

Understanding galaxy-halo connection using machine learning

by

Abhishek Jana

M.S., Indian Institutes of Science Education and Research Kolkata, 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

2024

Abstract

In this dissertation, we explore the galaxy-halo connection using advanced machine learning techniques to improve the accuracy and efficiency of modeling small-scale galaxy clustering and constraining Halo Occupation Distribution (HOD) parameters. The galaxy-halo connection, which describes how galaxies are distributed within dark matter halos, is a critical aspect of understanding cosmic structure formation. Traditional methods for studying this connection rely heavily on statistical models like HOD, but these approaches are computationally expensive and often lead to biased parameter estimates due to the complex nature of galaxy formation.

To address these challenges, we leverage a range of machine learning algorithms, including artificial neural networks (ANNs), random forests, and Bayesian regression models. Our analysis reveals that while many machine learning algorithms report good statistical fits, they can yield biased likelihood contours in both mean values and variances relative to the true model parameters. This underscores the importance of careful data processing, algorithm selection, and the transformation of training data to achieve unbiased and robust predictions. Among the methods tested, ANNs outperform random forests and ridge regression when the HOD parameter space is appropriately restricted. These machine-learning tools offer a promising approach to exploring the HOD parameter space with significantly reduced computational costs compared to traditional brute-force methods. Using our ANN-based pipeline, we successfully recreate standard results from the literature by analyzing key metrics such as the projected two-point correlation function ($w_p(r_p)$), angular multipoles of the correlation function ($\xi_\ell(r)$), and the void probability function (VPF). We show that while combining $w_p(r_p)$ and VPF improves parameter constraints, adding the multipoles ξ_0 , ξ_2 , and ξ_4 to $w_p(r_p)$ does not significantly enhance these constraints.

Additionally, we explore two auxiliary projects that use deep learning in fields outside of cosmology: a deep learning-based fingerprint verification system and a zero-shot classification framework for bioacoustics, specifically bird call recognition. Although these projects are not directly related to galaxy-halo studies, they demonstrate the adaptability of machine-learning techniques across diverse fields. The methods developed for biometric verification and species classification could inspire similar approaches in cosmology, particularly for handling sparse data and complex classification problems.

In summary, this dissertation advances the field of cosmology by integrating modern machine-learning approaches into the study of the galaxy-halo connection. Our results not only improve the efficiency and accuracy of modeling galaxy clustering but also highlight the broader applicability of machine learning techniques, offering new directions for future research in both cosmology and other scientific domains.

Constraining galaxy-halo connection using machine learning

by

Abhishek Jana

M.S., Indian Institutes of Science Education and Research Kolkata, 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

2024

Approved by:

Major Professor
Lado Samushia

Copyright

© Abhishek Jana 2024.

Abstract

In this dissertation, we explore the galaxy-halo connection using advanced machine learning techniques to improve the accuracy and efficiency of modeling small-scale galaxy clustering and constraining Halo Occupation Distribution (HOD) parameters. The galaxy-halo connection, which describes how galaxies are distributed within dark matter halos, is a critical aspect of understanding cosmic structure formation. Traditional methods for studying this connection rely heavily on statistical models like HOD, but these approaches are computationally expensive and often lead to biased parameter estimates due to the complex nature of galaxy formation.

To address these challenges, we leverage a range of machine learning algorithms, including artificial neural networks (ANNs), random forests, and Bayesian regression models. Our analysis reveals that while many machine learning algorithms report good statistical fits, they can yield biased likelihood contours in both mean values and variances relative to the true model parameters. This underscores the importance of careful data processing, algorithm selection, and the transformation of training data to achieve unbiased and robust predictions. Among the methods tested, ANNs outperform random forests and ridge regression when the HOD parameter space is appropriately restricted. These machine-learning tools offer a promising approach to exploring the HOD parameter space with significantly reduced computational costs compared to traditional brute-force methods. Using our ANN-based pipeline, we successfully recreate standard results from the literature by analyzing key metrics such as the projected two-point correlation function ($w_p(r_p)$), angular multipoles of the correlation function ($\xi_\ell(r)$), and the void probability function (VPF). We show that while combining $w_p(r_p)$ and VPF improves parameter constraints, adding the multipoles ξ_0 , ξ_2 , and ξ_4 to $w_p(r_p)$ does not significantly enhance these constraints.

Additionally, we explore two auxiliary projects that use deep learning in fields outside of cosmology: a deep learning-based fingerprint verification system and a zero-shot classification framework for bioacoustics, specifically bird call recognition. Although these projects are not directly related to galaxy-halo studies, they demonstrate the adaptability of machine-learning techniques across diverse fields. The methods developed for biometric verification and species classification could inspire similar approaches in cosmology, particularly for handling sparse data and complex classification problems.

In summary, this dissertation advances the field of cosmology by integrating modern machine-learning approaches into the study of the galaxy-halo connection. Our results not only improve the efficiency and accuracy of modeling galaxy clustering but also highlight the broader applicability of machine learning techniques, offering new directions for future research in both cosmology and other scientific domains.

Table of Contents

List of Figures	xiii
List of Tables	xxi
Acknowledgements	xxiii
Dedication	xxv
1 Introduction	1
1.1 Background and Motivation	1
1.2 Galaxy-Halo Connection Models	4
1.3 Halo Occupation Distribution (HOD) Model	11
1.3.1 Baseline HOD	11
1.3.2 Decorated Halo Occupation Distribution Models	13
1.4 Measurements in the Galaxy-Halo Connection	15
1.5 Application of Machine Learning in the Galaxy-Halo Connection	16
1.6 Mock Catalog Generation	17
1.7 Dissertation Structure	17
2 Clustering Statistics in the Galaxy-Halo Connection	20
2.1 Two-Point Correlation Function	20
2.1.1 Two-point Projected Correlation Function ($w_p(r_p)$)	22
2.1.2 Multipoles of the Redshift-space Correlation Function ($\xi_\ell(s)$)	24
2.2 Void Probability Function (VPF)	27
2.3 Counts-in-Cylinders (CiC)	29

2.4	Counts-in-Annuli (CiA)	30
3	Machine Learning Approaches in Galaxy-Halo Connection	32
3.1	Introduction	32
3.2	Overview of Machine Learning Techniques	33
3.2.1	Supervised Learning	33
3.2.2	Unsupervised Learning	34
3.2.3	Semi-Supervised Learning	35
3.2.4	Deep Learning	35
3.2.5	Transfer Learning	35
3.3	Supervised Learning Overview	36
3.3.1	Key Concepts of Supervised Learning	36
3.3.2	Types of Supervised Learning Algorithms	42
3.4	Application of Machine Learning in the Galaxy-Halo Connection	45
3.4.1	Emulating Galaxy Formation Models	45
3.4.2	Predicting Galaxy Properties from Halo Properties	46
3.4.3	Quantifying the Scatter in the Galaxy-Halo Connection	47
3.4.4	Constraining Galaxy-Halo Model Parameters	48
4	Constraining the HOD Model Using Machine Learning	49
4.1	Introduction	50
4.2	Creating Mock Galaxy Catalogs	52
4.2.1	Simulation Details	52
4.2.2	HOD Model	53
4.2.3	Galaxy Catalogs	54
4.3	Measurements	56
4.3.1	Projected Two-Point Correlation Function	56
4.3.2	Void Probability Function	56

4.3.3	Three-Dimensional Redshift-Space Two-Point Correlation Function	57
4.4	Covariance	59
4.5	Machine Learning Methods	60
4.5.1	Check the Distribution of the Data	61
4.5.2	Data Preprocessing	61
4.5.3	Split the Data into Training and Testing Sets	63
4.5.4	Fit Different ML Models to the Training Data	63
4.5.5	Hyperparameter Tuning	63
4.5.6	Perform Accuracy Test	65
4.5.7	Likelihood Analysis	66
4.6	Constraining HOD parameters	68
4.6.1	Effect of Data Transformations	69
4.6.2	Algorithm Choice	69
4.6.3	Number of Samples	71
4.6.4	Prior Parameter Ranges	74
4.7	Results and Conclusions	75
	Appendices	80
4.A	Machine Learning Overview	80
4.A.1	Artificial Neural Networks	80
4.A.2	Random Forest Regression	80
4.A.3	Ridge Regression (BR)	81
4.B	Evaluating Machine Learning Model	81
4.B.1	Mean Squared Error	82
4.B.2	Mean Absolute Error	82
4.B.3	R-squared	82
4.B.4	Adjusted R-Squared	83
4.B.5	Choice of Polynomial Features	85

5	Extended Analysis of Galaxy-Halo Connection with Machine Learning	86
5.1	Introduction	86
5.2	N-body Simulations	86
5.2.1	AbacusSummit Simulation Details	88
5.3	Machine Learning Pipeline for HOD Parameter Prediction	90
5.3.1	Overview of the ML Methods	92
5.3.2	Performance Evaluation and Results	94
5.3.3	Likelihood Contours and Results	95
5.4	Additional Results: Full Box Analysis	100
5.5	Conclusion	103
6	Other Project I: Using Deep Learning in Fingerprint Verification	105
6.1	Introduction	106
6.2	Preliminary Concepts: The Architecture of a Neural Fuzzy Extractor	109
6.2.1	The Classifier	109
6.2.2	The Expander	110
6.2.3	The Secure Sketch	110
6.2.4	The Hash	112
6.2.5	Security Considerations	113
6.3	An Instantiation of NFE Retrofit	116
6.3.1	Classifier and Expander Architecture	116
6.3.2	The Error-Correction Code	120
6.3.3	System Configuration Guidelines	121
6.4	NFE Evaluation	122
6.4.1	Datasets	122
6.4.2	Experimental Setup	123
6.4.3	Training of the Classifiers	124
6.4.4	Results	125

6.5	Conclusions	128
6.6	Acknowledgements	129
7	Other Project II: Few-Shot Classification in Bioacoustics	130
7.1	Introduction	130
7.2	Methods	131
7.3	Results	134
7.4	Conclusion	139
8	Conclusion	141
	Bibliography	143

List of Figures

1.1	Figure from Wechsler and Tinker ¹ shows approaches to modeling the galaxy-halo connection. The left panel shows the dark matter distribution in a $90 \times 90 \times 30 h^{-1}$ Mpc slice from a cosmological simulation, illustrating the cosmic web structure. The right panel presents the corresponding galaxy distribution using an abundance matching model, tuned to match the observed galaxy clustering properties. The table below categorizes different modeling techniques on a spectrum from physical models, which rely on direct simulations and detailed physical prescriptions (left), to empirical models, which use flexible parameterizations constrained by observational data (right).	5
3.1	The tradeoff between training error and validation error as model complexity increases, highlighting the risk of overfitting	38
3.2	Illustration of k-fold cross-validation, where the dataset is split into k equal folds. Each fold is used as the validation set exactly once, while the remaining $k - 1$ folds are used for training. This technique ensures a robust evaluation of the model's performance. Image sourced from the Scikit-learn documentation	40
4.1	HOD parameter distribution of the full box after filtering. The red boxes show the small box range (see Section 4.2.3).	55
4.2	Small-scale clustering measurements at the fiducial value from fitting the HOD model at $z = 1.1$. Top left: Void probability function as a function of r ; top right: projected correlation function as a function of r_p ; bottom: correlation function multipoles as a function of s	58

4.3	Jackknife covariance calculated at the fiducial value. The data vector is arranged from left to right as follows: bins 0–30 correspond to $w_p(r_p)$, bins 30–60 to $VPF(r)$, bins 60–70 to ξ_0 , bins 70–80 to ξ_2 , and bins 80–90 to ξ_4	59
4.4	M_{cut} distribution on full and small boxes (see Section 4.2.3). HOD parameters are drawn uniformly for both the full box and small box, but after applying filtering to the full box, the distribution becomes nearly Gaussian.	62
4.5	Exploring the impact of data transformation on ANN Method within the small Box. Left: Equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . The plot shows the convergence between transformed and non-transformed data. Right: Density plot between Transformed vs. Non-transformed Data as a function of Error/σ . The σ is calculated from the jackknife covariance matrix.	70
4.6	Exploring the impact of data transformation on BR Method within the small Box. Left: Equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . The plot shows the convergence between transformed and non-transformed data. Right: Density plot between Transformed vs. Non-transformed Data as a function of Error/σ . The σ is calculated from the jackknife covariance matrix.	70
4.7	Exploring the impact of different ML methods within the small box. Left: Equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . The plot shows the convergence test between the RF, BR, and ANN models. Right: Density plot between the RF, BR, and ANN models as a function of Error/σ . The σ is calculated from the jackknife covariance matrix.	71

4.8	Effect of increasing data points on Bayesian Ridge Regressor for the small box. This plot compares the density plot between the training set of 800, 1600, and 5400 for unscaled, SC, YJ, and BC transformation as a function of Error/σ . The σ is calculated from the jackknife covariance matrix.	72
4.9	Effect of increasing data points on ANN for the small box. This plot compares the density plot between the training set of 800, 1600, and 5400 for unscaled, SC, YJ, and BC transformation as a function of Error/σ . The σ is calculated from the jackknife covariance matrix.	73
4.10	Error Comparison Between full Box and small box. Left: Equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . This plot shows the convergence between different box sizes. Right: Density plot between small box and full box as a function of Error/σ . The σ is calculated from the jackknife covariance matrix.	74
4.11	Equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$. The ANN models are trained on the small box for w_p , VPF, ξ_0 , ξ_2 , and ξ_4 . The dashed lines denote the expected values of the HOD parameters. Contours are generated using the best-performing combinations of the ML models.	76
4.12	Effect of 1σ constraint on HOD parameters for a different combination of statistics. The blue lines denote the expected values of the HOD parameters. The white marks at the center of the red bars represent the mean of the MCMC samples after discarding 200 steps as burn-in.	77

4.B.1	Comparison of average percent RMSE Scores between unscaled, SC, BC, and YJ on BR and ANN as a function of polynomial features on w_p for the small box train and test set. We calculate the average percent RMSE (see Section 4.5.6 for definition) as a function of degree polynomial. The left plot shows the average percent RMSE of the BR model. The left plot shows the average percent RMSE of the ANN model.	84
4.B.2	Comparison of average percent RMSE Scores between BR and ANN for SC, YJ, and BC on w_p test set. The left, middle, and right plot shows the average percent RMSE of SC, YJ, and BC transformations respectively.	84
5.2.1	A visualization of the AbacusSummit base-resolution boxes taken from Maksimova et al. ² , showing progressive zoom-ins from the full box ($2h^{-1}$ Gpc) down to the cluster scale ($20h^{-1}$ Mpc). The 139 boxes that comprise the base-resolution suite are shown as tiles in the background. The renderings display the AbacusSummit base c000_ph000 simulation at $z = 0.1$, and the projections are $10h^{-1}$ Mpc deep. These state-of-the-art simulations, totaling roughly 60 trillion particles with particle mass $m_p \approx 2 \times 10^9 h^{-1} M_\odot$, span 97 cosmological models and provide a powerful tool for studying the evolution of the Universe and the formation of galaxies and clusters.	89
5.3.1	Flowchart illustrating the process of constraining HOD parameters using the AbacusSummit simulations and ML models. Left shows the brute force method of constraining HOD parameters by generating mock catalogs while running the MCMC chain. On the right, an ML model predicts clustering statistics directly from the HOD parameters, bypassing the need for generating mock catalogs and speeding up the parameter inference process.	91

5.3.2 Input parameter distribution on full and small boxes (see Section 4.2.3). HOD parameters are drawn uniformly for both the full box and small box, but after applying filtering to the full box, the distribution of M_{cut} becomes nearly Gaussian.	92
5.3.3 Density plot of VPF, ξ_0, ξ_2, ξ_4 between Transformed vs. Non-transformed Data as a function of Error/ σ . The σ is calculated from the jackknife covariance matrix. The plot shows the impact of data transformation on the BR Method within the small Box.	93
5.3.4 Density plot of VPF, ξ_0, ξ_2, ξ_4 between Transformed vs. Non-transformed Data as a function of Error/ σ . The σ is calculated from the jackknife covariance matrix. The plot shows the impact of data transformation on the ANN Method within the small Box.	93
5.3.5 Density plot of VPF, ξ_0, ξ_2, ξ_4 between the BR and ANN models as a function of Error/ σ . The σ is calculated from the jackknife covariance matrix. The plot shows the impact of different ML methods within the small box.	93
5.3.6 Comparison of predicted versus actual values for different observables (w_p, VPF, ξ_0, ξ_2 , and ξ_4) using the BR model. Each scatter plot displays the predicted values on the x-axis and the actual values on the y-axis, with colors representing different data bins. Histograms above the scatter plots represent the distribution of predicted values, while histograms on the right represent the distribution of actual values. We show the best-performing models for each statistic.	94

5.3.7 Comparison of predicted versus actual values for different observables (w_p , VPF , ξ_0 , ξ_2 , and ξ_4) using the ANN model. Each scatter plot displays the predicted values on the x-axis and the actual values on the y-axis, with colors representing different data bins. Histograms above the scatter plots represent the distribution of predicted values, while histograms on the right represent the distribution of actual values. We show the best-performing models for each statistic.	95
5.3.8 Exploring the impact of data transformation on BR Method within the small Box. Figure shows equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . The plot shows the convergence between different transformed data. The contours represent the 68% and 95% confidence intervals.	96
5.3.9 Exploring the impact of data transformation on ANN Method within the small Box. Figure shows equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . The plot shows the convergence between different transformed data. The contours represent the 68% and 95% confidence intervals.	97
5.3.10 Exploring the impact of BR and MLP methods within the small box. The figure shows equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . The plot compares the convergence between the BR and MLP Box-Cox methods. The contours represent the 68% and 95% confidence intervals.	98
5.3.11 Error comparison between the full box and small box. The figure shows equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . This plot compares the convergence between the full box (dark blue contours) and the small box (red contours). The contours represent the 68% and 95% confidence intervals.	99

5.4.1 Exploring the impact of data transformation on ANN Method within the full Box. Density plot between Transformed vs. Non-transformed Data as a function of Error/ σ . The σ is calculated from the jackknife covariance matrix.	101
5.4.2 1σ constraint on HOD parameters for a different combination of statistics on the full box. The blue lines denote the expected values of the HOD parameters. The white marks at the center of the red bars represent the mean of the MCMC samples after discarding 200 steps as burn-in.	103
6.1.1 Neural Fuzzy Extractor architecture.	107
6.2.1 Secure sketch construction. Left: The large circle contains all available embeddings from all users; the small circle for the authentic user (AU) is chosen to optimize the trade-off between false positives and false negatives (different AUs may have different radii). Middle: A codebook is built with a Voronoi region congruent to the AU's decision region; the center of the AU's decision region (r_i) is decoded to the nearest codeword (c_i), and the difference between the center and this codeword (d_i) is saved in the AU's record, along with the hash of the center of the decision region. Right: The AU provides a new sample for authentication (b); subtracting the difference vector (d_i) and decoding to the nearest codeword (c_i) retrieves the previously identified codeword. Adding the difference to this codeword (d_i) yields the center of the AU's decision region (r_i), which is hashed and compared to the AU's stored hash.	112
6.3.1 Triplet Loss architecture with the Siamese network: Three images "Anchor", "Positive", and "Negative") are processed through the same CNN simultaneously, generating a 128-dimensional vector in the final layer. These vectors are then passed through the triplet loss function, which minimizes the distance between "Anchor" and "Positive" while maximizing the distance between "Anchor" and "Negative".	118

