-astro-082812-140951.

[211] Mark Vogelsberger, Federico Marinacci, Paul Torrey, and Ewald Puchwein. Cosmological simulations of galaxy formation. *Nat. Rev. Phys.*, 2(1):42, January 2020. doi: 10.1038/s42254-019-0127-2.

[212] Risa H. Wechsler and Jeremy L. Tinker. The connection between galaxies and their dark matter halos. *ARA&A*, 56(1):435, Sep 2018. ISSN 1545-4282. doi: 10.1146/annurev-astro-081817-051756. URL http://dx.doi.org/10.1146/annurev-astro-081817-051756.

[213] Dylan Nelson, Annalisa Pillepich, Volker Springel, Rainer Weinberger, Lars Hernquist, Rüdiger Pakmor, Shy Genel, Paul Torrey, Mark Vogelsberger, Guinevere Kauffmann, Federico Marinacci, and Jill Naiman. First results from the illustristng simulations: the galaxy colour bimodality. *MNRAS*, 475(1):624, Nov 2017. ISSN 1365-2966. doi: 10.1093/mnras/stx3040. URL http://dx.doi.org/10.1093/mnras/stx3040.

[214] 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. *CompAC*, 6:2, 05 2019. doi: 10.1186/s40668-019-0028-x.

[215] Annalisa Pillepich, Volker Springel, Dylan Nelson, Shy Genel, Jill Naiman, Rüdiger Pakmor, Lars Hernquist, Paul Torrey, Mark Vogelsberger, Rainer Weinberger, and Federico Marinacci. Simulating galaxy formation with the illustristng model. *MNRAS*,

473(3):4077, Oct 2017. ISSN 1365-2966. doi: 10.1093/mnras/stx2656. URL http://dx.doi.org/10.1093/mnras/stx2656.

[216] Volker Springel, Rüdiger Pakmor, Annalisa Pillepich, Rainer Weinberger, Dylan Nelson, Lars Hernquist, Mark Vogelsberger, Shy Genel, Paul Torrey, Federico Marinacci, and Jill Naiman. First results from the IllustrisTNG simulations: matter and galaxy clustering. *MNRAS*, 475(1):676, Dec 2017. doi: 10.1093/mnras/stx3304. URL https://doi.org/10.1093%2Fmnras%2Fstx3304.

[217] Federico Marinacci, Mark Vogelsberger, Rüdiger Pakmor, Paul Torrey, Volker Springel, Lars Hernquist, Dylan Nelson, Rainer Weinberger, Annalisa Pillepich, Jill Naiman, and Shy Genel. First results from the illustristng simulations: radio haloes and magnetic fields. *MNRAS*, 480(4):5113, 08 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty2206. URL https://doi.org/10.1093/mnras/sty2206.

[218] Jill P Naiman, Annalisa Pillepich, Volker Springel, Enrico Ramirez-Ruiz, Paul Torrey, Mark Vogelsberger, Rüdiger Pakmor, Dylan Nelson, Federico Marinacci, Lars Hernquist, Rainer Weinberger, and Shy Genel. First results from the IllustrisTNG simulations: a tale of two elements – chemical evolution of magnesium and europium. *MNRAS*, 477(1):1206, Mar 2018. doi: 10.1093/mnras/sty618. URL https://doi.org/10.1093%2Fmnras%2Fsty618.

[219] Mélanie Habouzit, Masafusa Onoue, Eduardo Bañados, Marcel Neeleman, Daniel Anglés-Alcázar, Fabian Walter, Annalisa Pillepich, Romeel Davé, Knud Jahnke, and Yohan Dubois. Co-Evolution of Massive Black Holes and their Host Galaxies at High Redshift: Discrepancies From Six Cosmological Simulations and the Key Role of JWST. *MNRAS*, 511(3):3751, Jan 2022. doi: 10.1093/mnras/stac225. URL https://doi.org/10.1093%2Fmnras%2Fstac225.