6.3.2 Triplet Loss architecture: This architecture minimizes the distance between the “Anchor” and “Positive” while maximizing the distance between the “Anchor” and “Negative”.	120
6.3.3 Error rate vs. noise level for the Latin-Square LDLC with row/column degree d .	121
7.2.1 Overview of the classification pipeline, showcasing the key steps from data acquisition to final bird call detection.	131
7.2.2 Average spectrograms of vocalizations from the TBP (left) and PIP (right). The TBP shows a distinct low-frequency call with a clear curvature, while the PIP displays a flatter frequency distribution, indicating a broader and less variable frequency range.	132
7.2.3 Comparison of raw, bandpass filtered, and denoised spectrograms of a TBP bird call. Spectrograms are shown on a (0 - 80)dB scale. The raw spectrogram (left) displays the full frequency range of the original recording. The bandpass spectrogram (center) isolates the target frequency band of 150 Hz - 600 Hz, removing noise from other frequencies. The denoised spectrogram (right) further reduces the background noise, highlighting the essential features of the bird call.	132
7.2.4 Spectrogram of the uncertain audio samples after applying a bandpass filter and noise reduction. The red-highlighted regions indicate the call’s detected vocalization events.	133
7.3.1 Normalized Scores Heatmap of Clustering Quality Indices for various Deep Learning Models. The heatmap presents the performance of models across three key metrics: Silhouette Score (Silh), Davies-Bouldin Index (DB), and Dunn Index (Dunn). The values have been normalized for comparability.	135
7.3.2 Scatter plot matrix of the top 5 PCA components of Perch model	137
7.3.3 Scatter plot matrix of the top 5 PCA components of BirdNET model	138

List of Tables

4.1	Five-parameter HOD model. Uniform samples are drawn to populate the HOD model (see Section 4.2.3).	52
4.2	Hyperparameters for BR on Small Box	64
4.3	Hyperparameters for ANN on Small Box	64
4.4	Comparison of RF, BR, and ANN regression model performance on w_p for the small box test set. We calculate the average percent RMSE, average percent MAE, and average adjusted R^2 score to compare model performance (see Section 4.5.6 for definitions).	66
4.5	Performance for the 5 best and worst outputs of w_p on Box-Cox transformation for the small box test set. We compare RF, BR, and ANN models to report the percent RMSE.	67
4.6	Adjusted R^2 score of w_p on various transformations for the small box test set and full box test set. We compare RF, BR, and ANN models to report the adjusted R^2 score.	67
4.7	1σ error for the small box. The values are calculated by running the MCMC chain and using the ANN models for w_p , VPF, ξ_0 , ξ_2 , and ξ_4 .	75
5.4.1	Hyperparameters of BR for full box.	100
5.4.2	Hyperparameters for ANN on full box	100
5.4.3	1σ error for full box. The values are calculated by running the MCMC chain and using the ANN models for w_p , VPF, ξ_0 , ξ_2 , and ξ_4 .	102

6.4.1 Equal error rates (EER) and areas under the receiver operating characteristics (ROC) curve (AUROC) for our experiments using the FVC2006 database, with different neural network architectures.	125
6.4.2 Equal error rates (EER) and areas under the receiver operating characteristics (ROC) curve (AUROC) for our experiments using the PolyU database, with different neural network architectures.	126
6.4.3 Training and running times for our four experiments, under each of the three architectures, when using the FVC2006 database.	127
6.4.4 Training and running times for our four experiments, under each of the three architectures, when using the PolyU database.	127
7.3.1 Performance Metrics of Various Models Using Clustering Quality Indices. The table summarizes the overall clustering performance for different deep-learning models. 'N Components' is the number of components used in PCA.	136
7.3.2 Evaluation Metrics for Tooth-billed Pigeon (TBP) Using BirdNET and Perch Models.	137

Acknowledgments

As I stand at the threshold of completing my thesis, I find myself reflecting on the incredible journey that has brought me here. This story of academic pursuit and personal growth would not have been possible without the support, guidance, and encouragement of numerous individuals who have touched my life along the way.

First and foremost, I would like to express my deepest gratitude to my major professor, Dr. Lado Samushia. His guidance in laying the fundamental groundwork was invaluable, but what I appreciated most was the freedom he gave me to explore and carve my own path. That blend of structure and autonomy helped me grow not only as a researcher but also as an independent thinker.

I would also like to thank Dr. Pascal Hitzler and Dr. George Amariuca for giving me the opportunity to apply my deep learning knowledge in computer science. Dr. Hitzler's belief in my abilities provided the gateway, while Dr. Amariuca constantly checked in, supporting me throughout the project, and ensuring I stayed on track. I owe a great deal to Bipin, Monireh, and Sarker for their invaluable assistance along the way.

A special thanks to Dr. Leah Brickson from Colossal Biosciences for an unforgettable summer, where I had the chance to explore the application of deep learning in conservation—a learning experience that has left a lasting impression. And to Felipe, Ayusman, and Stefano, thank you for making that summer not only intellectually stimulating but incredibly fun.

I must also extend my appreciation to Dr. Bharat Ratra, whose effort in arranging graduate meetings with esteemed professors motivated me at key points during my journey.

I am indebted to Kim Coy for her unwavering support and for making me feel welcome when I first arrived in the USA. Her kindness, treating students like her own children, remembering my birthdays, and providing vital information at key stages of my graduation process have been deeply appreciated.

My heartfelt thanks to the Physics Department for providing the necessary resources and support that ensured my success. The department created an environment where learning, research, and collaboration thrived, and for that, I am incredibly grateful.

The sleepless quantum homework nights shared with Tomthin and Doyeong will forever be a cherished memory, as will the countless board games with Anbu, Thiba, Justin, Tyler, and Nelson. My sincere thanks to Siddharta and Ratan for the lively conversations and activities that made this journey enjoyable. And to Nelson and Nithin—thank you for the volleyball sessions that brought much-needed breaks to the routine. Special mention to Rahul and Sudeep, who kept me in the loop with current affairs and shared a steady stream of fun memes that brightened my days. I am deeply grateful to my department colleagues Hanyu, Jayashree, David, Joe, Narayan, and Giorgi for their camaraderie and support.

My heartfelt thanks to my thesis committee members—Dr. Lado Samushia, Dr. Glenn Horton-Smith, Dr. Bharat Ratra, and Dr. Pascal Hitzler—for their time, insights, and guidance. Their expertise has been pivotal in shaping this work.

Finally, I want to thank my parents, Mr. Rabin Jana and Mrs. Rupa Jana, and my wife, Ms. Jayashree Behera, for their unwavering support, motivation, and belief in me. They have been my anchor throughout this journey, and I could not have achieved this without them.

Thank you all for being a part of this adventure, one that has truly been a collective effort.

Dedication

This thesis is dedicated to the pillars of my life:

To my parents, Mr. Rabin Jana and Mrs. Rupa Jana,
whose unwavering support and sacrifices have been the foundation of my academic journey.
Your belief in my potential has been a constant source of inspiration.

And to my wife, Ms. Jayashree Behera, whose love, patience, and encouragement have been
my guiding light through the challenges of research and writing.
Your partnership has made this achievement possible.

Your collective faith in me has been the driving force behind this work.
This thesis stands as a testament to your enduring support and the values you've instilled
in me.

With deepest gratitude and love,
I dedicate this work to you.

Chapter 1

Introduction

1.1 Background and Motivation

In contemporary cosmological theories, approximately 83% of the universe's mass is thought to be composed of dark matter.³ This dark matter forms the essential framework or skeleton upon which galaxies form, evolve, and interact. Supported by a wide range of observational evidence, these well-established models suggest that small fluctuations in the matter distribution were generated in the universe's earliest moments during a rapid inflationary period. Over time, these initial perturbations were amplified by gravitational instability. At first, gas and dark matter were evenly mixed, but as the universe continued to evolve, gas cooled and began to accumulate at the centers of dark matter halos. In sufficiently massive halos, this gas could cool rapidly, leading to the formation of stars and the emergence of proto-galaxies. The power spectrum of matter indicates that smaller structures formed first, which then merged to create more massive ones. As dark matter halos merged, the galaxies within them grew through both internal star formation (in situ) and mergers with other galaxies (ex-situ). Processes such as star formation and feedback from supernovae within galaxies further influence their surroundings, affecting future gas accretion and star formation rates.

Within this theoretical framework, it becomes evident that the formation, internal properties, and spatial distribution of galaxies are intimately linked to the properties of their

host dark matter halos. In the cold dark matter (CDM) model, dark matter organizes itself into halos, while luminous matter forms galaxies. The relationship between the properties of galaxies and halos—both in terms of their physical characteristics and their statistical distributions—is referred to as the galaxy-halo connection. This concept encompasses the full range of properties of both dark matter halos and the galaxies that form within them. Understanding this connection is fundamental to addressing key questions in astrophysics and cosmology, such as:

Understanding the Physics of Galaxy Formation: What mechanisms control gas cooling, star formation, and feedback within galaxies? How can we leverage observational data, such as the spatial distribution of galaxies based on their characteristics, to infer these underlying physical processes? Statistical analyses of the galaxy-halo connection can integrate diverse datasets to answer these fundamental questions.

Constraining Cosmological Parameters: The latest generation of expansive imaging and spectroscopic surveys provides new opportunities to constrain cosmological models more tightly. These surveys can differentiate between models that involve a cosmological constant, dark energy, or modified gravity, estimate the mass of neutrinos, and place constraints on the inflationary potential. To fully utilize these measurements across different scales, it is crucial to robustly account for uncertainties in the galaxy-halo connection.

Exploring Dark Matter Properties and Distribution: How can galaxy surveys be employed to deduce the evolution of matter distribution and the characteristics of dark matter halos? Observations of galaxies and their spatial patterns can potentially map the distribution of dark matter and differentiate between various dark matter models. These observations typically depend on both the properties of dark matter particles and the galaxy-halo connection.

The concept of the galaxy-halo connection was first proposed based on early models of galaxy formation within the CDM framework, recognizing that the spatial distribution of galaxies could provide insights into their formation processes. For example, Peebles⁴ investigated galaxy clustering statistics, and early studies noted that massive galaxies and clusters exhibited different clustering properties compared to smaller, more typical galaxies.⁵⁻⁸

These differences in clustering behavior offer clues about the masses of the dark matter halos hosting these galaxies, as halo clustering is heavily influenced by halo mass.^{9;10} Additionally, researchers recognized early on that this relationship could be complex and vary with scale.^{11;12}

However, it was not until the late 1990s that cosmological simulations had the resolution needed to identify substructures within larger dark matter halos. Concurrently, large-scale galaxy surveys began. The APM survey¹³ was among the first to measure the galaxy correlation function for a large sample. Around the same time, Kravtsov and Klypin¹⁴ and Colín et al.¹⁵ analyzed substructures within cosmological simulations, finding power-law correlation functions consistent with the APM measurements. Major progress in the field was marked by the Two-degree Field Galaxy Redshift Survey (2dFGRS; Colless et al.¹⁶) and the Sloan Digital Sky Survey (SDSS; York et al.¹⁷), which provided groundbreaking measurements of the spatial clustering properties of large galaxy samples, allowing for separation based on properties like luminosity, color, stellar mass, and star formation rate (SFR). Observations of galaxies at high redshifts¹⁸ further enabled the first studies of galaxy clustering at earlier epochs.

Building upon prior initiatives, the Dark Energy Spectroscopic Instrument (DESI)¹⁹ is currently the largest ongoing spectroscopic survey targeting galaxies. DESI’s primary aim is to unravel the mysteries of dark energy and achieve highly accurate measurements of essential cosmological parameters. Over the next five years, DESI plans to obtain spectra for about 40 million galaxies and quasars, covering over 14,000 square degrees of the sky. This effort will significantly expand the scope of data collection, offering a 20-fold increase compared to earlier surveys such as the Baryon Oscillation Spectroscopic Survey (BOSS) and its extension, eBOSS, both conducted under the Sloan Digital Sky Survey (SDSS).¹⁷ The data collected by DESI are expected to provide valuable insights into the relationship between galaxies and their dark matter halos, as well as improve our broader understanding of the universe’s structure and evolution.

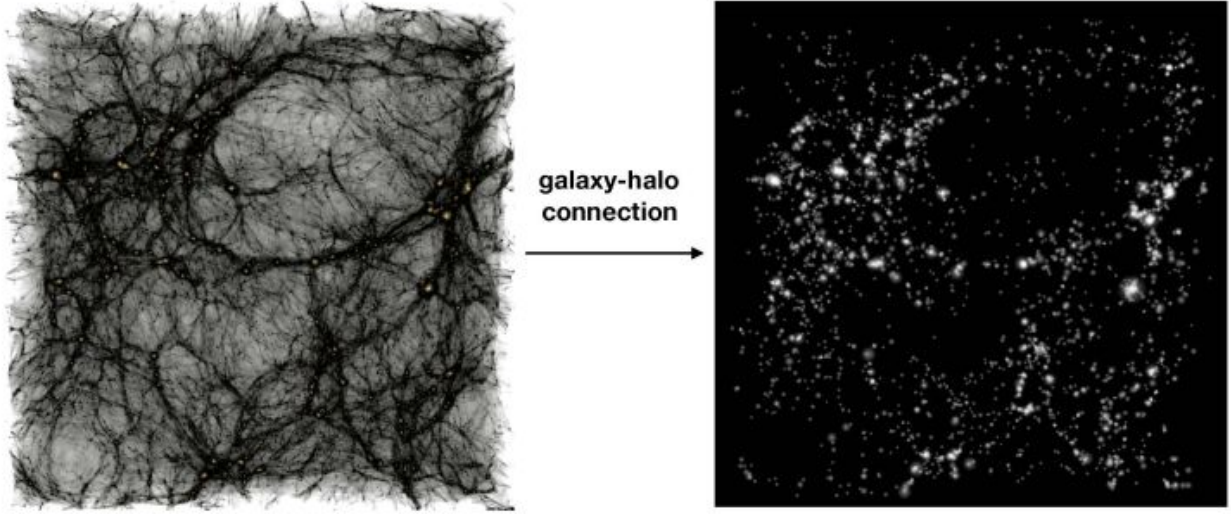
These advancements—the ability of numerical simulations to resolve dark matter substructures and the development of extensive galaxy surveys—have facilitated new methods

for statistically connecting galaxies and halos over the last 15 years. This review focuses on the statistical inference of the galaxy-halo connection made possible by these developments. We will highlight (a) theoretical methods to understand this connection, (b) key insights gained from studying the galaxy-halo connection, and (c) unresolved issues. The galaxy-halo connection is also intricately related to the halo model^{20–23}, an analytical framework for calculating non-linear dark matter clustering using halo properties (e.g., abundance and spatial clustering). The halo model can be integrated with galaxy-halo connection models to predict galaxy clustering behaviors.

The statistical relationship between galaxies and halos also benefits from another key development: the establishment of the standard cosmological model, Λ CDM, which suggests that the universe is composed of 5% baryonic matter, 25% dark matter, and 70% dark energy. The parameters of this model are now determined with high precision^{3;24–26}, allowing for robust predictions of cosmic structure growth and the formation and evolution of dark matter halos using numerical simulations. Although this review relies heavily on predictions from the Λ CDM model and the properties of dark matter halos, these topics have been comprehensively reviewed elsewhere.²⁷ The cosmological simulations that underpin many of the predictions discussed here were reviewed by Kuhlen et al.²⁸ The current status of physical models of galaxy formation is reviewed by Somerville and Davé²⁹, while unresolved theoretical issues are discussed by Naab and Ostriker³⁰; galaxy cluster formation is reviewed by Kravtsov and Borgani³¹. Modern cosmological probes using galaxy surveys were reviewed in Weinberg et al.³²; our focus here is on studies requiring a deep understanding of the galaxy-halo connection. The interaction between the galaxy-halo connection and small-scale dark matter models is thoroughly covered by Bullock and Boylan-Kolchin³³, so we will only briefly address these topics.

1.2 Galaxy-Halo Connection Models

To describe the connection between galaxies and halo mass, we explore various models, which can be broadly categorized into two main types: empirical models and physical models.



Approaches to modeling the galaxy-halo connection

← physical models		empirical models →		
Hydrodynamical Simulations	Semi-analytic Models	Empirical Forward Modeling	Subhalo Abundance Modeling	Halo Occupation Models
Simulate halos & gas; Star formation & feedback recipes	Evolution of density peaks plus recipes for gas cooling, star formation, feedback	Evolution of density peaks plus parameterized star formation rates	Density peaks (halos & subhalos) plus assumptions about galaxy—(sub)halo connection	Collapsed objects (halos) plus model for distribution of galaxy number given host halo properties

Figure 1.1: Figure from Wechsler and Tinker¹ shows approaches to modeling the galaxy-halo connection. The left panel shows the dark matter distribution in a $90 \times 90 \times 30 h^{-1}$ Mpc slice from a cosmological simulation, illustrating the cosmic web structure. The right panel presents the corresponding galaxy distribution using an abundance matching model, tuned to match the observed galaxy clustering properties. The table below categorizes different modeling techniques on a spectrum from physical models, which rely on direct simulations and detailed physical prescriptions (left), to empirical models, which use flexible parameterizations constrained by observational data (right).

Empirical models use observational data to derive parameters that describe the relationship between galaxies and dark matter halos, either at a specific epoch or over time. In contrast, physical models focus on simulating or parameterizing the physical processes that drive galaxy formation, such as gas cooling, star formation, and feedback mechanisms.

Figure 1.1 provides a schematic overview of these modeling techniques, showing an example of galaxy and dark matter distributions for one model and highlighting the key components of each approach.

It is important to understand that these modeling approaches lie on a continuum. As one moves to the right in the figure, there is a transition from models heavily grounded in physical theory to those more reliant on empirical data. Data-driven models are typically less predictive since they do not directly simulate fundamental physics but offer greater flexibility in adjusting parameters to fit observational data. These models are also less computationally demanding than those that attempt to directly model detailed physical processes.

Understanding the relationship between galaxies and their dark matter halos is crucial for developing a complete theory of galaxy formation and evolution. Several modeling approaches have been developed to connect observable galaxy properties to the underlying dark matter distribution predicted by cosmological simulations. We will discuss some of the popular models here and encourage readers to refer to Wechsler and Tinker¹ for a deeper understanding of the models.

1. Hydrodynamical Simulations: Hydrodynamical simulations are designed to directly simulate galaxy formation by solving the equations governing gravity and fluid dynamics within a cosmological framework. These simulations incorporate essential physical mechanisms such as gas cooling, star formation, feedback from stars, and feedback from active galactic nuclei (AGN). Although they require significant computational resources, hydrodynamical simulations like EAGLE³⁴, IllustrisTNG³⁵, and FIRE³⁶ have advanced considerably in generating realistic representations of galaxy populations. They provide detailed predictions about the baryonic content within halos and enable an in-depth examination of the physical processes that drive the galaxy-halo connection.

Some key advantages of hydrodynamical simulations include:

- Self-consistent modeling of baryonic physics and its impact on the structure of dark matter.
- The ability to track detailed gas dynamics and thermodynamic properties.
- Providing predictions for a range of observable galaxy properties, such as stellar mass, gas content, and metallicity.
- Sufficient resolution to examine internal structures within galaxies.

Initial research into halo occupation in hydrodynamical simulations occurred before empirical models had thoroughly constrained this aspect^{37–39}, building on earlier work that examined halo occupation in a semi-analytic framework, as seen in Benson et al.⁴⁰ These studies provided valuable early insights into modeling the halo occupation distribution (HOD). For instance, Zheng et al.⁴¹ employed smoothed particle hydrodynamics (SPH) simulations and semi-analytic models to propose initial forms for the HOD and the conditional luminosity function (CLF). These models were later refined using precise clustering data from the Sloan Digital Sky Survey. More recently, studies by Simha et al.⁴² and Chaves-Montero et al.⁴³ have used modern cosmological hydrodynamical simulations to test critical assumptions underlying the subhalo abundance matching technique.

While hydrodynamical simulations provide significant insights into the physical processes of galaxy formation, uncertainties remain due to the use of subgrid models for unresolved physics. It is essential to carefully compare these simulations to observational data to validate and refine the physical models they employ. The relationship between hydrodynamical simulations and empirical models is mutually beneficial: simulations can help shape empirical parameterizations, while empirical constraints can guide improvements in the subgrid physics used in simulations.