[220] Masayuki Tanaka, Masato Onodera, Rhythm Shimakawa, Kei Ito, Takumi Kakimoto, Mariko Kubo, Takahiro Morishita, Sune Toft, Francesco Valentino, and Po-Feng Wu. A protocluster of massive quiescent galaxies at z = 4. *ApJ*, 970(1):59, Jul 2023. doi:

10.3847/1538-4357/ad5316. URL https://ui.adsabs.harvard.edu/abs/2024ApJ...970...59T/abstract.

[221] Shalini Kurinchi-Vendhan, Marion Farcy, Michaela Hirschmann, and Francesco Valentino. A two-phase model of galaxy formation: II. The size-mass relation of dynamically hot galaxies. , Jun 2024. doi: 10.48550/arXiv.2310.03083. URL https://ui.adsabs.harvard.edu/abs/2023arXiv231003083K/abstract.

[222] Tiago Castro, Stefano Borgani, Klaus Dolag, Valerio Marra, Miguel Quartin, Alexandro Saro, and Emiliano Sefusatti. On the impact of baryons on the halo mass function, bias, and cluster cosmology. *MNRAS*, 500(2):2316, Nov 2020. doi: 10.1093/mnras/staa3473. URL https://doi.org/10.1093%2Fmnras%2Fstaa3473.

[223] Dhayaa Anbajagane, August E Evrard, and Arya Farahi. Baryonic imprints on DM haloes: population statistics from dwarf galaxies to galaxy clusters. *MNRAS*, 509(3):3441, Nov 2021. doi: 10.1093/mnras/stab3177. URL https://doi.org/10.1093%2Fmnras%2Fstab3177.

[224] Roan Haggar, Frazer R Pearce, Meghan E Gray, Alexander Knebe, and Gustavo Yepes. The three hundred project: Substructure in hydrodynamical and dark matter simulations of galaxy groups around clusters. *MNRAS*, 502(1):1191, Jan 2021. doi: 10.1093/mnras/stab064. URL https://doi.org/10.1093%2Fmnras%2Fstab064.

[225] Stephen D Riggs, Jon Loveday, Peter A Thomas, Annalisa Pillepich, Dylan Nelson, and Benne W Holwerda. Exploring the effect of baryons on the radial distribution of satellite galaxies with GAMA and IllustrisTNG. *MNRAS*, 514(4):4676, Jun 2022. doi: 10.1093/mnras/stac1591. URL https://doi.org/10.1093%2Fmnras%2Fstac1591.

[226] 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, Jun 2021. doi: 10.1093/mnras/stab1755. URL https://doi.org/10.1093%2Fmnras%2Fstab1755.

[227] Planck Collaboration. Planck 2015 Results. XIII. Cosmological Parameters. *A&A*, 594:A13, Sep 2016. doi: 10.1051/0004-6361/201525830. URL https://doi.org/10.1051%2F0004-6361%2F201525830.

[228] Martin Chalela, Emanuel Sillero, Luis Pereyra, Mario Alejandro García, Juan B. Cabral, Marcelo Lares, and Manuel Merchán. GriSPy: A Python Package for Fixed-Radius Nearest Neighbors Search. *Astronomy & Computing*, 34:100443, December 2021. doi: https://doi.org/10.1016/j.ascom.2020.100443.

[229] Jingjing Shi, Huiyuan Wang, Houjun Mo, Mark Vogelsberger, Luis C. Ho, Min Du, Dylan Nelson, Annalisa Pillepich, and Lars Hernquist. The formation history of subhalos and the evolution of satellite galaxies. *ApJ*, 893(2):139, Apr 2020. ISSN 1538-4357. doi: 10.3847/1538-4357/ab8464. URL http://dx.doi.org/10.3847/1538-4357/ab8464.

[230] T. Sousbie. The persistent cosmic web and its filamentary structure - i. theory and implementation. *MNRAS*, 414(1):350, Apr 2011. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2011.18394.x. URL http://dx.doi.org/10.1111/j.1365-2966.2011.18394.x.

[231] Antonio D Montero-Dorta, Jonás Chaves-Montero, M Celeste Artale, and Ginevra Favole. On the influence of halo mass accretion history on galaxy properties and assembly bias. *MNRAS*, 508(1):940, Sep 2021. ISSN 1365-2966. doi: 10.1093/mnras/stab2556. URL http://dx.doi.org/10.1093/mnras/stab2556.

[232] Charlie Conroy, James Gunn, and Martin White. The Propagation of Uncertainties in Stellar Population Synthesis Modeling. I. The Relevance of Uncertain Aspects of Stellar Evolution and the Initial Mass Function to the Derived Physical Properties of Galaxies. *ApJ*, 699:486, 07 2009. doi: 10.1088/0004-637X/699/1/486.

[233] 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, Mar 2010. ISSN 1538-4357. doi: 10.1088/0004-637x/712/2/833. URL http://dx.doi.org/10.1088/0004-637X/712/2/833.

[234] D. Nelson, A. Pillepich, S. Genel, M. Vogelsberger, V. Springel, P. Torrey, V. Rodriguez-Gomez, D. Sijacki, G.F. Snyder, B. Griffen, F. Marinacci, L. Blecha, L. Sales, D. Xu, and L. Hernquist. The illustris simulation: Public data release. *Astronomy and Computing*, 13:12, Nov 2015. doi: 10.1016/j.ascom.2015.09.003. URL https://doi.org/10.1016%2Fj.ascom.2015.09.003.

[235] Vicente Rodriguez-Gomez, Shy Genel, Mark Vogelsberger, Debora Sijacki, Annalisa Pillepich, Laura V. Sales, Paul Torrey, Greg Snyder, Dylan Nelson, Volker Springel, Chung-Pei Ma, and Lars Hernquist. The merger rate of galaxies in the illustris simulation: a comparison with observations and semi-empirical models. *MNRAS*, 449(1):49, Mar 2015. ISSN 0035-8711. doi: 10.1093/mnras/stv264. URL http://dx.doi.org/10.1093/mnras/stv264.

[236] Peter S. Behroozi, Risa H. Wechsler, Hao-Yi Wu, Michael T. Busha, Anatoly A. Klypin, and Joel R. Primack. Gravitationally Consistent Halo Catalogs and Merger Trees for Precision Cosmology. *ApJ*, 763(1):18, Dec 2012. doi: 10.1088/0004-637x/763/1/18. URL https://doi.org/10.1088%2F0004-637x%2F763%2F1%2F18.

[237] 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, dec 2012. doi: 10.1088/0004-637x/762/2/109. URL https://doi.org/10.1088%2F0004-637x%2F762%2F2%2F109.

[238] Annalisa Pillepich, Dylan Nelson, Lars Hernquist, Volker Springel, Rüdiger Pakmor, Paul Torrey, Rainer Weinberger, Shy Genel, Jill P Naiman, Federico Marinacci, and Mark Vogelsberger. First results from the illustristng simulations: the stellar mass content of groups and clusters of galaxies. *MNRAS*, 475(1):648, Dec 2017. ISSN

1365-2966. doi: 10.1093/mnras/stx3112. URL http://dx.doi.org/10.1093/mnras/stx3112.

[239] Kun Ting Eddie Chua, Annalisa Pillepich, Mark Vogelsberger, and Lars Hernquist. Shape of dark matter haloes in the illustris simulation: effects of baryons. *MNRAS*, 484(1):476, Jan 2019. doi: 10.1093/mnras/sty3531. URL https://doi.org/10.1093%2Fmnras%2Fsty3531.

[240] Kun Ting Eddie Chua, Mark Vogelsberger, Annalisa Pillepich, and Lars Hernquist. The impact of galactic feedback on the shapes of dark matter haloes. *MNRAS*, 515 (2):2681, Jul 2022. doi: 10.1093/mnras/stac1897. URL https://doi.org/10.1093%2Fmnras%2Fstac1897.