2. Semi-analytic Models (SAMs): Semi-analytic models (SAMs) offer a computationally efficient approach to studying galaxy formation by applying analytic formulas to dark matter-only N-body simulations. These models trace the formation and merging history of dark matter halos and use parameterized equations to represent key physical processes like

gas cooling, star formation, feedback from stars and active galactic nuclei (AGN), and other baryonic physics. Because SAMs are less computationally demanding than hydrodynamical simulations, they allow for extensive exploration of parameter space and the creation of large statistical samples.

Historically, the earliest SAMs were developed in the 1990s^{44–49}, and they have since been refined to include more detailed treatments of processes such as gas cooling, star formation, feedback mechanisms, chemical enrichment, and morphological evolution. Modern SAMs have achieved success in replicating a broad spectrum of observed galaxy properties and scaling relations over cosmic time.

Some key advantages of SAMs include:

- Efficiency allows the exploration of parameter space and large statistical samples.
- Modular nature facilitates testing different physical prescriptions.
- Capability to run on large cosmological volumes.

Semi-analytic models offer computational efficiency but face limitations compared to hydrodynamical simulations. They lack self-consistent modeling of baryonic effects on dark matter structure and often use simplified assumptions for complex baryonic processes. With 10 to 30 parameters, SAMs are challenging to constrain fully using observations. Recent efforts employ Monte Carlo Markov chains^{50–52} to better align parameters with data, but reconciling models with clustering and spatial statistics remains difficult. Notable SAMs include that of Somerville et al.⁵³, focusing on cosmic star formation and the stellar mass function; Henriques et al.⁵⁴, using MCMC to fit multiple observables; and SAGE⁵⁵, emphasizing modularity and efficiency. These models have provided insights into quenching mechanisms, merger impacts, and environmental effects on galaxy evolution over cosmic time.

As observational data sets become more precise, there is an increasing need for galaxy-halo models to evolve and incorporate more sophisticated physics to fully leverage the available data for both understanding galaxy formation processes and constraining cosmological

parameters. SAMs, when used in conjunction with hydrodynamical simulations and empirical models, provide a comprehensive toolkit for advancing our understanding of galaxy formation and the galaxy-halo connection in a cosmological context.

3. Subhalo Abundance Matching (SHAM): Subhalo Abundance Matching (SHAM) is an empirical method used to associate galaxies with dark matter halos and subhalos in N-body simulations by assuming a monotonic relationship between galaxy properties (such as stellar mass or luminosity) and halo properties (like mass or maximum circular velocity).^{56;57} This approach maintains the observed number density of galaxies by matching them to halos and subhalos ranked by their respective properties.

Key features of SHAM:

- Provides a straightforward method to link galaxies to both halos and subhalos with minimal assumptions.
- Effectively reproduces observed galaxy clustering across different scales and epochs.⁵⁸
- Efficient method for populating large cosmological simulations with galaxies.

SHAM typically ranks halos or subhalos by mass or V_{max} and matches them to galaxies ranked by stellar mass or luminosity, either at a single epoch or across time.⁵⁹ While SHAM is less physically detailed than hydrodynamical simulations or semi-analytic models, it is highly effective in replicating observed galaxy distributions and clustering patterns.⁶⁰ It is especially useful for creating mock galaxy catalogs for surveys. However, it makes simplifying assumptions, such as ignoring scatter between galaxy and halo properties and assuming the same relations for central and satellite galaxies.

Despite these simplifications, SHAM remains a valuable tool for studying the galaxy-halo connection and complements more detailed physical models. As models become more sophisticated and data more precise, SHAM continues to play a crucial role in testing and refining our understanding of galaxy formation and cosmology.

4. Halo Occupation Distribution (HOD) Models: Halo Occupation Distribution models specify the probability $P(N|M)$ for a halo of mass M to host N galaxies of a given

type. The HOD is typically parameterized separately for central and satellite galaxies. For central galaxies, a step-like function is usually assumed, while for satellites, a power-law is common. Under these assumptions, the HOD is fully characterized by its mean occupation number $\langle N|M \rangle$.^{22;61}

Some key features of HOD models include:

- Efficiency in exploring parameter space and generating large statistical samples.
- Flexibility to be constrained directly by clustering measurements.
- Ability to make predictions analytically when combined with the halo model.

The connection between modern HOD models and galaxy clustering measurements began to be explored in the early 2000s.^{21;22;40;41;61–64} The functional form of the HOD for mass- or luminosity-selected galaxies is generally assumed to be similar to that of dark matter subhalos within their hosts. Kravtsov et al.⁵⁶ found that the HOD for subhalo samples is well described by a power law $N \sim M$ for satellites, including the presence of a central galaxy.

HOD models have been widely used to interpret galaxy clustering measurements and constrain the galaxy-halo connection. Zheng et al.⁴¹ used both hydrodynamic simulations and semi-analytic models to propose functional forms for the HOD that were later constrained with clustering data from the Sloan Digital Sky Survey. More recent work has extended HOD modeling to incorporate assembly bias⁶⁵ and to jointly model multiple galaxy properties.^{66–68}

While less physically motivated than hydrodynamical simulations or semi-analytic models, HOD models provide a flexible framework for empirically connecting galaxies to halos. They serve as an important complement to more physical approaches and play a key role in cosmological analyses that aim to marginalize uncertainties in the galaxy-halo connection.

This thesis will primarily explore the relationships between galaxies and their dark matter halos using the Halo Occupation Distribution (HOD) framework. A comprehensive overview of the HOD approach will be provided in the next section.

1.3 Halo Occupation Distribution (HOD) Model

The Halo Occupation Distribution (HOD) model is a statistical framework used to describe how galaxies populate dark matter halos. It defines the probability $P(N|M)$ that a halo of mass M contains N galaxies of a particular type. The HOD model distinguishes between central galaxies, which reside at the centers of halos, and satellite galaxies, which orbit within halos.

1.3.1 Baseline HOD

Given a halo mass (M_{halo}), the model specifies the expected number of central (N_{cen}) and satellite (N_{sat}) galaxies. We consider central and satellite galaxies separately, as central galaxies reside at the potential well of the host halo, while satellite galaxies follow different formation physics. For Luminous Red Galaxies (LRGs), the HOD model is defined as a five-parameter model.⁶⁹

The average number of galaxies $\langle N(M_{\text{halo}}) \rangle$ in a halo of mass M_{halo} is given by

$$\langle N(M_{\text{halo}}) \rangle = \langle N_{\text{cen}}(M_{\text{halo}}) \rangle + \langle N_{\text{sat}}(M_{\text{halo}}) \rangle. \quad (1.1)$$

Central Galaxies:

The mean occupation function of central galaxies, $\langle N_{\text{cen}}(M_{\text{halo}}) \rangle$, is a step-like function with a characteristic minimum halo mass M_{cut} . It is given by

$$\langle N_{\text{cen}}(M_{\text{halo}}) \rangle = \frac{1}{2} p_{\text{max}} \left[1 + \text{erf} \left(\frac{\log_{10} M_{\text{halo}} - \log_{10} M_{\text{cut}}}{\sqrt{2} \sigma_{\text{M}}} \right) \right], \quad (1.2)$$

where p_{max} is the saturation level that controls the probability of hosting a central galaxy. For LRG-only samples, p_{max} is set to 1, meaning that most massive halos are guaranteed to possess a central galaxy. M_{cut} is the characteristic minimum mass of halos to host a central galaxy, and the parameter σ_{M} adjusts the shape of the central galaxy probability function by controlling the sharpness of the transition around M_{cut} . A higher σ_{M} results in a smoother

transition between 0 and 1 occupation probabilities.

Satellite Galaxies:

The mean occupation function of satellite galaxies, $\langle N_{\text{sat}}(M_{\text{halo}}) \rangle$, follows a power law with index α and characteristic satellite mass M_1 . It is given by

$$\langle N_{\text{sat}}(M_{\text{halo}}) \rangle = \langle N_{\text{cen}}(M_{\text{halo}}) \rangle \left(\frac{M_{\text{halo}} - \kappa M_{\text{cut}}}{M_1} \right)^\alpha, \quad (1.3)$$

for $M_{\text{halo}} > \kappa M_{\text{cut}}$, and zero otherwise. Here, κ determines the cut-off mass in units of M_{cut} .

Three additional parameters κ , α , and M_1 are used to determine the satellite probability:

- M_1 is the characteristic satellite mass.
- α is the power-law index of the satellite distribution.
- κ determines the cut-off mass relative to M_{cut} .

The baseline HOD model has been extensively utilized in numerous galaxy surveys to explore the galaxy-halo connection and interpret clustering measurements. In the Sloan Digital Sky Survey (SDSS), the HOD model was employed to analyze galaxy clustering across various environments.^{70–75} In the Baryon Oscillation Spectroscopic Survey (BOSS), it was used to model the distribution of Luminous Red Galaxies (LRGs), providing insights into the large-scale structure of the universe.⁷¹ The Dark Energy Survey (DES) applied the HOD framework to constrain cosmological models based on galaxy clustering data.⁷⁶ Similarly, the Kilo-Degree Survey (KiDS) utilized the HOD model for galaxy-galaxy lensing studies, helping to map the distribution of dark matter.⁷⁷ In the Hyper Suprime-Cam (HSC) Survey, the model was implemented to gain a deeper understanding of galaxy formation and evolution through clustering analysis.⁷⁸

The HOD model has proven particularly effective in characterizing the clustering properties of Luminous Red Galaxies (LRGs) in both the BOSS and eBOSS surveys. Beyond LRGs, it has also been applied to other cosmic tracers such as Bright Galaxy Samples (BGS), Emission Line Galaxies (ELGs)⁷⁹, and Quasi-Stellar Objects (QSOs).^{80;81} These applications have

successfully modeled the clustering behavior of diverse galaxy populations across multiple surveys, enhancing our understanding of the large-scale structure and the underlying dark matter distribution.

1.3.2 Decorated Halo Occupation Distribution Models

While the standard HOD model is successful at reproducing many observed galaxy clustering statistics, it makes simplifying assumptions that may not fully capture the complexity of galaxy formation physics. Decorated HOD models aim to extend the standard HOD by incorporating additional dependencies on halo properties beyond just mass.

The key idea behind decorated HOD models is to allow galaxy occupation to depend on secondary halo properties in addition to mass. This is typically implemented by modifying the standard HOD functions for central and satellite galaxies based on a secondary halo property S :

$$P(N|M, S) = P(N|M) [1 + \delta(S)], \quad (1.4)$$

where $\delta(S)$ represents the adjustment that modulates the occupation based on S . Common choices for S include halo concentration, formation time, or local environment.

Assembly Bias

One of the primary motivations for decorated HOD models is to account for assembly bias—the phenomenon where halo clustering depends on properties other than mass. In a decorated HOD framework, assembly bias can be incorporated by allowing the central galaxy occupation to depend on halo concentration c :

$$\langle N_{\text{cen}}|M, c \rangle = \langle N_{\text{cen}}|M \rangle [1 + A_{\text{cen}}(c - \bar{c})], \quad (1.5)$$

where $\bar{c}$ is the mean concentration at mass M , and A_{cen} controls the strength of the assembly bias effect. Similar adjustments can be applied to the satellite occupation.

Several studies have utilized decorated HOD models to search for evidence of assembly bias in observed galaxy clustering. For example, Zentner et al.⁸² applied a decorated HOD to SDSS clustering measurements and found that some luminosity-threshold samples preferred models with nonzero assembly bias. However, the strength and even the direction of the effect varied with the galaxy sample. More recently, Zehavi et al.⁸³ used SDSS DR7 data to constrain decorated HOD models, finding only weak evidence for assembly bias in most galaxy samples.

On the simulation side, Wechsler et al.⁸⁴ used N-body simulations to show that halo clustering depends on formation time at fixed mass, supporting the inclusion of assembly bias in HOD models. The IllustrisTNG hydrodynamical simulations have also been used to study galaxy assembly bias, with Bose et al.⁸⁵ finding significant assembly bias effects, especially for low-mass halos.

Velocity Bias

Standard HOD models typically assume that satellite galaxies trace the velocity distribution of dark matter within halos. However, evidence suggests that galaxies may be biased tracers of the underlying velocity field. Decorated HODs can account for velocity bias by modifying the velocity distribution of satellites:

$$\sigma_{\text{sat}} = \sigma_{\text{DM}} [1 + B_{\text{vel}}(M)], \quad (1.6)$$

where σ_{DM} is the velocity dispersion of dark matter, and $B_{\text{vel}}(M)$ introduces a mass-dependent velocity bias. This modification allows satellites to have systematically higher or lower velocities compared to dark matter.

Guo et al.⁸⁶ used SDSS redshift-space clustering measurements to constrain HOD models with velocity bias, finding evidence for nonzero velocity bias that varies with galaxy luminosity. The EAGLE hydrodynamical simulations have also been used to study velocity bias, with Armitage et al.⁸⁷ finding that satellite galaxies have lower velocity dispersions than dark matter, especially in massive clusters.

Conformity Bias

Galactic conformity refers to correlations between central and satellite galaxy properties beyond what would be expected from halo mass alone. Decorated HOD models can capture conformity effects by introducing correlations between central and satellite occupations:

$$P(N_{\text{sat}}|M, N_{\text{cen}}) = P(N_{\text{sat}}|M) [1 + A_{\text{conf}}(N_{\text{cen}} - \langle N_{\text{cen}} \rangle)], \quad (1.7)$$

where A_{conf} controls the strength of the conformity signal. This allows the number of satellites to depend on whether a central galaxy is present, in addition to halo mass.

Kauffmann et al.⁸⁸ found evidence for large-scale conformity in SDSS galaxies, motivating the inclusion of conformity in HOD models. However, Tinker et al.⁸⁹ showed that much of this signal was due to selection effects. More recently, Calderon et al.⁹⁰ used SDSS DR7 data to constrain decorated HOD models with conformity, finding only weak evidence for conformity on small scales.

In summary, decorated HOD models provide a flexible framework for extending the standard HOD approach to capture more complex galaxy-halo relationships. By incorporating effects like assembly bias, velocity bias, and conformity, these models aim to offer a more comprehensive description of how galaxies populate dark matter halos. As observational constraints on galaxy clustering continue to improve, decorated HODs may become increasingly important tools for interpreting data and connecting observations to models of galaxy formation.

1.4 Measurements in the Galaxy-Halo Connection

Clustering statistics play a crucial role in analyzing galaxy distributions and constraining the galaxy-halo connection. Key statistics include:

- **Two-point projected correlation function $w_p(r_p)$:** Measures the excess probability of finding galaxy pairs at a given projected separation, integrating along the line of sight to mitigate redshift-space distortions.

- **Multipoles of the redshift-space correlation function $\xi_\ell(s)$:** Characterize anisotropies in clustering due to peculiar velocities.
- **Void probability function (VPF):** Quantifies the probability of finding regions devoid of galaxies.
- **Counts-in-cylinders (CIC) and counts-in-annuli (CIA):** Measure galaxy density in cylindrical or annular volumes around other galaxies.

These statistics probe different aspects of the galaxy distribution on various scales. The two-point functions are most commonly used to constrain halo occupation models. Higher-order statistics like the three-point function provide additional information on galaxy-halo relations, while void statistics are sensitive to low-density regions.

By comparing observed clustering to predictions from galaxy-halo models, these statistics constrain how galaxies populate dark matter halos of different masses. This enables inferring the statistical relationship between galaxy and halo properties.

More details on clustering statistics and their application to the galaxy-halo connection will be provided in Chapter 2.

1.5 Application of Machine Learning in the Galaxy-Halo Connection

Machine learning (ML) is increasingly applied to study the galaxy-halo connection, offering powerful tools for uncovering complex relationships between galaxy properties and their host dark matter halos. Traditionally, models like the Halo Occupation Distribution (HOD) describe how galaxies populate halos based primarily on mass. However, ML techniques can go beyond this by considering a multitude of halo properties—such as concentration, velocity dispersion, or formation history—thus enabling more accurate predictions of galaxy clustering and formation patterns.

Random forest regressions and neural networks have been particularly effective in identifying secondary halo properties that correlate with galaxy features like stellar mass or star formation rates. These ML models have been trained on large cosmological simulations like IllustrisTNG and AbacusSummit to improve the accuracy of mock galaxy catalogs by better predicting the halo-galaxy occupation at different redshifts and halo masses. Moreover, ML-based approaches allow for flexible modeling of the stochastic scatter between galaxy and halo properties, capturing subtleties that traditional models might miss. This is especially valuable for next-generation galaxy surveys, such as those conducted by DESI and LSST, where the precision required for interpreting large datasets is paramount.

In short, machine learning offers an efficient and scalable method to constrain galaxy-halo connections, advancing beyond conventional models by integrating complex, high-dimensional data from simulations and observations.

1.6 Mock Catalog Generation

The HOD model is a fundamental tool, based on a probabilistic framework, for generating mock galaxy catalogs by populating dark matter halos from N-body simulations. Applying the HOD to halo catalogs enables the creation of mock galaxy catalogs that accurately reproduce observed galaxy clustering statistics. These catalogs are essential for estimating covariance matrices, testing analysis pipelines, and assessing systematic errors in large-scale structure studies. The HOD model’s flexibility allows it to be calibrated to match a wide range of galaxy properties and clustering measurements, making it an indispensable tool for interpreting galaxy surveys and constraining cosmological parameters.

1.7 Dissertation Structure

This dissertation is organized into eight chapters, each addressing different aspects of machine learning applications in both the galaxy-halo connection and other interdisciplinary projects. Below is an outline of each chapter:

Chapter 2: Clustering Statistics in the Galaxy-Halo Connection This chapter introduces key statistical tools used to examine the clustering of galaxies within the cosmic structure. It covers methods such as the two-point correlation function, void probability function (VPF), and counts-in-cylinders (CiC), which are used to measure galaxy clustering at various scales. These tools provide the statistical foundation for understanding how galaxies populate dark matter halos, which is crucial for connecting galaxy formation to the larger cosmic web.

Chapter 3: Machine Learning Approaches in Galaxy-Halo Connection Chapter 3 explores the application of ML techniques to study the galaxy-halo connection. Various machine learning algorithms, such as artificial neural networks (ANN), random forests, and supervised learning methods, are discussed. The chapter examines how these approaches are used to model the relationship between galaxy properties and their host dark matter halos, improving upon traditional models in terms of accuracy and computational efficiency.

Chapter 4: Constraining the HOD Model Using Machine Learning This chapter focuses on the HOD model, which is used to statistically connect galaxies to dark matter halos. The use of machine learning to constrain HOD parameters is examined, providing more efficient ways to model small-scale galaxy clustering. The chapter discusses the creation of mock galaxy catalogs and the use of different machine-learning methods to predict clustering statistics and derive cosmological constraints.

Chapter 5: Extended Analysis of Galaxy-Halo Connection with Machine Learning Chapter 5 expands upon the previous chapters by presenting additional results from applying machine learning techniques to large-scale cosmological simulations. It provides an in-depth analysis of the AbacusSummit simulations and evaluates the performance of different machine learning models in predicting HOD parameters. This chapter highlights the strengths and limitations of using machine learning for large-scale galaxy-halo studies.

Chapter 6: Other Project I - Using Deep Learning in Fingerprint Verification

This chapter explores an application of deep learning techniques in biometric security, specifically fingerprint verification. It introduces the architecture of a neural fuzzy extractor and evaluates its performance on various fingerprint datasets. The chapter demonstrates how deep learning models can be adapted for secure and efficient fingerprint verification.

Chapter 7: Other Project II - Zero-Shot Classification in Bioacoustics

Chapter 7 presents a project focused on the zero-shot classification of bird calls in bioacoustics. The challenges of training machine learning models for species detection with limited or no labeled data are addressed. The chapter showcases how clustering methods and machine learning models are applied to bioacoustic data to classify bird species based on their vocalizations, even in the absence of large training datasets.

Chapter 8: Conclusion The final chapter summarizes the key findings and contributions made throughout the dissertation. It emphasizes the role of machine learning in improving the understanding of the galaxy-halo connection and highlights its potential applications in ongoing and future cosmological surveys. The chapter concludes by discussing the broader implications of this research for the field of cosmology and outlines potential future research directions.

Chapter 2

Clustering Statistics in the Galaxy-Halo Connection

2.1 Two-Point Correlation Function

The two-point correlation function is a powerful statistical tool used to study the clustering and distribution of galaxies in the universe. It measures the excess probability of finding a pair of galaxies separated by a given distance compared to a random distribution. Estimating the two-point correlation function accurately is crucial for constraining cosmological models and understanding the dynamics of our universe.

The two-point correlation function (2PCF), denoted $\xi(r)$, quantifies the excess probability, compared to a random distribution, of finding a pair of galaxies separated by a distance r . Mathematically, the probability dP of finding two galaxies within volumes dV_1 and dV_2 , separated by r , can be written as:

$$dP = n^2 [1 + \xi(r)] dV_1 dV_2, \quad (2.1)$$

where dP is the probability of finding a galaxy in a volume element dV_1 and another galaxy in a volume element dV_2 , separated by a distance r , and n is the mean number density of

galaxies. When $\xi(r) > 0$, there is a higher likelihood of finding galaxy pairs at that separation compared to a random distribution. When $\xi(r) < 0$, the pairs are less likely to be found.

Various estimators have been proposed to calculate the two-point correlation function from galaxy catalogs. A study by Kerscher et al.⁹¹ compared nine of the most important estimators using a rigorous criterion based on subsamples of the Virgo Hubble Volume Simulation cluster catalog. The estimators were ranked by their ability to return a value within a certain tolerance of the “real” correlation function determined from the full survey. Some of the estimators compared in the study include the Davis & Peebles estimator⁹², the Hewett estimator⁹³, the Hamilton estimator⁹⁴, and the Landy & Szalay estimator (original⁹⁵ and grid version⁹⁶).

The comparison revealed that the Landy & Szalay estimator, in its original or grid version (Szapudi & Szalay), outperformed most of the other estimators and had a performance almost indistinguishable from the Hamilton estimator.

The Landy & Szalay estimator has emerged as the preferred choice for estimating the two-point correlation function in astrophysical applications. It is defined as:

$$\xi(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)}, \quad (2.2)$$

where $DD(r)$ is the number of data-data pairs, $DR(r)$ is the number of data-random pairs, and $RR(r)$ is the number of random-random pairs at separation r . This form effectively corrects for edge effects and minimizes variance. The factor of 2 in front of the $DR(r)$ term accounts for the reduction of bias associated with random pairings. The Landy & Szalay estimator has become the standard for most modern cosmological surveys due to its accuracy and robustness, especially in cases where survey volumes are not simple geometrical shapes and where the selection function is nontrivial.

Recent work by Keihänen et al.⁹⁷ presents a method called Linear Construction (LC) for fast evaluation of the covariance matrix for the Landy & Szalay estimator. The LC method estimates the covariance for small random catalogs and constructs the covariance for arbitrary random catalog sizes as a linear combination, providing a significant speedup

compared to the standard method.

2.1.1 Two-point Projected Correlation Function ($w_p(r_p)$)

The two-point projected correlation function, denoted as $w_p(r_p)$, is a fundamental tool in cosmology for studying the large-scale structure of the universe. It plays a crucial role in investigating the galaxy-halo connection within the Halo Occupation Distribution (HOD) framework.⁵

The 2PCF gives the excess probability, compared to a random distribution, of finding two galaxies separated by a distance r . In redshift surveys, galaxies' positions are measured in redshift space, which introduces distortions along the line of sight due to their peculiar velocities.

To correct for this, the correlation function is first expressed as $\xi(r_p, \pi)$, where r_p is the separation of galaxy pairs projected onto the plane of the sky, and π is the separation along the line of sight. This separates the total distance r into:

$$r^2 = r_p^2 + \pi^2. \quad (2.3)$$

The projected correlation function $w_p(r_p)$ is calculated by integrating the two-point correlation function $\xi(r_p, \pi)$ along the line-of-sight direction π , up to some maximum separation $\pi_{\max}$. This removes the distortions caused by galaxy velocities along the line of sight, focusing on the clustering signal in the plane of the sky:

$$w_p(r_p) = 2 \int_0^{\pi_{\max}} \xi(r_p, \pi) d\pi. \quad (2.4)$$

In this equation, the factor of 2 accounts for integrating along both positive and negative π values, and $\pi_{\max}$ is the maximum line-of-sight distance considered for the integration, typically chosen based on the dataset (e.g., $\pi_{\max} \sim 40 - 60 \text{ Mpc}/h$). The choice of $\pi_{\max}$ is important: too large a value introduces noise, while too small a value might exclude true correlations.⁹⁸ This projection effectively minimizes the impact of peculiar velocities, making

$w_p(r_p)$ a cleaner measure of real-space clustering.

Role of $w_p(r_p)$ in the HOD

The connection between galaxies and their host dark matter halos is a central focus in cosmology. The projected correlation function $w_p(r_p)$ serves as a powerful tool for constraining the parameters of the HOD model by capturing the clustering of galaxies at different scales. The clustering of galaxies on small and large scales reflects two distinct physical processes:

One-Halo Term: This term describes galaxy pairs within the same dark matter halo and dominates at small separations ($r_p \lesssim 1 \text{ Mpc}/h$).⁹⁹ It is primarily sensitive to the number and distribution of satellite galaxies in a halo. The one-halo term helps constrain key parameters such as the mean number of satellites per halo and the distribution of halo sizes.

Two-Halo Term: At larger scales ($r_p \gtrsim 1 \text{ Mpc}/h$), the clustering is dominated by galaxy pairs residing in different halos.¹⁰⁰ This term provides insights into how the overall galaxy population traces the large-scale distribution of dark matter and informs the bias of dark matter halos.

By fitting the observed $w_p(r_p)$ from galaxy surveys to HOD-based theoretical models, we can infer key parameters about the galaxy-halo connection^{101;102}, such as:

- The minimum halo mass required to host central galaxies.
- The fraction of halos that contain satellite galaxies and the average number of satellites as a function of halo mass.
- The scatter in the luminosity of central galaxies relative to halo mass.

These constraints allow us to probe how galaxies populate dark matter halos based on properties like mass, luminosity, and color.¹⁰³ For instance, studies show that the amplitude of $w_p(r_p)$ increases with galaxy luminosity and stellar mass, indicating that more luminous and massive galaxies reside in more massive halos.¹⁰⁴ Moreover, the shape of $w_p(r_p)$ can reveal additional details about the satellite fraction and the spatial distribution of galaxies within halos.¹⁰⁵

The HOD modeling of $w_p(r_p)$ has been applied across multiple galaxy surveys, including the Sloan Digital Sky Survey (SDSS)¹⁰⁶, the VIMOS Public Extragalactic Redshift Survey (VIPERS)¹⁰⁴, and the PRISM Multi-object Survey (PRIMUS)¹⁰³. These surveys span different redshift ranges and galaxy populations, providing diverse insights into the galaxy-halo connection. These studies underscore the versatility of $w_p(r_p)$ in connecting galaxy populations to their dark matter halos, offering invaluable insights into the evolution of large-scale structures in the universe.

2.1.2 Multipoles of the Redshift-space Correlation Function ($\xi_\ell(s)$)

In redshift space, the 2PCF becomes anisotropic due to the peculiar velocities of galaxies, which cause redshift-space distortions (RSD). To describe this anisotropy, the 2PCF is often expanded into multipole moments using Legendre polynomials. This decomposition enables cosmologists to separate the isotropic and anisotropic components of galaxy clustering.

The 2PCF in redshift space, $\xi(s, \mu)$, depends not only on the separation s between galaxy pairs but also on $\mu = \cos(\theta)$, where θ is the angle between the line of sight and the separation vector. The anisotropic 2PCF can be expanded into a series of Legendre polynomials $P_\ell(\mu)$, yielding the multipole moments $\xi_\ell(s)$:

$$\xi(s, \mu) = \sum_{\ell=0,2,4,\dots} \xi_\ell(s) P_\ell(\mu). \quad (2.5)$$

Here, ℓ represents the order of the multipole moment, typically taking even values (0, 2, 4, ...) because the correlation function must remain symmetric about the line of sight.

The multipole moments $\xi_\ell(s)$ can be obtained by projecting $\xi(s, \mu)$ onto Legendre polynomials using an integration over μ . The formula for the multipole moments is:

$$\xi_\ell(s) = \frac{2\ell + 1}{2} \int_{-1}^1 \xi(s, \mu) P_\ell(\mu) d\mu. \quad (2.6)$$

In this expression, $\xi(s, \mu)$ is the anisotropic 2PCF; $P_\ell(\mu)$ is the Legendre polynomial of order ℓ ; $\mu = \cos(\theta)$; and ℓ is typically even (0, 2, 4, ...) due to symmetry.

Each $\xi_\ell(s)$ corresponds to a specific angular component of the 2PCF:

- $\xi_0(s)$ is the monopole, representing the isotropic component of the clustering signal.
- $\xi_2(s)$ is the quadrupole, capturing the leading-order anisotropy due to redshift-space distortions.
- $\xi_4(s)$ is the hexadecapole, providing further detail on the anisotropic structure of the galaxy distribution.

The multipole moments of the 2PCF have been widely used in the analysis of galaxy surveys, such as the extended Baryon Oscillation Spectroscopic Survey (eBOSS). In a study by Bautista et al. ¹⁰⁷, the authors used the multipole moments $\xi_\ell(s)$ with $\ell = 0, 2, 4$ to measure the clustering of the eBOSS quasar sample and constrain cosmological parameters.

Application of $\xi_\ell(s)$ in HOD

The 2PCF is widely used in conjunction with the HOD model to study the anisotropic clustering of galaxies and constrain cosmological parameters. Here are some key applications of the 2PCF multipoles in the context of HOD modeling:

Small-Scale Clustering: One-Halo Term: At small separations ($s \lesssim 1 \text{ Mpc}/h$), the one-halo term dominates, corresponding to galaxies residing within the same halo. The random motions of satellite galaxies within halos, particularly those with large velocity dispersions, produce Finger-of-God distortions. These distortions are reflected in the quadrupole $\xi_2(s)$ and, to a lesser extent, the hexadecapole $\xi_4(s)$.

By fitting the small-scale quadrupole to observations, we can infer the velocity dispersion of satellites within halos, which is directly related to the halo mass. In this way, the multipole expansion helps constrain the satellite occupation and the dynamical structure of dark matter halos.

Large-Scale Clustering: Two-Halo Term: At large separations ($s \gtrsim 10 \text{ Mpc}/h$), the two-halo term dominates, describing galaxy pairs residing in different halos. The Kaiser effect, which enhances clustering along the line of sight due to coherent infall toward overdense

regions, primarily affects the quadrupole $\xi_2(s)$. By analyzing this effect, we can measure the linear galaxy bias b_g , which connects the galaxy distribution to the underlying matter distribution. This bias is a key parameter in the HOD model, determining how galaxies trace the dark matter distribution.

In the Baryon Oscillation Spectroscopic Survey (BOSS), researchers analyzed the monopole and quadrupole moments of the redshift-space correlation function to fit the HOD model. By fitting $\xi_0(s)$ and $\xi_2(s)$, they constrained the bias of luminous red galaxies (LRGs), finding that these galaxies are hosted in halos with masses $M_{\min} \sim 10^{13} M_{\odot}$ and exhibit significant velocity dispersions that affect the small-scale quadrupole.¹⁰⁸

Constraining HOD Parameters: The measured multipoles, particularly the monopole ξ_0 and quadrupole ξ_2 , can be used to constrain the HOD parameters. By comparing the observed multipoles with theoretical predictions from the HOD model, one can infer key parameters such as the mean occupation number of central and satellite galaxies as a function of halo mass.¹⁰⁹

For example, Zhai et al.⁷² used the monopole and quadrupole of the 2PCF measured from the SDSS-IV eBOSS LRG sample to constrain the HOD parameters. They found that the combination of ξ_0 and ξ_2 effectively breaks the degeneracy between the satellite fraction and the galaxy bias, allowing for robust constraints on the HOD.

Measuring the Growth Rate of Structure: The quadrupole moment of the 2PCF, ξ_2 , is sensitive to the redshift-space distortions caused by the peculiar velocities of galaxies. By modeling the RSD effect within the HOD framework, one can measure the growth rate of structure, $f\sigma_8$, which quantifies the rate at which cosmic structures grow due to gravitational instability.¹⁰⁹

In a study by Alam et al.⁷⁴, the authors used the multipole moments of the 2PCF measured from the SDSS-IV eBOSS LRG sample to constrain the growth rate of structure. They modeled the multipoles using the HOD approach, accounting for the RSD effect, and obtained precise measurements of $f\sigma_8$ at the effective redshift of the sample.

Testing Gravity Models: The 2PCF multipoles can also be used to test alternative gravity models within the HOD framework. By comparing the observed multipoles with predictions

from different gravity models, one can constrain deviations from general relativity.¹¹⁰

For instance, Armijo et al.¹¹¹ proposed a test of gravity using the marked correlation function, which assigns a density-dependent weight to galaxies in the estimation of the 2PCF. They applied this test to SDSS-III LRG samples, building mock catalogs using the HOD model for both general relativity and $f(R)$ gravity. Although the current data were insufficient to distinguish between the models, the authors demonstrated the potential of the method for future surveys.

2.2 Void Probability Function (VPF)

The Void Probability Function (VPF), $P_0(r)$, is a key statistical tool for studying the clustering properties of galaxies and the distribution of voids in the universe. It measures the probability of finding empty regions (voids) of a given radius r in a galaxy distribution. As a higher-order clustering statistic, it is sensitive to non-Gaussian features of large-scale structures and provides insights into galaxy formation and evolution, particularly in the context of Halo Occupation Distribution (HOD) models.

The VPF $P_0(r)$ is defined as the probability that a randomly placed spherical region of radius r contains no galaxies. For a Poisson-distributed random field of galaxies, this can be written as:

$$P_0(r) = \exp(-\bar{n}V), \quad (2.7)$$

where $\bar{n}$ is the mean number density of galaxies, and $V = \frac{4}{3}\pi r^3$ is the volume of the sphere.

For a clustered galaxy distribution, $P_0(r)$ is modified by galaxy correlations. In terms of the two-point correlation function $\xi_2(r_{12})$, it can be expressed as:

$$P_0(r) = \exp\left(-\bar{n}V \left[1 + \frac{1}{V} \int_V d^3\mathbf{r}_1 \int_V d^3\mathbf{r}_2 \xi_2(|\mathbf{r}_1 - \mathbf{r}_2|)\right]\right). \quad (2.8)$$

Here, the integral accounts for galaxy clustering within the spherical region, with $\xi_2(r_{12})$ representing the excess probability of finding galaxy pairs separated by r_{12} . As clustering

increases, the VPF decreases, reflecting the reduced likelihood of finding voids in the distribution.¹¹²

Application of VPF in HOD

The VPF is widely used in HOD models to constrain the galaxy-halo connection by probing underdense regions of the universe. In the HOD framework, galaxy populations within halos are modeled as a function of halo mass, describing both central and satellite galaxies. The VPF is sensitive to several HOD parameters, including:

- **Halo mass threshold ($M_{\min}$):** The minimum mass for halos to host galaxies. A higher $M_{\min}$ leads to a higher VPF, as fewer halos are populated by galaxies, increasing the number of voids.
- **Satellite fraction:** A higher number of satellite galaxies decreases the VPF because voids are less likely to occur in galaxy-rich environments.

In a seminal study, Tinker et al.¹¹³ used the VPF and the projected correlation function from the Sloan Digital Sky Survey (SDSS) to test whether galaxy occupation depends on the local environment. Their results showed that the standard HOD model, where galaxy occupation depends only on halo mass, provides a good fit for the observed VPF. This ruled out alternative models in which galaxy formation efficiency is significantly reduced in low-density environments.

Another application comes from Croton et al.¹¹⁴, who measured the VPF for red and blue galaxies in the 2dF Galaxy Redshift Survey (2dFGRS). They found that faint red galaxies significantly deviate from the “universal” negative binomial void probabilities followed by other populations, suggesting that the spatial distribution of galaxies depends on their color and luminosity. This study demonstrated that different galaxy populations create distinct void statistics, which can be linked to their formation histories and environments.

The evolution of the VPF with cosmic time provides insights into the growth of structure and the evolution of galaxy clustering. Conroy et al.¹¹⁵ measured the VPF at $z \sim 1$ using data

from the DEEP2 Redshift Survey and compared it to the VPF at $z \sim 0$ from the SDSS. They found that voids become larger at lower redshifts, consistent with predictions from Λ CDM simulations. This evolution in void size reflects the hierarchical growth of structure, where smaller voids at early times merge into larger voids as cosmic structures evolve.

The VPF provides a complementary statistic to the two-point correlation function $\xi(r)$, which focuses on galaxy-rich regions. Since the VPF probes voids, it is sensitive to higher-order correlation functions and non-Gaussian features of the large-scale structure. The VPF thus captures information about the three-point correlation function (3PCF) and higher moments of the galaxy distribution, making it a powerful tool for testing models of galaxy formation and evolution that rely on the large-scale distribution of dark matter and galaxies.

2.3 Counts-in-Cylinders (CiC)

The Counts-in-Cylinders (CiC) method, introduced by Reid et al. ¹¹⁶, measures the probability distribution of the number of galaxies within a cylinder of fixed radius and height centered on a galaxy. The cylinder is oriented along the line of sight to minimize the impact of redshift-space distortions. The CiC statistic is defined as:

$$P(N|r, h) = \frac{1}{N_{\text{gal}}} \sum_{i=1}^{N_{\text{gal}}} \delta_{N, N_i(r, h)}, \quad (2.9)$$

where N_{gal} is the total number of galaxies, $N_i(r, h)$ is the number of galaxies within a cylinder of radius r and height h centered on the i -th galaxy, and δ is the Kronecker delta function.

The CiC statistic is sensitive to higher-order moments of the galaxy density field and can be used to constrain the HOD model. By comparing the observed CiC distribution with predictions from HOD models, one can infer the occupation number of galaxies as a function of halo mass. ¹¹⁷

Application of CiC in HOD

The CiC method has been used to constrain HOD models and study the galaxy-halo connection. For example, Reid and Spergel¹¹⁸ used the CiC statistic measured from the SDSS to constrain the HOD of luminous red galaxies (LRGs). They modeled the CiC using a combination of the Navarro-Frenk-White (NFW) halo profile and the HOD framework, finding that the CiC provides strong constraints on the satellite fraction and the concentration of the satellite galaxy distribution within halos.

Similarly, Berrier et al.¹¹⁹ employed the CiC method to study the small-scale clustering of galaxies in the SDSS. They used a halo-based model with a simple HOD prescription to interpret the CiC measurements, demonstrating that the CiC can be used to constrain the HOD and probe the galaxy-halo connection at small scales.

Guo et al.⁸⁶ applied the CiC method to the SDSS Main Galaxy Sample and found that the satellite fraction decreases with increasing galaxy luminosity, consistent with predictions of the HOD model.

2.4 Counts-in-Annuli (CiA)

The Counts-in-Annuli (CiA) method is an extension of the CiC method that counts galaxies within annular regions instead of cylinders. The annuli are defined by a set of radial bins, allowing the measurement of the galaxy density field on different scales. The CiA statistic is defined as:

$$P(N|r_{\min}, r_{\max}, h) = \frac{1}{N_{\text{gal}}} \sum_{i=1}^{N_{\text{gal}}} \delta_{N, N_i(r_{\min}, r_{\max}, h)}, \quad (2.10)$$

where $r_{\min}$ and $r_{\max}$ are the inner and outer radii of the annulus, respectively.

The CiA method provides a more detailed characterization of the galaxy density field compared to the CiC method. It can be used to study the scale dependence of galaxy clustering and to constrain HOD model parameters, such as the satellite fraction and the concentration of satellite galaxies within halos.

Application of CiA in HOD

The CiA method has been applied to study the galaxy-halo connection and constrain HOD models. Watson et al. ¹²⁰ used the CiA statistic to measure the clustering of galaxies in the SDSS as a function of luminosity and color. They modeled the CiA using a halo model approach with a luminosity-dependent HOD, finding that the CiA provides valuable constraints on HOD parameters and the satellite fraction as a function of galaxy luminosity.