[241] Volker Springel, Simon D. M. White, Giuseppe Tormen, and Guinevere Kauffmann. Populating a cluster of galaxies - i. results at z=0. , 328(3):726, Dec 2001. doi: 10.1046/j.1365-8711.2001.04912.x. URL https://doi.org/10.1046%2Fj.1365-8711.2001.04912.x.

[242] Julio F. Navarro, Carlos S. Frenk, and Simon D. M. White. The structure of cold dark matter halos. *ApJ*, 462:563, May 1996. doi: 10.1086/177173. URL https://doi.org/10.1086%2F177173.

[243] D. H. Zhao, Y. P. Jing, H. J. Mo, and G. Börner. Accurate universal models for the mass accretion histories and concentrations of dark matter halos. *ApJ*, 707(1):354, Nov 2009. doi: 10.1088/0004-637x/707/1/354. URL https://doi.org/10.1088%2F0004-637x%2F707%2F1%2F354.

[244] Francisco Prada, Anatoly A. Klypin, Antonio J. Cuesta, Juan E. Betancort-Rijo, and Joel Primack. Halo concentrations in the standard lcdm cosmology. *MNRAS*, 423(4):3018, Jun 2012. doi: 10.1111/j.1365-2966.2012.21007.x. URL https://doi.org/10.1111%2Fj.1365-2966.2012.21007.x.

[245] Philip Mansfield and Camille Avestruz. How biased are halo properties in cosmological simulations? *MNRAS*, 500(3):3309, Jan 2021. ISSN 1365-2966. doi: 10.1093/mnras/staa3388. URL http://dx.doi.org/10.1093/mnras/staa3388.

[246] Kai Hoffmann, Susana Planelles, Enrique Gaztañaga, Alexander Knebe, Frazer R. Pearce, Hanni Lux, Julian Onions, Stuart I. Muldrew, Pascal Elahi, Peter Behroozi, Yago Ascasibar, Jiaxin Han, Michal Maciejewski, Manuel E. Merchan, Mark Neyrinck, Andrés N. Ruiz, and Mario A. Sgro. Subhaloes gone notts: subhaloes as tracers of the dark matter halo shape. , 442(2):1197, Jun 2014. doi: 10.1093/mnras/stu933. URL https://doi.org/10.1093%2Fmnras%2Fstu933.

[247] Julian Onions, Alexander Knebe, Frazer R. Pearce, Stuart I. Muldrew, Hanni Lux, Steffen R. Knollmann, Yago Ascasibar, Peter Behroozi, Pascal Elahi, Jiaxin Han, Michal Maciejewski, Manuel E. Merchán, Mark Neyrinck, Andrés N. Ruiz, Mario A. Sgró, Volker Springel, and Dylan Tweed. Subhaloes going notts: the subhalo-finder comparison project. *MNRAS*, 423(2):1200, Apr 2012. doi: 10.1111/j.1365-2966.2012.20947.x. URL https://doi.org/10.1111%2Fj.1365-2966.2012.20947.x.

[248] Jonathan J Davies, Robert A Crain, Benjamin D Oppenheimer, and Joop Schaye. The quenching and morphological evolution of central galaxies is facilitated by the feedback-driven expulsion of circumgalactic gas. *MNRAS*, 491(3):4462, Nov 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz3201. URL http://dx.doi.org/10.1093/mnras/stz3201.

[249] Daniele Sorini, Romeel Davé, Weiguang Cui, and Sarah Appleby. How baryons affect haloes and large-scale structure: a unified picture from the SIMBA simulation. , 516 (1):883–906, October 2022. doi: 10.1093/mnras/stac2214.

[250] Benjamin D. Oppenheimer, Romeel Davé, Dušan Kereš, Mark Fardal, Neal Katz, Juna A. Kollmeier, and David H. Weinberg. Feedback and recycled wind accretion: assembling the z= 0 galaxy mass function. , 406(4):2325, Aug 2010. ISSN 0035-

8711. doi: 10.1111/j.1365-2966.2010.16872.x. URL https://doi.org/10.1111/j.1365-2966.2010.16872.x.