Guo et al. ¹²¹ measured the projected two-point correlation function and the CiA statistic for galaxy samples from the SDSS and the CMASS sample of the Baryon Oscillation Spectroscopic Survey (BOSS). They constructed mock galaxy catalogs using both the HOD and Subhalo Abundance Matching (SHAM) approaches and compared the model predictions with the observed clustering measurements.

Chapter 3

Machine Learning Approaches in Galaxy-Halo Connection

3.1 Introduction

Machine learning has emerged as a powerful tool for exploring the galaxy-halo connection, a fundamental aspect of modern cosmology and galaxy formation theory. These algorithms excel at identifying patterns and relationships in large, high-dimensional datasets, making them well-suited for studying the complex, nonlinear relationships between galaxy properties and their host dark matter halos.

Machine learning techniques can predict galaxy properties based on halo properties, classify galaxies and halos, and identify new features and correlations that may not be apparent through traditional methods such as the Halo Occupation Distribution (HOD) and the Conditional Luminosity Function (CLF).¹²² By training models on simulations and applying them to observational data, machine learning enables more direct comparisons between theory and observations.¹²³

In this chapter, we provide an overview of machine learning techniques relevant to the study of the galaxy-halo connection, discuss their applications in supervised learning for galaxy clustering, and explore how these methods are advancing our understanding of the

relationship between galaxies and their dark matter halos.

3.2 Overview of Machine Learning Techniques

Machine learning (ML) is a rapidly advancing field focused on developing algorithms that learn from data to make predictions or decisions autonomously. In astrophysics, particularly in the study of the galaxy-halo connection, ML techniques are increasingly employed to extract insights from large, complex datasets. These methods offer the ability to model the intricate relationships between galaxies and their host dark matter halos, a task traditionally performed through time-consuming simulations. Below is an overview of the primary categories of ML techniques and their relevance to the study of the galaxy-halo connection.

3.2.1 Supervised Learning

Supervised learning is a fundamental approach in ML where algorithms are trained using labeled datasets, meaning the input features (e.g., halo mass, galaxy luminosity) are paired with known output labels (e.g., galaxy type, halo occupation). The goal is to train a model that can generalize to unseen data by learning the mapping between inputs and outputs.

Classification

In classification tasks, the model predicts discrete output labels, such as identifying galaxy types (elliptical, spiral, etc.) or determining halo membership based on galaxy properties. Algorithms such as decision trees, random forests, support vector machines (SVMs), and neural networks are commonly used for classification tasks. For example, random forests have been applied to classify galaxies based on photometric features¹²⁴, and SVMs have been effective in differentiating between central and satellite galaxies¹²⁵.

Regression

Regression is used for predicting continuous values, such as estimating halo masses or galaxy luminosities. Techniques such as linear regression, polynomial regression, and more sophisticated models like neural networks can be applied to these tasks. In galaxy-halo studies, regression models can be trained to predict properties like star formation rates or halo masses from observational data¹²⁶.

3.2.2 Unsupervised Learning

Unsupervised learning algorithms work on datasets without labeled outputs. These methods seek to uncover hidden patterns, structures, or groupings within the data. In the galaxy-halo context, unsupervised learning can identify clusters of similar galaxies or halos based on intrinsic properties.

Clustering

Clustering algorithms group similar data points based on their features. Techniques like k-means, hierarchical clustering, and Gaussian mixture models (GMMs) are popular for this task. Clustering methods are valuable in large-scale structure studies, where galaxies or halos are grouped based on shared properties like mass or environment.¹²⁷ For example, hierarchical clustering can help identify galaxy clusters based on their spatial distribution.

Dimensionality Reduction

Dimensionality reduction techniques like Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) reduce the number of features in a dataset while preserving important patterns. These methods are crucial when working with high-dimensional dataset, such as galaxy catalogs with many features. PCA has been employed to simplify galaxy clustering datasets, reducing computational load while retaining the essential structure of the data¹²⁸.

3.2.3 Semi-Supervised Learning

Semi-supervised learning combines both labeled and unlabeled data, making it especially useful when labeled data is scarce or expensive to obtain. In astrophysics, where obtaining precise labels (e.g., galaxy masses) from observations is often difficult, semi-supervised learning leverages a small set of labeled galaxies and a larger set of unlabeled galaxies to improve prediction accuracy. For instance, semi-supervised techniques can enhance galaxy-halo modeling by using simulation data (labeled) alongside real observational data (unlabeled)¹²⁹.

3.2.4 Deep Learning

Deep learning, a subset of machine learning, focuses on neural networks with many layers that automatically learn hierarchical representations of data. These models are particularly well-suited for large, high-dimensional datasets such as those found in galaxy surveys.

In galaxy-halo connection studies, convolutional neural networks (CNNs) have been used to predict galaxy properties directly from dark matter density fields, leveraging spatial patterns in cosmological simulations.¹³⁰ Recurrent neural networks (RNNs) and other deep architectures have also been explored for time-series data, such as tracking the evolution of galaxy properties over cosmic time.¹³¹

Deep learning models excel at modeling complex, nonlinear relationships between galaxies and their halos, making them increasingly popular for high-precision predictions in astrophysical research.¹³²

3.2.5 Transfer Learning

Transfer learning allows knowledge gained from one domain to be applied to another, which is particularly useful in astrophysics. For example, models trained on cosmological simulations can be adapted to work with observational data or knowledge from one galaxy survey can be transferred to another with minimal retraining. In the context of the galaxy-halo connection, transfer learning has been used to adapt simulation-trained models for predicting

galaxy clustering in observational datasets like SDSS or DESI.¹³³ This technique significantly reduces training time while still providing accurate predictions for new datasets.

As the size and complexity of astronomical datasets continue to grow, the application of ML methods will become increasingly critical for extracting meaningful insights into the galaxy-halo connection and for making rapid, accurate predictions based on observational data.

3.3 Supervised Learning Overview

Supervised learning is a machine learning paradigm where a model learns to map input data to corresponding output labels using a labeled training dataset. The objective is to learn a function $f : X \rightarrow Y$ that can predict the correct output $y \in Y$ for a given input $x \in X$.

3.3.1 Key Concepts of Supervised Learning

Below, we discuss some essential concepts for solving a supervised machine learning problem.

Loss Function

A loss function $L(y, \hat{y})$ quantifies the discrepancy between the predicted output $\hat{y}$ and the true output y . The objective of supervised learning is to minimize the expected loss of the training dataset:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim D} [L(y, f_{\theta}(x))] \quad (3.1)$$

where θ represents the model parameters and D is the data distribution. Different loss functions are used for different tasks:

- *Mean Squared Error (MSE)* for regression tasks, where $L(y, \hat{y}) = (y - \hat{y})^2$.
- *Cross-Entropy Loss* for classification tasks, especially in multi-class classification problems.

Model Optimization

Model parameters θ are optimized using gradient-based methods, such as stochastic gradient descent (SGD). The gradients of the loss function with respect to the parameters are computed using backpropagation:

$$\theta \leftarrow \theta - \alpha \nabla_{\theta} L(y, f_{\theta}(x)) \quad (3.2)$$

where α is the learning rate, controlling the step size during each update. Other optimization methods include *Adam* and *RMSProp*, which adaptively adjust the learning rate based on gradients.

Generalization

In supervised learning, generalization refers to the model's ability to perform well on unseen data, not just the specific training dataset it was trained on. A model with good generalization learns patterns that are robust across different datasets. Ensuring generalization is critical to avoid overfitting, where the model performs well on the training dataset but poorly on new data.

Overfitting and Regularization

Overfitting occurs when a model learns to fit the training dataset too closely, capturing noise and irrelevant patterns. This results in poor performance on new data. Regularization techniques are employed to reduce overfitting by adding constraints to the model. Some common regularization techniques include:

- *L1 Regularization (Lasso)*: Adds the sum of the absolute values of the model parameters to the loss function, encouraging sparsity in the learned weights.

$$L_{\text{reg}}(y, \hat{y}) = L(y, \hat{y}) + \lambda \sum_i |\theta_i| \quad (3.3)$$

- *L2 Regularization (Ridge)*: Adds the sum of the squares of the parameters to the loss function, preventing large parameter values.

$$L_{\text{reg}}(y, \hat{y}) = L(y, \hat{y}) + \lambda \sum_i \theta_i^2 \quad (3.4)$$

- *Dropout*: Randomly drops out neurons during training, forcing the network to become more robust by not relying too heavily on any one neuron.
- *Early Stopping*: Stops training when performance on a validation set starts to degrade, preventing the model from overfitting to the training dataset.

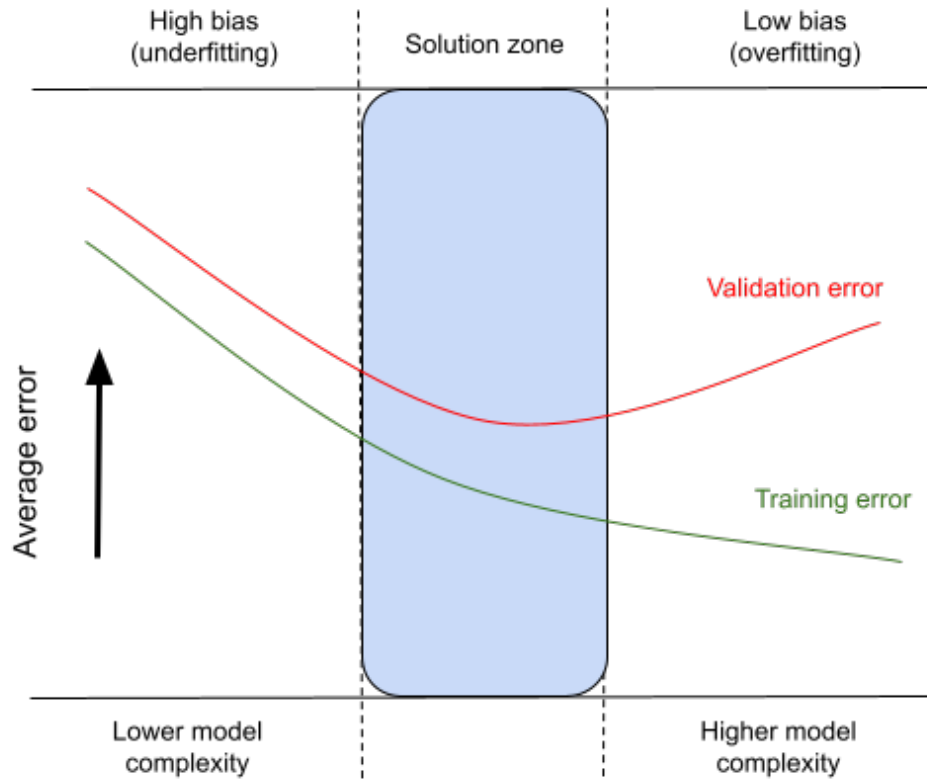


Figure 3.1: The tradeoff between training error and validation error as model complexity increases, highlighting the risk of overfitting

Bias-Variance Tradeoff

The bias-variance tradeoff is a fundamental concept in supervised learning. *Bias* refers to the error introduced by approximating real-world data with a simplified model, while *variance* refers to the model’s sensitivity to fluctuations in the training dataset.

- High-bias models (e.g., linear models) are often too simple and may underfit the data.
- High-variance models (e.g., deep neural networks) can overfit the training dataset and perform poorly on new data.

Balancing bias and variance is key to building models that generalize well. Figure 3.1 demonstrates how the training error (green curve) and validation error (red curve) behave as model complexity increases. Initially, both errors decrease, but after a certain point, the validation error starts to rise even as the training error continues to fall. This increase in validation error indicates overfitting, where the model becomes too specialized to the training dataset and loses its ability to generalize to unseen data. The shaded region represents the point of optimal model complexity, where both training and validation errors are reasonably low, indicating the best tradeoff between underfitting (when the model is too simple, like a linear fit) and overfitting (when the model is too complex, like a high-degree polynomial fit).

Cross-Validation

Cross-validation is a technique used to assess how well a model generalizes to unseen data. The dataset is split into multiple parts (or “folds”), and the model is trained on some folds while tested on others. K-fold cross-validation is a common method where the dataset is split into k parts, and the model is trained and tested k times, each time using a different fold for testing and the remaining $k - 1$ folds for training. This helps in tuning hyperparameters and avoiding overfitting.

As shown in Figure 3.2, in this example k-fold cross-validation divides the dataset into 5 equally sized folds. In each iteration, one fold is set aside as the validation set, and the model is trained on the remaining 4 folds. This process is repeated 5 times, ensuring that

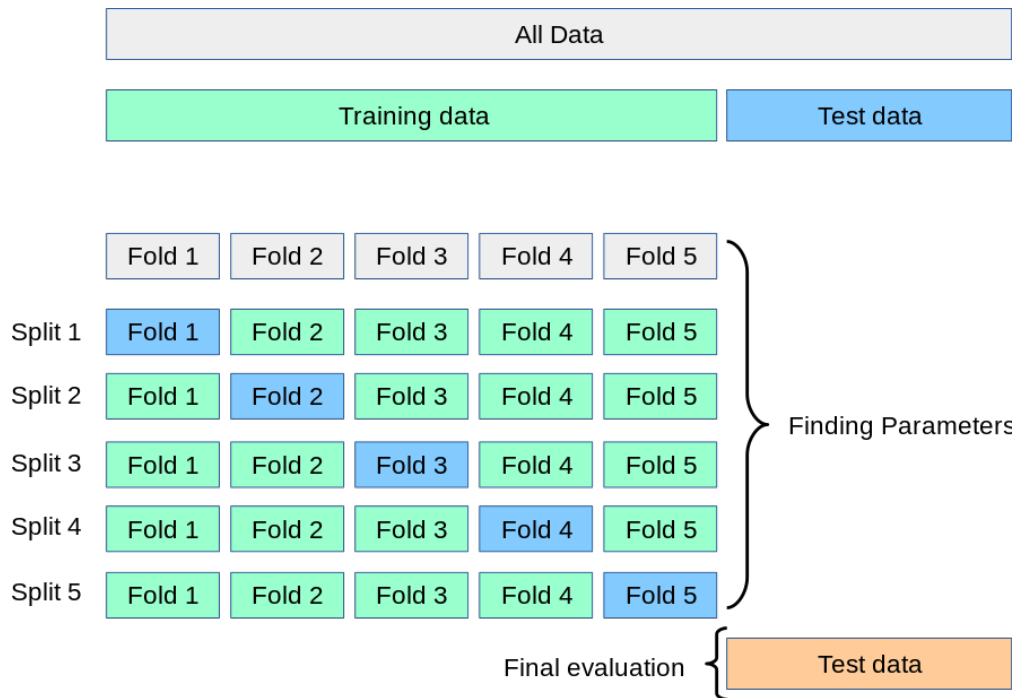


Figure 3.2: Illustration of k -fold cross-validation, where the dataset is split into k equal folds. Each fold is used as the validation set exactly once, while the remaining $k - 1$ folds are used for training. This technique ensures a robust evaluation of the model’s performance. Image sourced from the [Scikit-learn documentation](#).

every fold is used for validation exactly once. The final performance is averaged across all 5 runs, providing a more accurate estimate of the model’s ability to generalize to unseen data.

Hyperparameter Tuning

Models in supervised learning often have parameters that are not learned from the dataset but set manually before training, called *hyperparameters* (e.g., learning rate, regularization strength, number of layers in a neural network). Grid search and random search are commonly used for hyperparameter optimization, and more advanced techniques like Bayesian optimization can find the optimal set of hyperparameters more efficiently.

Feature Engineering

Feature engineering refers to the process of transforming raw data into meaningful input features that can be used to improve the model’s performance. Techniques include:

- *Normalization and Standardization*: Scaling features to a consistent range.
- *Polynomial Features*: Creating higher-order features by combining existing ones.
- *One-Hot Encoding*: Converting categorical variables into binary vectors.

Model Evaluation Metrics

Evaluating the performance of a supervised learning model is crucial. Some common metrics are explained below.

For classification tasks:

- *Accuracy*: Proportion of correctly classified instances.
- *Precision and Recall* (important for imbalanced classification problems):
 - *Precision* measures how many of the predicted positive instances are correct.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (3.5)$$

- *Recall* measures how many of the actual positive instances were correctly identified.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3.6)$$

- *F1 Score*: The harmonic mean of precision and recall, providing a single metric that balances both.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.7)$$

For regression tasks:

- *Mean Squared Error (MSE)*: Measures the average squared difference between the predicted values $\hat{y}_i$ and the true values y_i over all data points n :

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.8)$$

where y_i is the true value for the i^{th} data point, $\hat{y}_i$ is the predicted value, and n is the number of data points.

- *R-squared* (R^2): Explains the proportion of variance in the output that can be explained by the input features:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.9)$$

where $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ is the sum of squared residuals (unexplained variance), $\sum_{i=1}^n (y_i - \bar{y})^2$ is the total sum of squares (total variance), and $\bar{y}$ is the mean of the true values y_i . R^2 ranges from 0 to 1, where a value of 1 indicates perfect prediction and 0 means the model does not explain any variance. Negative values can occur if the model performs worse than simply predicting the mean.

Ensemble Methods

Ensemble methods combine multiple learning algorithms to improve prediction accuracy and robustness. Some common ensemble methods include:

- *Bagging*: Combines the predictions of multiple independent models trained on different subsets of the dataset. Random forests are a popular example.
- *Boosting*: Sequentially trains models where each new model focuses on correcting the errors of the previous ones. Gradient boosting and AdaBoost are common boosting algorithms.

3.3.2 Types of Supervised Learning Algorithms

Several supervised machine learning algorithms are commonly used. Below, we discuss some of the most popular methods.

Linear and Polynomial Regression

Linear regression models the relationship between input features and a continuous output variable using a linear function:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p, \quad (3.10)$$

where y is the predicted output, $x_1, x_2, \dots, x_p$ are the input features, and β_i are the coefficients that the model learns during training.

Polynomial regression is an extension of linear regression, where the relationship between input and output is modeled using higher-order terms. For instance, a second-degree polynomial regression would take the form:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_d x^d, \quad (3.11)$$

Polynomial regression can model more complex relationships but comes with the risk of overfitting if the degree d is too high.

Decision Trees and Random Forests

Decision trees are non-parametric supervised learning algorithms used for both classification and regression tasks. A decision tree splits the input space into regions by applying thresholds on feature values, with each split leading to a decision node. The leaves of the tree represent class labels or regression outputs. Decision trees use metrics like Gini impurity or information gain to decide where to split.¹³⁴

Random forests extend decision trees by training an ensemble of multiple trees, each on a random subset of the training dataset and features. The final prediction is obtained by averaging (in regression) or majority voting (in classification) across all the trees.^{135;136} Random forests help mitigate the overfitting problem inherent in individual decision trees

by reducing variance:

$$\hat{y}_{\text{final}} = \frac{1}{n} \sum_{i=1}^n \hat{y}_{i,\text{tree}}, \quad (3.12)$$

where n is the number of trees in the forest, and $\hat{y}_{i,\text{tree}}$ is the prediction from the i -th tree.

Support Vector Machines (SVMs)

Support Vector Machines (SVMs)¹³⁷ are used primarily for classification tasks, though they can also handle regression (Support Vector Regression). SVMs aim to find a hyperplane that best separates data points from different classes with the maximum margin. The optimization problem for a linear SVM is defined as:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{subject to} \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad \forall i, \quad (3.13)$$

where $\mathbf{w}$ is the normal vector to the hyperplane, b is the bias term, and y_i and $\mathbf{x}_i$ are the class label and feature vector for the i -th data point, respectively.

For non-linearly separable data, SVMs use kernel functions (e.g., polynomial, Gaussian RBF) to map data to a higher-dimensional space where a linear separator can be applied. This technique makes SVMs highly versatile for solving complex classification problems.