[251] Kareem El-Badry, Andrew Wetzel, Marla Geha, Philip F. Hopkins, Dusan Kereš, T. K. Chan, and Claude-André Faucher-Giguère. Breathing fire: How stellar feedback drives radial migration, rapid size fluctuations, and population gradients in low-mass galaxies. , 820(2):131, Mar 2016. ISSN 1538-4357. doi: 10.3847/0004-637x/820/2/131. URL http://dx.doi.org/10.3847/0004-637X/820/2/131.

[252] Daniel Anglés-Alcázar, Claude-André Faucher-Giguère, Dušan Kereš, Philip F. Hopkins, Eliot Quataert, and Norman Murray. The cosmic baryon cycle and galaxy mass assembly in the FIRE simulations. , 470(4):4698, Oct 2017. doi: 10.1093/mnras/stx1517.

[253] 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, 11 2014. ISSN 0035-8711. doi: 10.1093/mnras/stu2058. URL https://doi.org/10.1093/mnras/stu2058.

[254] Kartheik G Iyer, Sandro Tacchella, Shy Genel, Christopher C Hayward, Lars Hernquist, Alyson M Brooks, Neven Caplar, Romeel Davé, Benedikt Diemer, John C Forbes, Eric Gawiser, Rachel S Somerville, and Tjitske K Starkenburg. The diversity and variability of star formation histories in models of galaxy evolution. *MNRAS*, 498(1):430, Oct 2020. ISSN 1365-2966. doi: 10.1093/mnras/staa2150. URL http://dx.doi.org/10.1093/mnras/staa2150.

[255] Elad Zinger, Annalisa Pillepich, Dylan Nelson, Rainer Weinberger, Rüdiger Pakmor, Volker Springel, Lars Hernquist, Federico Marinacci, and Mark Vogelsberger. Ejective

and preventative: the IllustrisTNG black hole feedback and its effects on the thermodynamics of the gas within and around galaxies. *MNRAS*, 499(1):768, Aug 2020. ISSN 0035-8711. doi: 10.1093/mnras/staa2607. URL https://doi.org/10.1093/mnras/staa2607.

[256] Dan Walters, Joanna Woo, and Sara L Ellison. Quenching time-scales in the IllustrisTNG simulation. *MNRAS*, 511(4):6126, Feb 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac283. URL https://doi.org/10.1093/mnras/stac283.

[257] Asa Bluck, Roberto Maiolino, Joanna M Piotrowska, James Trussler, Sara L Ellison, Sebastian F Sánchez, Mallory D Thorp, Hossen Teimoorinia, Jorge Moreno, and Christopher J Conselice. How do central and satellite galaxies quench? – insights from spatially resolved spectroscopy in the manga survey. *MNRAS*, 499 (1):230, Sep 2020. ISSN 1365-2966. doi: 10.1093/mnras/staa2806. URL http://dx.doi.org/10.1093/mnras/staa2806.

[258] Martina Donnari, Annalisa Pillepich, Gandhali D Joshi, Dylan Nelson, Shy Genel, Federico Marinacci, Vicente Rodriguez-Gomez, Rüdiger Pakmor, Paul Torrey, Mark Vogelsberger, and Lars Hernquist. Quenched fractions in the IllustrisTNG simulations: the roles of AGN feedback, environment, and pre-processing. , 500(3):4004, Oct 2020. doi: 10.1093/mnras/staa3006. URL https://doi.org/10.1093%2Fmnras%2Fstaa3006.

[259] Martina Donnari, Annalisa Pillepich, Dylan Nelson, Federico Marinacci, Mark Vogelsberger, and Lars Hernquist. Quenched fractions in the IllustrisTNG simulations: comparison with observations and other theoretical models. , 506(4):4760, Jul 2021. doi: 10.1093/mnras/stab1950. URL https://doi.org/10.1093%2Fmnras%2Fstab1950.

[260] Christoph Engler, Annalisa Pillepich, Gandhali D Joshi, Dylan Nelson, Anna Pasquali, Eva K Grebel, Thorsten Lisker, Elad Zinger, Martina Donnari, Federico Marinacci, Mark Vogelsberger, and Lars Hernquist. The distinct stellar-to-halo mass relations of satellite and central galaxies: insights from the illustristng simulations. *MNRAS*,

500(3):3957, Nov 2020. ISSN 1365-2966. doi: 10.1093/mnras/staa3505. URL http://dx.doi.org/10.1093/mnras/staa3505.

[261] Devontae C Baxter, M C Cooper, Michael L Balogh, Timothy Carleton, Pierluigi Cerulo, Gabriella De Lucia, Ricardo Demarco, Sean McGee, Adam Muzzin, Julie Nantais, Irene Pintos-Castro, Andrew M M Reeves, Gregory H Rudnick, Florian Sarron, Remco F J van der Burg, Benedetta Vulcani, Gillian Wilson, and Dennis Zaritsky. The GOGREEN survey: constraining the satellite quenching time-scale in massive clusters at z 1. *MNRAS*, 515(4):5479, Aug 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac2149. URL https://doi.org/10.1093/mnras/stac2149.

[262] Zhiying Mao, Tadayuki Kodama, Jose Manuel Pérez-Martínez, Tomoko L. Suzuki, Naoaki Yamamoto, and Kouta Adachi. Revealing impacts of stellar mass and environment on galaxy quenching. *AA*, 666:A141, 2022. doi: 10.1051/0004-6361/202243733. URL https://doi.org/10.1051/0004-6361/202243733.

[263] Yin Li, Yueying Ni, Rupert A. C. Croft, Tiziana Di Matteo, Simeon Bird, and Yu Feng. AI-assisted superresolution cosmological simulations. *PNAS*, 118(19), May 2021. doi: 10.1073/pnas.2022038118. URL https://doi.org/10.1073%2Fpnas.2022038118.

[264] Yueying Ni, Yin Li, Patrick Lachance, Rupert A C Croft, Tiziana Di Matteo, Simeon Bird, and Yu Feng. AI-assisted superresolution cosmological simulations II. halo substructures, velocities, and higher order statistics. , 507(1):1021, Jul 2021. doi: 10.1093/mnras/stab2113. URL https://doi.org/10.1093%2Fmnras%2Fstab2113.

[265] Han Aung, Daisuke Nagai, Anatoly Klypin, Peter Behroozi, Mohamed H Abdullah, Tomoaki Ishiyama, Francisco Prada, Enrique Pérez, Javier López Cacheiro, and José Ruedas. The Uchuu-UniverseMachine Dataset: Galaxies In And Around Clusters. , 519(2):1648, Dec 2022. doi: 10.1093/mnras/stac3514. URL https://doi.org/10.1093%2Fmnras%2Fstac3514.

[266] Peter Behroozi, Risa H Wechsler, Andrew P Hearin, and Charlie Conroy. UniverseMa-

chine: The correlation between galaxy growth and dark matter halo assembly from z = 0-10. *MNRAS*, 488(3):3143, May 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz1182. URL http://dx.doi.org/10.1093/mnras/stz1182.

[267] 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. , 488(3):3143, May 2019. ISSN 0035-8711. doi: 10.1093/mnras/stz1182. URL https://doi.org/10.1093/mnras/stz1182.

[268] Kartheik G. Iyer, Joshua S. Speagle, Neven Caplar, John C. Forbes, Eric Gawiser, Joel Leja, and Sandro Tacchella. Stochastic modeling of star formation histories. iii. constraints from physically motivated gaussian processes. , 961(1):53, Jan 2024. ISSN 1538-4357. doi: 10.3847/1538-4357/acff64. URL http://dx.doi.org/10.3847/1538-4357/acff64.

[269] Yangyao Chen, H. J. Mo, Cheng Li, Huiyuan Wang, Xiaohu Yang, Youcai Zhang, and Kai Wang. Relating the structure of dark matter halos to their assembly and environment. , 899(1):81, Aug 2020. doi: 10.3847/1538-4357/aba597. URL https://doi.org/10.3847%2F1538-4357%2Faba597.