Neural Networks

Neural networks are powerful supervised learning models inspired by biological neurons.^{138;139} A neural network consists of an input layer, hidden layers, and an output layer. Each neuron in the hidden layers applies a non-linear activation function (such as ReLU, sigmoid, or tanh) to a weighted sum of inputs:

$$a = \sigma(\mathbf{w}^T \mathbf{x} + b), \quad (3.14)$$

where $\mathbf{w}$ are the weights, $\mathbf{x}$ is the input, b is the bias term, and σ is the activation function.

Deep neural networks (DNNs)¹⁴⁰ consist of multiple hidden layers and can capture complex, hierarchical relationships in the data. Training neural networks involves backpropagation, where gradients of the loss function with respect to weights are computed, and the

model parameters are updated using gradient descent:

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla_{\mathbf{w}} L(y, \hat{y}), \quad (3.15)$$

where α is the learning rate, $L(y, \hat{y})$ is the loss function, and $\hat{y}$ is the predicted output.

Neural networks are particularly effective for tasks such as image recognition, natural language processing, and other applications with large, high-dimensional datasets.

This overview of supervised machine learning techniques covers key concepts and algorithms. For readers interested in a deeper understanding, *Pattern Recognition and Machine Learning* by Bishop¹⁴¹ and *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* by Hastie et al.¹⁴² offer comprehensive insights into the theory and practical applications of supervised learning.

3.4 Application of Machine Learning in the Galaxy-Halo Connection

Machine learning techniques have been increasingly applied to study the galaxy-halo connection, leveraging their ability to uncover complex, high-dimensional relationships in large datasets. Below are key examples of how machine learning is being used to model and analyze the galaxy-halo connection.

3.4.1 Emulating Galaxy Formation Models

Neural networks and Gaussian processes are used to create fast, accurate emulators of computationally expensive galaxy formation models and N-body simulations.

Kamdar et al.^{143, 144} extended this approach to emulate the results of hydrodynamical simulations of galaxy formation, which are more computationally expensive than semi-analytic models (SAMs). They used artificial neural networks (ANNs) to learn the relationship between dark matter halo properties and galaxy properties from the Illustris simulation.

The trained emulators were able to accurately predict galaxy properties for halos within the simulation and for halos from independent simulations with different cosmological parameters. This showcased the flexibility and generalizability of machine learning-based emulators in exploring the impact of various physical processes and model assumptions on galaxy properties.

More recently, Lovell et al. ¹⁴⁵ introduced a hierarchical approach using normalizing flows to model the galaxy-halo relationship. Normalizing flows are a type of generative model that can learn the conditional probability distributions relating galaxy and halo properties from hydrodynamical simulations. By training on datasets from the CAMELS (Cosmology and Astrophysics with Machine Learning Simulations) suite of simulations—which cover a range of cosmological parameters and astrophysical models—the trained models could generate realizations of galaxy properties for arbitrary halo masses and cosmological parameters. The hierarchical structure of the normalizing flows allowed for capturing the complex, multi-scale nature of the galaxy-halo relationship.

3.4.2 Predicting Galaxy Properties from Halo Properties

Various machine learning algorithms have been applied to predict galaxy properties from halo properties in simulations like IllustrisTNG ¹⁴⁶. These models help elucidate the complex, stochastic mapping between galaxies and halos and can be used to populate dark matter-only simulations with realistic galaxy populations.

Delgado et al. ¹⁴⁷ demonstrated the successful application of machine learning techniques to model the complex galaxy-halo connection in the IllustrisTNG hydrodynamical simulation. By training models on a large datasets of galaxy and halo properties, they accurately predicted galaxy properties from halo properties, capturing the intricate, stochastic nature of the relationship. The machine learning models outperformed traditional halo occupation distribution models in describing the scatter in the galaxy-halo connection, highlighting the limitations of simple parametric models. The study revealed the importance of halo mass, concentration, and local density in predicting galaxy properties, providing valuable insights

into the physical processes that govern galaxy formation and evolution.

3.4.3 Quantifying the Scatter in the Galaxy-Halo Connection

Ensembles of neural networks with a Gaussian loss function have been used to predict probability distributions of galaxy properties given halo properties, rather than point estimates. This models the intrinsic scatter in the galaxy-halo connection and quantifies its dependence on halo properties, providing insights into the diversity of galaxy properties at fixed halo mass.

In Stiskalek et al. ¹²⁸, the authors applied machine learning to study the galaxy-halo connection in cosmological simulations (Horizon-AGN and IllustrisTNG100-1). They designed an ensemble of neural networks with a Gaussian loss function to predict galaxy property distributions from halo properties, modeling the scatter in the galaxy-halo connection. By extracting galaxy and halo variables from simulations, they quantified how well halo properties predict galaxy properties. They found that halo properties beyond mass explain up to 50% of the scatter in the halo-to-stellar mass relation, but additional halo properties do not significantly improve the prediction of stellar half-mass radius or total gas mass.

Munshi et al. ¹⁴⁸ predicted the stellar mass–halo mass (SMHM) relationship for dwarf galaxies using simulations across halo masses from 10^7 to 10^{11} solar masses. They found that the scatter in the SMHM relation increases for lower-mass halos, exceeding 1 dex at a halo mass of 10^8 solar masses. Using a probabilistic model, they populated “dark” halos below the resolution limit and predicted an intrinsic slope and scatter for the SMHM relation. They showed that at the faintest end of the SMHM relation, a galaxy cannot be assigned a unique halo mass based solely on luminosity. Instead, they provided a stochastic formula to populate low-mass halos, demonstrating that increasing scatter steepens the faint-end slope of the predicted stellar mass function.

3.4.4 Constraining Galaxy-Halo Model Parameters

Machine learning is being integrated into Bayesian inference frameworks to constrain the parameters of galaxy-halo connection models using observational data. By incorporating machine learning techniques, these methods offer a statistically rigorous way to link galaxy formation models to observable quantities. For example, recent studies have used machine learning to automatically identify giant radio galaxies and infer their number density and size distribution, helping to refine galaxy-halo models with greater precision. This approach improves the accuracy of empirical and physical models by directly modeling the relationship between galaxy and halo properties, offering more robust constraints on key parameters.

In Mitra et al. ¹⁴⁹, the authors present significant improvements to the Basilisk method, a Bayesian hierarchical model used to infer the galaxy-halo connection from satellite galaxy kinematics. Key advancements include updated galaxy selection processes, more accurate handling of interlopers and impurities, and extending kinematic modeling to higher orders by incorporating the kurtosis of the line-of-sight velocity distribution. Additionally, satellite abundance is used as a supplementary constraint. These enhancements enable the method to recover the conditional luminosity function for both central and satellite galaxies with unprecedented precision. Validated with mock data and applied to SDSS-DR7, the improved Basilisk method yields tighter constraints than previous approaches. It also breaks the mass-anisotropy degeneracy, providing accurate global measurements of satellite orbital velocity anisotropy, showing mildly radial orbits consistent with dark matter simulations. This paper establishes Basilisk as a highly effective tool for constraining galaxy-halo model parameters using satellite galaxy kinematics.

These methods enable the analysis of much larger datasets and more complex models than traditional approaches. They are particularly promising for current and upcoming large surveys like LSST, DESI, Euclid, and Roman.

Chapter 4

Constraining the HOD Model Using Machine Learning

This chapter is part of a literature which is currently under review as

Abhishek Jana, Lado Samushia - Constraining Galaxy-Halo Connection Using Machine Learning

Abstract

We investigate the potential of machine learning (ML) methods to model small-scale galaxy clustering for constraining Halo Occupation Distribution (HOD) parameters. Our analysis reveals that while many ML algorithms report good statistical fits, they often yield likelihood contours that are significantly biased in both mean values and variances relative to the true model parameters. This highlights the importance of careful data processing and algorithm selection in ML applications for galaxy clustering, as even seemingly robust methods can lead to biased results if not applied correctly. ML tools offer a promising approach to exploring the HOD parameter space with significantly reduced computational costs compared to traditional brute-force methods if their robustness is established. Using our ANN-based pipeline, we successfully recreate some standard results from recent literature.

Properly restricting the HOD parameter space, transforming the training data, and carefully selecting ML algorithms are essential for achieving unbiased and robust predictions. Among the methods tested, artificial neural networks (ANNs) outperform random forests (RF) and ridge regression in predicting clustering statistics, when the HOD prior space is appropriately restricted. We demonstrate these findings using the projected two-point correlation function ($w_p(r_p)$), angular multipoles of the correlation function ($\xi_\ell(r)$), and the void probability function (VPF) of Luminous Red Galaxies from Dark Energy Spectroscopic Instrument mocks. Our results show that while combining $w_p(r_p)$ and VPF improves parameter constraints, adding the multipoles ξ_0 , ξ_2 , and ξ_4 to $w_p(r_p)$ does not significantly improve the constraints.

4.1 Introduction

Due to gravitational instability, the clustering of matter on small scales is a highly nonlinear process that can only be accurately modeled using time-consuming N-body simulations. These simulations trace the evolution of a large number of dark matter particles over cosmic timescales. They can be used to study the gravitational collapse of dark matter particles into virialized halos. Multiple high-performance N-body simulation codes are available, allowing us to study the clustering properties of halos, the distribution of their properties, and the distribution of dark matter inside them with sub-percent precision.^{2;150–163}

Modeling the formation of galaxies inside dark matter halos is a more complicated process. To simulate it, we need to account for multiple intricate baryonic processes occurring during galaxy formation and evolution. Although several codes for simulating these baryonic processes are available, they are significantly more computationally demanding.^{34;164–168} The largest current pure dark matter N-body simulations have cumulative volumes of up to $(11.19 \text{ Gpc})^3$, whereas the most advanced simulations including baryons cover volumes of only $(740 \text{ Mpc})^3$.

One way to circumvent this problem is to populate the N-body simulations with galaxies based on a simplified prescription. One such approach is the Halo Occupation Distribution (HOD).^{21–23;41;61;63;169–171} The main assumption of the HOD approach is that the statistical

properties of galaxies within halos depend on the properties of the halos, primarily their mass. In the simplest version of the HOD, we assign galaxies to dark matter halos based on the host halo mass. The host halo mass determines the probability of that halo hosting a central galaxy and a certain number of satellite galaxies. The positions and velocities of these galaxies are assigned accordingly. The exact dependence of these probabilities on the halo mass is tuned to match the observed clustering of galaxies. More advanced HOD prescriptions also make the probabilities depend on other properties of the local environment, such as the local density of halos. These extra dependencies, collectively known as the assembly bias, are important in simulations^{83;172–175}, but have not been convincingly detected in real data¹⁷⁶.

To derive cosmological constraints from small-scale clustering, we need to marginalize over all possible HOD parameters that are consistent with the data at a given cosmology. This process can be very time-consuming. The space of HOD parameters is at least five-dimensional, and the statistics used to describe small-scale clustering are computationally intensive to compute. An increasingly popular method to speed up this process is to use machine learning to predict clustering as a function of HOD parameters.

The application of machine learning (ML) methods to predict cosmological observables has become a popular area of research. Artificial neural networks (ANN) have been used in Agarwal et al.¹⁷⁷; Gómez-Vargas et al.¹⁷⁸; Nygaard et al.¹⁷⁹; Perez et al.¹⁸⁰ to efficiently model galaxy-halo properties, emulate cosmological calculations, and constrain cosmological parameters. Recent works on deriving HOD constraints using ML methods mostly focus on techniques such as symbolic regression, random forest regression, and k-nearest neighbors.^{181–186}

In this work, we train ANN to predict key clustering measures as functions of the HOD parameters. This enables us to run Markov Chain Monte Carlo (MCMC) over the HOD parameters that converge in a few minutes, whereas a brute-force approach would take several days. We also investigate the potential of other ML techniques like random forests (RF) and Bayesian ridge regression (BR) to accelerate HOD modeling. We focus on modeling two well-known clustering metrics—the projected two-point correlation function $w_p(r_p)$ and the void probability function (VPF) measured from mock galaxy catalogs. We also measure

HOD Parameters					
Parameters	Fiducial	Full Box		Small Box	
		Max	Min	Max	Min
$\log M_{\text{cut}}$	13.0	14.5	11.5	13.17	12.83
$\log M_1$	14.24	16.24	12.24	14.33	14.14
σ	0.6	2	0	0.84	0.36
κ	0.98	2	0	1.36	0.59
α	0.4	2	0	0.59	0.2

Table 4.1: Five-parameter HOD model. Uniform samples are drawn to populate the HOD model (see Section 4.2.3).

the monopole (ξ_0), quadrupole (ξ_2), and hexadecapole (ξ_4) to compare with $w_p(r_p)$. Our goal is to train ML algorithms to rapidly predict these clustering signals for any point in HOD parameter space.

4.2 Creating Mock Galaxy Catalogs

To create mock galaxy catalogs from N-body simulations, we perform the steps described in the following subsections.

4.2.1 Simulation Details

For our analysis, we use the `AbacusSummit_base_c132_ph000` simulation from the broader emulator grid of the AbacusSummitⁱ N-body simulation suite.^{187–191} The simulation is conducted in a flat Λ CDM cosmology characterized by the following parameters: $\omega_b = 0.02237$, $\omega_{\text{cdm}} = 0.1200$, $h = 0.6736$, $A_s = 2.1791 \times 10^{-9}$, and $n_s = 0.9049$. The simulation box has a side length of $2000 h^{-1}$ Mpc and contains 6912^3 particles, totaling

ⁱ<https://abacussummit.readthedocs.io/en/latest/>

approximately 330 billion particles, with a particle mass of about $2 \times 10^9 M_\odot/h$.

Our study specifically focuses on box number 132 within this simulation suite, corresponding to a redshift of $z = 1.1$. We populate this box with galaxies using the HOD model, employing the parameters listed in Table 4.1. These galaxies are used for subsequent clustering analyses as part of our broader study of large-scale structures.

4.2.2 HOD Model

We use the HOD approach to generate galaxy catalogs from the AbacusSummit simulation. Given a halo mass (M_{halo}), the model specifies the expected number of central (N_{cen}) and satellite (N_{sat}) galaxies. We consider central and satellite galaxies separately, as central galaxies reside at the potential well of the host halo, while satellite galaxies follow different formation physics. The mean occupation function of central galaxies, $\langle N_{\text{cen}}(M_{\text{halo}}) \rangle$, is a step-like function with a characteristic minimum halo mass M_{cut} . The mean occupation function of satellite galaxies, $\langle N_{\text{sat}}(M_{\text{halo}}) \rangle$, follows a power law with index α and characteristic satellite mass M_1 . For Luminous Red Galaxies (LRGs), the HOD model is defined as a five-parameter model.⁶⁹ The average number of galaxies $\langle N(M_{\text{halo}}) \rangle$ in a halo of mass M_{halo} is given by

$$\langle N(M_{\text{halo}}) \rangle = \langle N_{\text{cen}}(M_{\text{halo}}) \rangle + \langle N_{\text{sat}}(M_{\text{halo}}) \rangle, \quad (4.1)$$

where

$$\langle N_{\text{cen}}(M_{\text{halo}}) \rangle = \frac{1}{2} p_{\text{max}} \operatorname{erfc} \left(\frac{\log_{10} M_{\text{cut}} - \log_{10} M_{\text{halo}}}{\sqrt{2} \log_{10}(e) \sigma_{\text{M}}} \right), \quad (4.2)$$

$$\langle N_{\text{sat}}(M_{\text{halo}}) \rangle = \mathcal{H} \left(\frac{M_{\text{halo}}}{\kappa M_{\text{cut}}} \right) \left(\frac{M_{\text{halo}} - \kappa M_{\text{cut}}}{M_1} \right)^\alpha \langle N_{\text{cen}}(M_{\text{halo}}) \rangle, \quad (4.3)$$

and

$$\mathcal{H} \left(\frac{M_{\text{halo}}}{\kappa M_{\text{cut}}} \right) = \begin{cases} 1, & M_{\text{halo}} > \kappa M_{\text{cut}}, \\ 0, & \text{otherwise.} \end{cases} \quad (4.4)$$

Here, p_{max} is the saturation level that controls the probability of hosting a central galaxy. For LRG-only samples, p_{max} is set to 1, meaning that most of the massive halos are guaranteed

to possess a central galaxy. M_{cut} is the characteristic minimum mass of halos to host a central galaxy, and the parameter σ_{M} adjusts the shape of the central galaxy probability function by controlling the sharpness of the transition around M_{cut} . A higher σ_{M} results in a smoother transition between 0 and 1 occupation probabilities.

Three additional parameters— κ , α , and M_1 —are used to determine the satellite probability. M_1 is the characteristic satellite mass, α is the power-law index of the satellite distribution, and κ determines the cut-off mass in units of M_{cut} .

4.2.3 Galaxy Catalogs

We generate 20,000 sets of HOD parameters by randomly selecting them within the bounds in Table 4.1 labeled as the “full box,” with uniform probability. We discard the HOD parameter sets that result in a galaxy sample with a density of less than $\bar{n} = 4.25 \times 10^{-4} h^3 \text{ Mpc}^{-3}$. At such low number densities, the clustering measurements are very noisy, making reliable interpolations difficult. The main galaxy samples relevant for cosmological analysis from large spectroscopic surveys, either currently available or expected in the near future (such as the Baryon Oscillations Spectroscopic Survey¹⁹², Dark Energy Spectroscopic Survey¹⁹³, Euclid satellite mission¹⁹⁴, and Roman Space Telescope¹⁹⁵) have number densities that are larger than our thresholds. Computing clustering measurements from a very large number of galaxies is time-consuming. To avoid extremely large computational times, we also discard the HOD parameter sets that result in a sample with a density higher than $\bar{n} = 1.75 \times 10^{-2} h^3 \text{ Mpc}^{-3}$. Both of these extreme number densities are sufficiently far from the samples obtained in realistic galaxy surveys to be of practical importance. After imposing additional cuts on the number density, we are left with 8,337 mocks out of 20,000.

To achieve a more reliable interpolation over a smaller parameter space, we separately generate 6,000 galaxy samples within tighter HOD parameter bounds. This reduced parameter range is labeled as the “small box” in Table 4.1. The range is chosen such that the galaxy number density is always well within the aforementioned range of $4.25 \times 10^{-4} h^3 \text{ Mpc}^{-3} < \bar{n} < 1.75 \times 10^{-2} h^3 \text{ Mpc}^{-3}$.

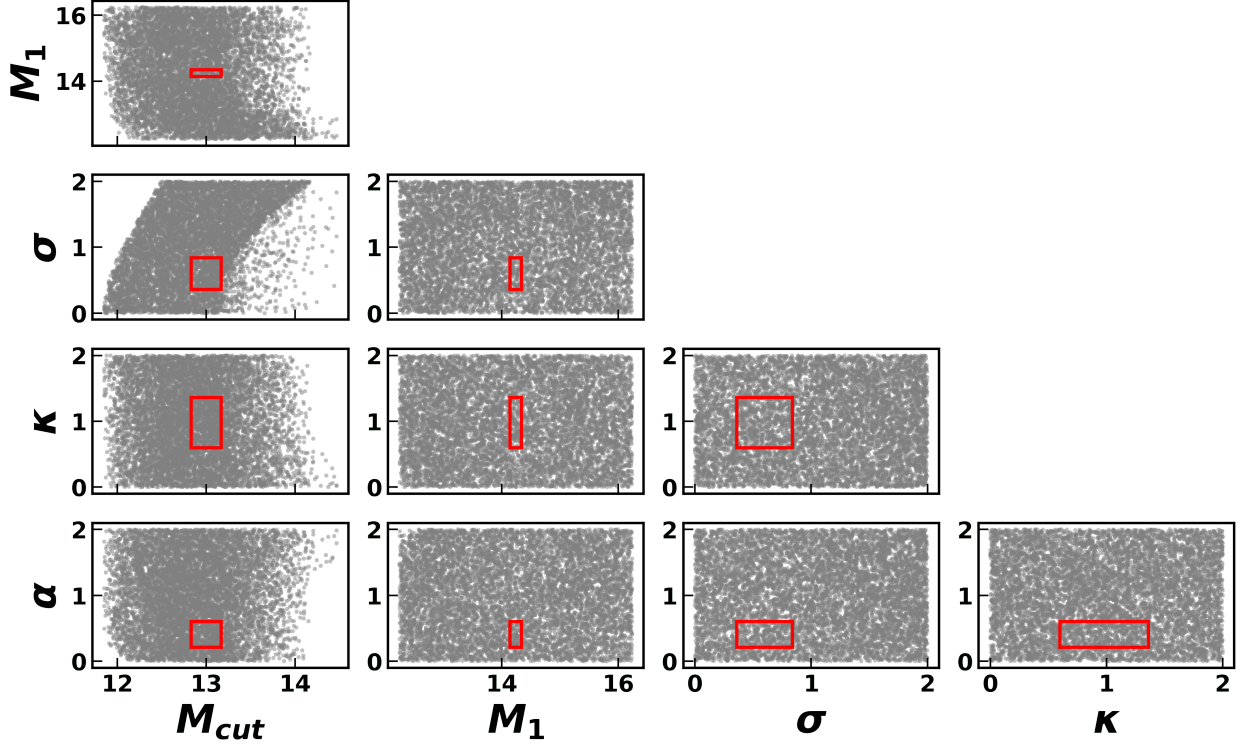


Figure 4.1: HOD parameter distribution of the full box after filtering. The red boxes show the small box range (see Section 4.2.3).