[270] M Celeste Artale, Idit Zehavi, Sergio Contreras, and Peder Norberg. The impact of assembly bias on the halo occupation in hydrodynamical simulations. *MNRAS*, 480(3):3978, Aug 2018. ISSN 1365-2966. doi: 10.1093/mnras/sty2110. URL http://dx.doi.org/10.1093/mnras/sty2110.

[271] Antonio D. Montero-Dorta, Facundo Rodriguez, M. Celeste Artale, Rory Smith, and Jonas Chaves-Montero. Tracking the evolution of satellite galaxies: mass stripping and dark-matter deficient galaxies. *e-print arXiv: 2212.12090*, 2022. doi: 10.48550/arXiv.2212.12090. URL https://arxiv.org/abs/2212.12090.

[272] Sownak Bose, Daniel J Eisenstein, Lars Hernquist, Annalisa Pillepich, Dylan Nelson, Federico Marinacci, Volker Springel, and Mark Vogelsberger. Revealing the galaxy–

halo connection in illustristng. *MNRAS*, 490(4):5693, Sep 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz2546. URL http://dx.doi.org/10.1093/mnras/stz2546.

[273] Dylan Nelson, Annalisa Pillepich, Volker Springel, Rüdiger Pakmor, Rainer Weinberger, Shy Genel, Paul Torrey, Mark Vogelsberger, Federico Marinacci, and Lars Hernquist. First results from the TNG50 simulation: galactic outflows driven by supernovae and black hole feedback. , 490(3):3234, Aug 2019. doi: 10.1093/mnras/stz2306. URL https://doi.org/10.1093%2Fmnras%2Fstz2306.

[274] Annalisa Pillepich, Dylan Nelson, Volker Springel, Rüdiger Pakmor, Paul Torrey, Rainer Weinberger, Mark Vogelsberger, Federico Marinacci, Shy Genel, Arjen van der Wel, and Lars Hernquist. First results from the TNG50 simulation: the evolution of stellar and gaseous discs across cosmic time. , 490(3):3196, Sep 2019. doi: 10.1093/mnras/stz2338. URL https://doi.org/10.1093%2Fmnras%2Fstz2338.

[275] Sownak Bose, Boryana Hadzhiyska, Monica Barrera, Ana Maria Delgado, Fulvio Ferlito, Carlos Frenk, César Hernández-Aguayo, Lars Hernquist, Rahul Kannan, Rüdiger Pakmor, Volker Springel, and Simon D M White. The MillenniumTNG Project: the large-scale clustering of galaxies. , 524(2):2579, Jul 2023. ISSN 0035-8711. doi: 10.1093/mnras/stad1097. URL https://doi.org/10.1093/mnras/stad1097.

[276] César Hernández-Aguayo, Volker Springel, Rüdiger Pakmor, Monica Barrera, Fulvio Ferlito, Simon D. M. White, Lars Hernquist, Boryana Hadzhiyska, Ana Maria Delgado, Rahul Kannan, Sownak Bose, and Carlos Frenk. The millenniumtng project: high-precision predictions for matter clustering and halo statistics. , 524(2):2556, Jul 2023. ISSN 1365-2966. doi: 10.1093/mnras/stad1657. URL http://dx.doi.org/10.1093/mnras/stad1657.

[277] Ruediger Pakmor, Volker Springel, Jonathan P. Coles, Thomas Guillet, Christoph Pfrommer, Sownak Bose, Monica Barrera, Ana Maria Delgado, Fulvio Ferlito, Carlos Frenk, Boryana Hadzhiyska, César Hernández-Aguayo, Lars Hernquist, Rahul Kannan, and Simon D. M. White. The millenniumtng project: the hydrodynamical full physics

simulation and a first look at its galaxy clusters. , 524(2):2539, Jul 2023. ISSN 1365-2966. doi: 10.1093/mnras/stac3620. URL http://dx.doi.org/10.1093/mnras/stac3620.