The full box covers a very wide range of possible HOD parameter values that can describe all the galaxy samples from current and future surveys relevant to cosmological analysis. The small box covers a range of values smaller than the precision currently allowed by clustering measurements. We expect the interpolation within the small box to be much more reliable. We will see in Section 4.6.3, however, that even within the small box the interpolation sometimes encounters serious problems. The size of the small box that is safe enough for interpolation will likely change for extended HOD models. Gray points in Figure 4.1 show the HOD parameter values used in this work. The outer edges of the plots correspond to the full box cuts, while the smaller red rectangles correspond to the small box cuts. Some areas in the parameter space are not covered because they were ruled out by the constraints imposed on the number density.

For each point in the parameter space, we create a galaxy sample from the AbacusSummit simulation. We place a central galaxy at the center of mass of the dark matter halo based

on a probability computed with Equation 4.2. We assign the mean velocity of halo particles to this galaxy. We determine how many satellite galaxies to place in each halo by randomly drawing a number from a Poisson distribution with the mean given by Equation 4.3. We then randomly select that many dark matter particles from the halo and assign their positions and velocities to the satellites.

4.3 Measurements

For each generated galaxy sample, we compute several key clustering measures. Each measurement is computed in redshift space by transforming the galaxy coordinates from x, y, z to $x, y, z + v_z/(aH)$, where we choose the z -direction to be our line of sight. Here, v_z is the z -component of the galaxy velocity, a is the scale factor (0.47619), and H is the Hubble parameter ($67.36 \text{ km s}^{-1} \text{ Mpc}^{-1}$).

4.3.1 Projected Two-Point Correlation Function

The main statistic used in our tests is the projected two-point correlation function,

$$w_p(r_p) = 2 \int_0^{\pi_{\max}} \xi(r_p, \pi) d\pi, \quad (4.5)$$

where $\xi(r_p, \pi)$ quantifies the excess probability over random of finding another galaxy at a separation defined by the projected distance r_p and the line-of-sight distance π . We choose $\pi_{\max} = 60 h^{-1} \text{ Mpc}$ to include most correlated pairs while suppressing noise from distant uncorrelated pairs, resulting in a stable measurement.

We use 30 logarithmically spaced bins ranging from $r_p = 0.1 h^{-1} \text{ Mpc}$ to $r_p = 31.6 h^{-1} \text{ Mpc}$.

4.3.2 Void Probability Function

To calculate the void probability function (VPF), we randomly place N_{sphere} spheres of radius r in the simulation box and count the number N_{void} of spheres containing no galaxies (void

spheres). The VPF is defined as

$$\text{VPF}(r) = \frac{N_{\text{void}}(r)}{N_{\text{sphere}}(r)}. \quad (4.6)$$

We use $N_{\text{sphere}} = 10^5$ spheres. To compute $\text{VPF}(r)$, we use 30 logarithmically spaced radii from $r = 1 h^{-1} \text{ Mpc}$ to $r = 10 h^{-1} \text{ Mpc}$. We employ `Halotools`ⁱⁱ (v0.7)¹⁹⁶ to compute both VPF and w_p .

4.3.3 Three-Dimensional Redshift-Space Two-Point Correlation Function

We compute the auto-correlation function using the Landy-Szalay estimator⁹⁵:

$$\xi(s, \mu) = \frac{DD(s, \mu) - 2DR(s, \mu) + RR(s, \mu)}{RR(s, \mu)}, \quad (4.7)$$

where $DD(s, \mu)$, $DR(s, \mu)$, and $RR(s, \mu)$ are the counts of galaxy-galaxy, galaxy-random, and random-random pairs, respectively, as functions of the separation s and the cosine of the angle to the line of sight μ .

We then compute the angular multipoles of the correlation function:

$$\xi_\ell(s) = \frac{2\ell + 1}{2} \int_{-1}^1 \xi(s, \mu) \mathcal{L}_\ell(\mu) d\mu, \quad (4.8)$$

where $\mathcal{L}_\ell(\mu)$ is the Legendre polynomial of order ℓ . We measure the correlation function monopole $\xi_0(s)$, quadrupole $\xi_2(s)$, and hexadecapole $\xi_4(s)$. We use 10 equally spaced bins in s between $2 h^{-1} \text{ Mpc}$ and $20 h^{-1} \text{ Mpc}$, with 100 bins in μ .

We utilize `Corrfunc`^{iii 197;198} for calculating ξ_ℓ .

ⁱⁱ<https://halotools.readthedocs.io/en/latest/>

ⁱⁱⁱ<https://corrfunc.readthedocs.io/en/master/>

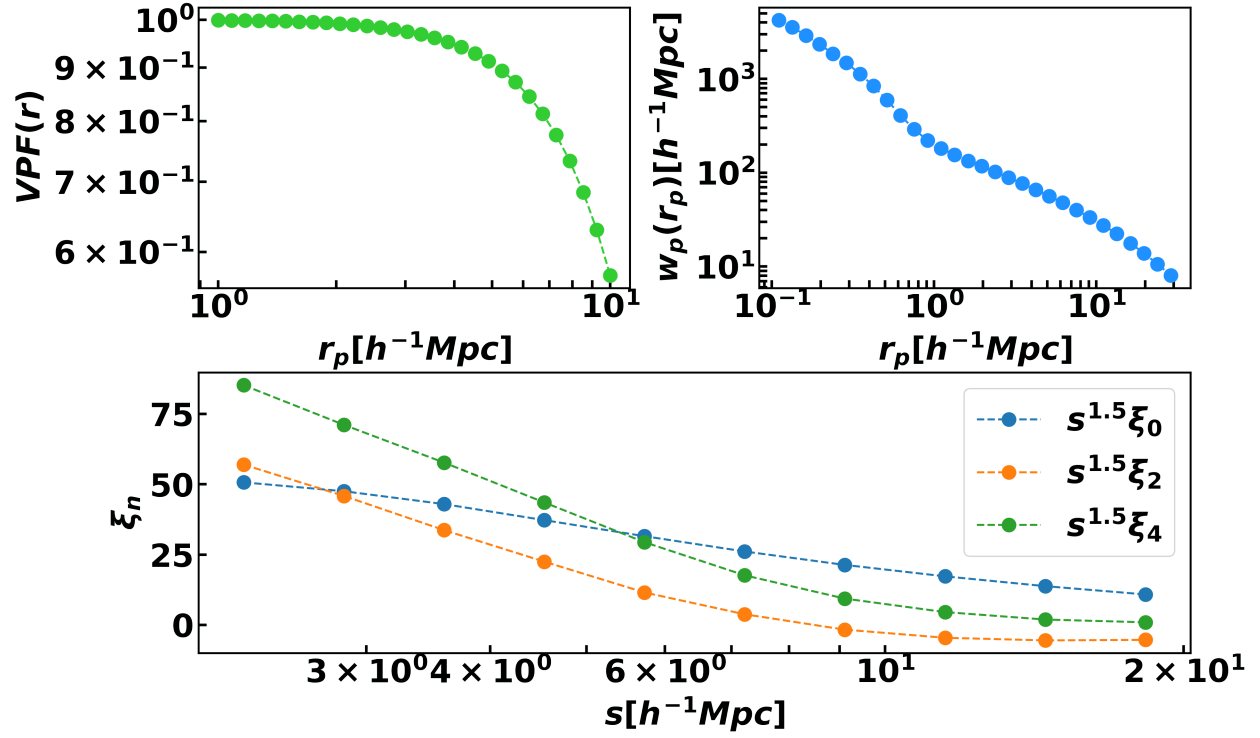


Figure 4.2: Small-scale clustering measurements at the fiducial value from fitting the HOD model at $z = 1.1$. Top left: Void probability function as a function of r ; top right: projected correlation function as a function of r_p ; bottom: correlation function multipoles as a function of s .

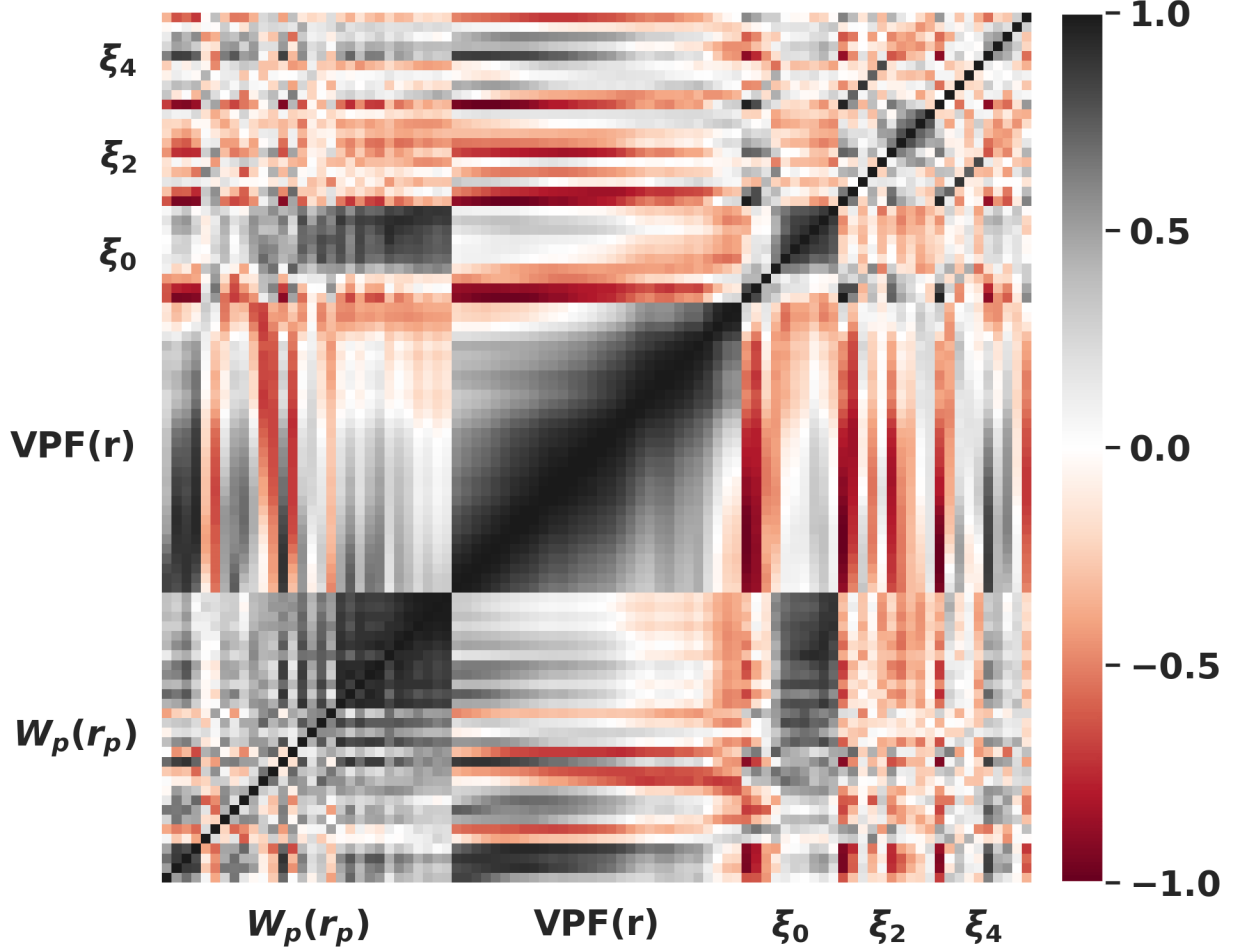


Figure 4.3: Jackknife covariance calculated at the fiducial value. The data vector is arranged from left to right as follows: bins 0–30 correspond to $w_p(r_p)$, bins 30–60 to $VPF(r)$, bins 60–70 to ξ_0 , bins 70–80 to ξ_2 , and bins 80–90 to ξ_4 .

4.4 Covariance

We use the jackknife resampling technique^{199–203} to estimate the covariance matrix of our measurements. We divide the simulation box into overlapping cuboidal sub-volumes of dimensions $200 h^{-1}$ Mpc along two axes and $2000 h^{-1}$ Mpc along the third axis. By rotating the longer axis among the x , y , and z directions, we create three configurations. In each configuration, we partition the two shorter axes into 10 segments of $200 h^{-1}$ Mpc, resulting in $10 \times 10 = 100$ sub-volumes per configuration and a total of 300 overlapping sub-volumes. We compute the clustering measurements described in Section 4.3, each time excluding one

sub-volume, generating 300 jackknife realizations.

We arrange the measurements into a data vector in the following order: w_p measurements, followed by VPF, ξ_0 , ξ_2 , and ξ_4 . Within each measurement type, data points are ordered from smallest to largest scales. The jackknife estimate of the covariance matrix is then computed using

$$C_{ij} = \frac{1}{300} \sum_{k=1}^{300} (X_i^k - \bar{X}_i)(X_j^k - \bar{X}_j), \quad (4.9)$$

$$\bar{X}_i = \frac{1}{300} \sum_{k=1}^{300} X_i^k, \quad (4.10)$$

where X_i^k denotes the i^{th} element of the measurement vector from the k^{th} jackknife realization.

The reduced covariance matrix, defined as

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}, \quad (4.11)$$

is shown in Figure 4.3. The w_p and VPF measurements between neighboring bins are strongly correlated, as are the ξ_ℓ measurements with different ℓ but similar separations. The VPF is not correlated with other clustering measures, aligning with theoretical expectations.

These covariance matrices are not intended for robust cosmological analysis. Our goal is to study the convergence properties of machine learning-based interpolation methods, and for this purpose, any reasonable positive-definite matrix that approximates a typical covariance matrix of clustering measurements suffices.

4.5 Machine Learning Methods

We employed supervised machine learning (ML) regression for our analysis. A more detailed discussion of ML methods is provided in Appendix 4.A. Our goal is to predict measures of small-scale clustering as functions of Halo Occupation Distribution (HOD) parameters.

The input features are the five HOD parameters, and the target variables are the projected two-point correlation function w_p , the void probability function (VPF), and the correlation function multipoles ξ_ℓ at different scales. We followed these steps:

1. Check the distribution of the data.
2. Preprocess the data.
3. Split the data into training and testing sets.
4. Fit different ML models to the training data.
5. Perform hyperparameter tuning.
6. Evaluate model accuracy.
7. Select the best model.

4.5.1 Check the Distribution of the Data

The distributions of the HOD parameters for both the full and small boxes were initially uniform. However, due to the filtering method applied to the full box, the M_{cut} parameter exhibits a nearly Gaussian distribution, as shown in Figure 4.4. We suspect that some inconsistencies in the full box may arise from this filtering (see Section 4.6.4).

4.5.2 Data Preprocessing

We used `Scikit-Learn`^{iv 204} for all ML-related computations. To improve the model’s performance, we increased the number of input features by using the *PolynomialFeatures* module, followed by the *StandardScaler* module from Scikit-Learn. For the small box, we chose `PolynomialFeatures(degree=2, interaction_only=True)`, resulting in a total of 16 input features (discussed in Appendix 4.B). The purpose of using *StandardScaler* is to center the data by subtracting the mean and scaling it to have unit variance. This

^{iv}<https://scikit-learn.org/stable/index.html>

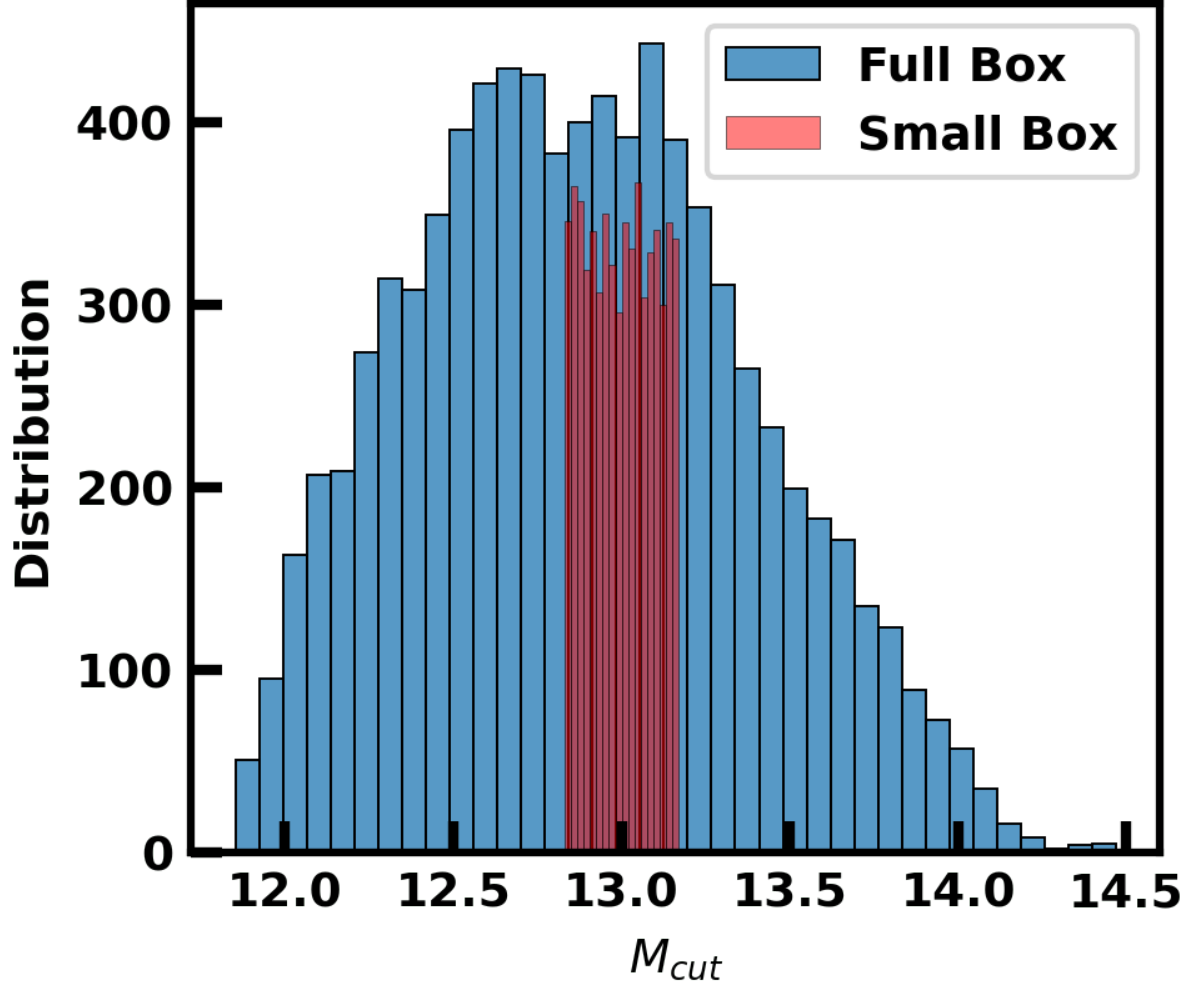


Figure 4.4: M_{cut} distribution on full and small boxes (see Section 4.2.3). HOD parameters are drawn uniformly for both the full box and small box, but after applying filtering to the full box, the distribution becomes nearly Gaussian.

preprocessing step is crucial for standardizing the features and ensuring that they contribute equally to the learning process, especially when features have different scales or variances.

The target variables, representing the small-scale clustering statistics, are highly skewed, which can introduce bias in model training. To mitigate this issue, we applied several common data transformation methods to the raw data, including log transformation (LT), standard scaling (SC), Yeo-Johnson (YJ), and Box-Cox (BC). These transformations help reduce biases, and their importance is further discussed in Section [4.6.1](#).

4.5.3 Split the Data into Training and Testing Sets

We split the data into two sets: a training set and a test set, with the test set typically representing 20% of the total data. We trained the model using the training set and evaluated its accuracy on the test set. In our study, we used a training set of 6,660 samples and a test set of 1,666 samples for the full box. For the small box, we varied the sample size to investigate its effect on the model’s performance (see Section [4.6.3](#)), with training set sizes of 800, 1,600, and 5,400, and a test set size of 600.

4.5.4 Fit Different ML Models to the Training Data

To achieve better accuracy and robustness, we evaluated several regression models. We tested the performance of three models: random forest (RF) regression, Bayesian ridge (BR) regression, and artificial neural networks (ANN). Each method is discussed in detail in Appendix [4.A](#).

We found that ANN performed the best among the three methods. A more detailed comparison is provided in Section [4.6.2](#).

4.5.5 Hyperparameter Tuning

Hyperparameter tuning is a crucial step in machine learning, as it can regularize a model to prevent overfitting or underfitting. Using the Scikit-Learn library, we employed specific hyperparameter settings for the three ML models.

Small Box Hyperparameters for Bayesian Ridge Regressor					
Statistics	w_p	VPF	ξ_0	ξ_2	ξ_4
lambda_init	0.01	0.1	0.1	0.1	0.1

Table 4.2: Hyperparameters for BR on Small Box

Small Box Hyperparameters for Neural Network					
Statistics	w_p	VPF	ξ_0	ξ_2	ξ_4
hidden_layer_sizes	500	500	500	500	500
alpha	0.01	0.01	0.01	0.01	0.01
activation	tanh	tanh	tanh	tanh	tanh
learning_rate_init	0.0005	0.0005	0.0005	0.0005	0.0005
learning_rate	constant	constant	constant	constant	constant
solver	adam	adam	adam	adam	adam
random_state	1	1	1	1	1
validation_fraction	0.1	0.1	0.2	0.2	0.2
early_stopping	True	True	True	True	True
max_iter	2000	5000	2000	2000	2000

Table 4.3: Hyperparameters for ANN on Small Box

For the small box, we used the *BayesianRidge* module for the BR model, with hyperparameters given in Table 4.2; other parameters were kept at their default values. For the ANN, we used the *MLPRegressor* module, with hyperparameters listed in Table 4.3. We only tested the RF model on w_p using the *RandomForestRegressor* module with `n_estimators=200`, `criterion='squared_error'`, `bootstrap=False`, `random_state=0`, `warm_start=True`, while keeping other parameters at their default values. Although RF is a popular method and has been used in various cosmological problems^{132;205–207}, it is not suitable for our problem because it relies on a series of decision rules based on individual features, which may not capture the full complexity of the data.

4.5.6 Perform Accuracy Test

Several performance measures are employed to assess the efficacy of an ML model, such as the root mean squared error (RMSE), mean absolute error (MAE), and adjusted R^2 score (see Appendix 4.B). Since each bin in our clustering statistics has a different range, we utilize modified versions of RMSE and MAE, namely the average percent RMSE and average percent MAE, respectively.

The average percent RMSE is defined as:

$$(\%) \overline{RMSE} = \frac{1}{n_{bins}} \sum_{i=1}^{n_{bins}} \frac{RMSE_i}{\left(\frac{|y_{max}^i| + |y_{min}^i|}{2} \right)} \times 100. \quad (4.12)$$

The average percent MAE is defined as:

$$(\%) \overline{MAE} = \frac{1}{n_{bins}} \sum_{i=1}^{n_{bins}} \frac{MAE_i}{\left(\frac{|y_{max}^i| + |y_{min}^i|}{2} \right)} \times 100, \quad (4.13)$$

where n_{bins} is the number of bins, and for the i^{th} bin, y_{max}^i and y_{min}^i represent the maximum and minimum true values, respectively.

Similarly, the average adjusted R^2 (Adj $\overline{R^2}$) score is calculated by summing the adjusted

Model	(%) $\overline{RMSE}$	(%) $\overline{MAE}$	Adj $\overline{R^2}$	(%) RMSE Range	(%) MAE Range
RF	2.55	1.57	0.982	4.86–1.26	3.00–0.80
BR	1.24	0.94	0.995	1.78–0.78	1.29–0.65
ANN	0.63	0.48	0.999	1.11–0.35	0.72–0.27

Table 4.4: Comparison of RF, BR, and ANN regression model performance on w_p for the small box test set. We calculate the average percent RMSE, average percent MAE, and average adjusted R^2 score to compare model performance (see Section 4.5.6 for definitions).

R^2 scores and dividing by n_{bins} . We present results solely for the projected correlation function on the small box; however, the same analysis was performed on other statistics when selecting an optimal model.

In Table 4.4, we compare $(\%)\overline{RMSE}$, $(\%)\overline{MAE}$, and average adjusted $\overline{R^2}$ score for RF, BR, and ANN. The ANN outperforms RF and BR by a considerable margin.

Table 4.5 displays the best five and worst five bins for $(\%)\overline{RMSE}$. All models exhibit superior performance at larger projected separations (r_p).

The results for the full box and the small box are presented in Table 4.6. We do not include results for YJ and BC for the full box due to their inconsistency with sparse data. For all models, except BR on the full box, performance scores are acceptable (adjusted R^2 score is above 0.95 on the test set). Typically, this would be considered satisfactory performance; however, we will later demonstrate that even with such a high-performance score, the models result in strongly biased posterior likelihood contours.

Furthermore, we examined the cross-validation scores for all statistics to assess the uniformity of the data and concluded that the variance was reduced with data transformation.

4.5.7 Likelihood Analysis

We performed Markov Chain Monte Carlo (MCMC) sampling²⁰⁸ to obtain the posterior probability distribution of the parameter space. The likelihood function is $\mathcal{L} \propto \exp(-\chi^2/2)$,

Model	Bin No.	5 Best (%) RMSE	Bin No.	5 Worst (%) RMSE
RF	30	1.26	1	4.88
	29	1.26	2	4.84
	28	1.28	3	4.78
	27	1.31	4	4.72
	26	1.34	5	4.66
BR	30	0.78	9	1.78
	29	0.80	6	1.78
	28	0.81	7	1.77
	26	0.82	5	1.76
	27	0.84	1	1.73
ANN	28	0.35	3	1.11
	29	0.36	5	1.02
	27	0.36	1	1.00
	30	0.38	2	1.00
	25	0.38	4	0.98

Table 4.5: Performance for the 5 best and worst outputs of w_p on Box-Cox transformation for the small box test set. We compare RF, BR, and ANN models to report the percent RMSE.

Adjusted R^2	Small Box			Full Box	
	RF	BR	ANN	BR	ANN
Unscaled	0.966	0.990	0.970	0.885	0.980
Standard Scaler	0.982	0.990	0.998	0.885	0.994
Yeo-Johnson	0.983	0.995	0.999	—	—
Box-Cox	0.982	0.995	0.999	—	—

Table 4.6: Adjusted R^2 score of w_p on various transformations for the small box test set and full box test set. We compare RF, BR, and ANN models to report the adjusted R^2 score.

where χ^2 is given by

$$\chi^2 = (x - \mu)^T C^{-1} (x - \mu). \quad (4.14)$$

In our analysis, x denotes the machine learning model predictions (the ‘theory’), while μ represents the actual simulation measurements (the ‘observations’). Specifically, we applied this framework to w_p , VPF, and ξ_ℓ , as well as their combinations (see Section 4.7). Here, $(x - \mu)$ is the difference between theory and observations, and C^{-1} is the inverse of the covariance matrix. We used `emcee`^v to estimate the posterior distribution of the model parameters. Several combinations of statistics were tested using 250 walkers, with the first 200 steps discarded as burn-in. Given the model’s efficiency, we achieved convergence within a minute. Our findings are discussed in detail in the next section.

4.6 Constraining HOD parameters

The ultimate objective of ML-based emulators is to be able to run MCMC chains on measured data to constrain HOD parameters. While all our approaches yield acceptable performance metrics, we will demonstrate that unless considerable care is exercised, they result in inaccurate likelihood surfaces. The cumulative effect of small inaccuracies significantly biases the best-fit values, and the uncertainties in derived model parameters are often significantly underestimated. This holds even when we constrain ourselves to the small box where the HOD parameter range is narrow. We will present results for the w_p , but the general conclusions are similar for all other statistics. Despite our model’s ability to accurately predict the full projected correlation function, we only selected $r_p = 1.77h^{-1}$ Mpc to $r_p = 17.78h^{-1}$ Mpc (bin 15 - bin 27) for our MCMC run. This was done to maintain consistency with the ξ_ℓ range ($s = 2h^{-1}$ Mpc – $20h^{-1}$ Mpc), as we will demonstrate how HOD constraints improve with the combination of clustering statistics (see Section 4.7).

^v<https://emcee.readthedocs.io/en/stable/>

4.6.1 Effect of Data Transformations

We find that data transformations are essential. Without them, the likelihood contours are strongly biased. Figure 4.5 shows the distribution of the inaccuracy in the ANN-based emulator prediction at different HOD parameter values (left panel), and the resulting one and two σ contours in the σ vs $\log M_1$ plane. We compute this for unscaled data and for three different data transformation algorithms: SC, BC, and, YJ. Without applying any transformations, the likelihood contours are strongly biased. Applying a data transformation corrects the bias. Despite the standard scalar transformation exhibiting an off-center error distribution, we find that all three approaches perform equally well when it comes to the likelihood contours. The bias in the computed likelihood function is much larger than what the left panel of Figure 4.5 would suggest. The error in the density plots, represented as $Error/\sigma$, is calculated as the difference between the predicted values from the machine learning models and the true values obtained from the simulation, normalized by the standard deviation (σ) derived from the jackknife covariance matrix. Specifically, for each predicted value y_i , the error is computed as $Error_i = y_i^{\text{pred}} - y_i^{\text{true}}$, where σ_i is the jackknife estimate of the standard deviation for the i^{th} data point. This normalization allows for a clearer interpretation of the prediction errors by accounting for the intrinsic variability in the data. All the Error calculations shown in this work are calculated on w_p . However, we found similar results for the other measurements (VPF , ξ_ℓ) as well.

Figure 4.6 shows similar computation for the BR method. The data scaling is essential here as well. Unlike the ANN, however, none of the transformations can produce unbiased likelihood contours, even though the YJ and the BC transformations lead to consistent results. Here as well, the bias in the likelihood contours is much more visible than what the error distribution plot would suggest.

4.6.2 Algorithm Choice

Figure 4.7 shows the error distribution and the HOD parameter likelihood contours for all three ML algorithms considered in this work. In all three cases, the BC transformation was

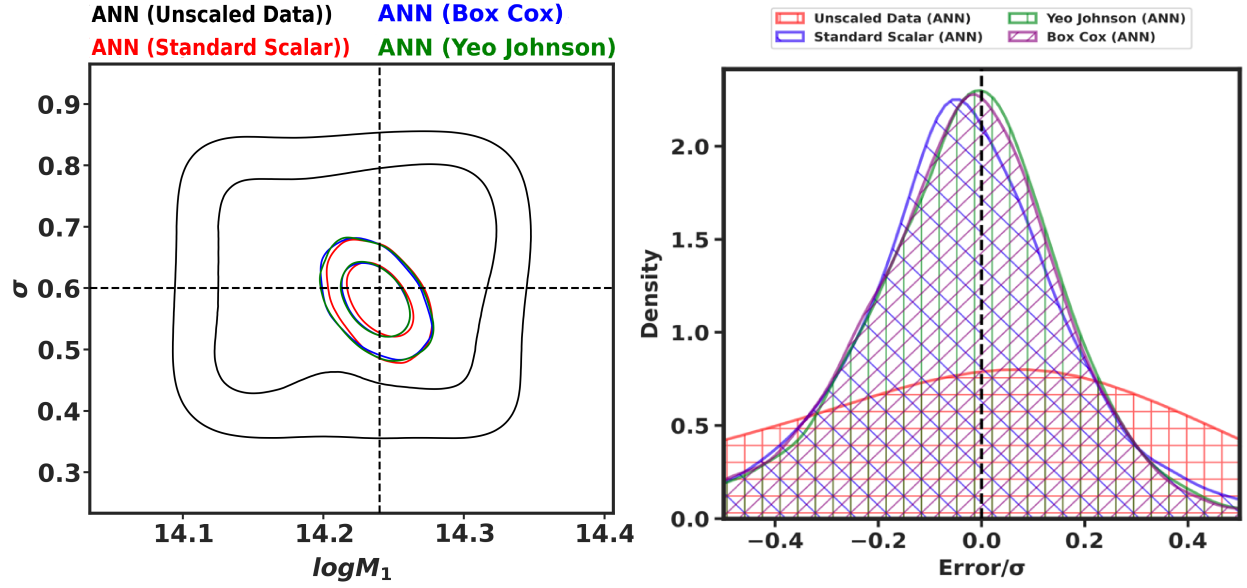


Figure 4.5: Exploring the impact of data transformation on ANN Method within the small Box. Left: Equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . The plot shows the convergence between transformed and non-transformed data. Right: Density plot between Transformed vs. Non-transformed Data as a function of Error/σ . The σ is calculated from the jackknife covariance matrix.

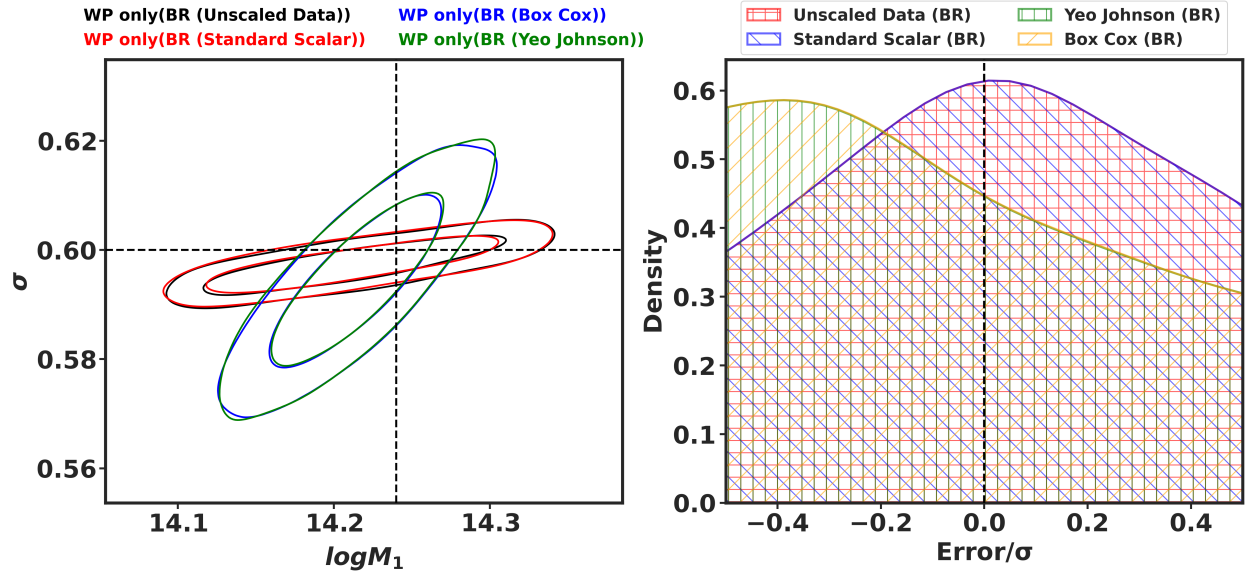


Figure 4.6: Exploring the impact of data transformation on BR Method within the small Box. Left: Equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . The plot shows the convergence between transformed and non-transformed data. Right: Density plot between Transformed vs. Non-transformed Data as a function of Error/σ . The σ is calculated from the jackknife covariance matrix.

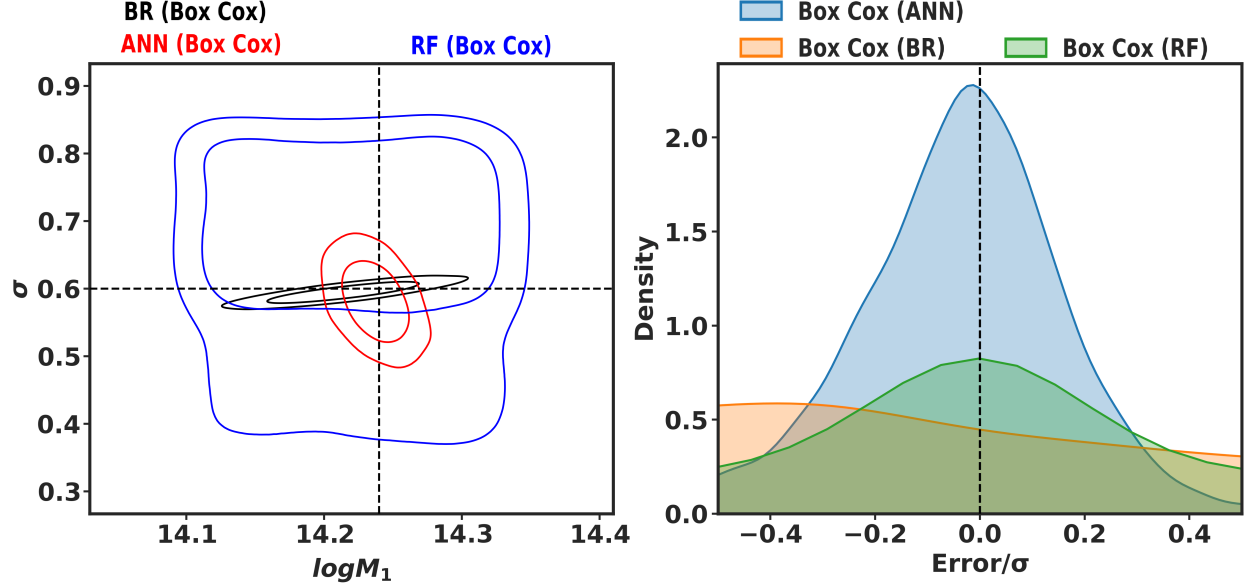


Figure 4.7: Exploring the impact of different ML methods within the small box. Left: Equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . The plot shows the convergence test between the RF, BR, and ANN models. Right: Density plot between the RF, BR, and ANN models as a function of Error/σ . The σ is calculated from the jackknife covariance matrix.

applied to the data. The right panel on Figure 4.7 shows that the ANN-based method is the least biased. The BR and RF methods result in significantly wider likelihood contours compared to the ANN, which is the one closest to the truth. In the case of RF, the likelihood contours are completely dominated by the prior parameters.

4.6.3 Number of Samples

We anticipate that all methods will eventually converge toward the truth as the size of the training sample increases. On the other hand, an excessively large training sample would defeat the primary purpose of the emulator, which is to obtain likelihoods with significantly less computational time.

Figure 4.8 shows the effect of increasing sample size on our results for the BR method. We see that increasing the sample size by a factor of two and then a factor of seven does not have a significant effect on the performance of the algorithm.

For the ANN algorithm, on the other hand, the improvement in the size of the train-

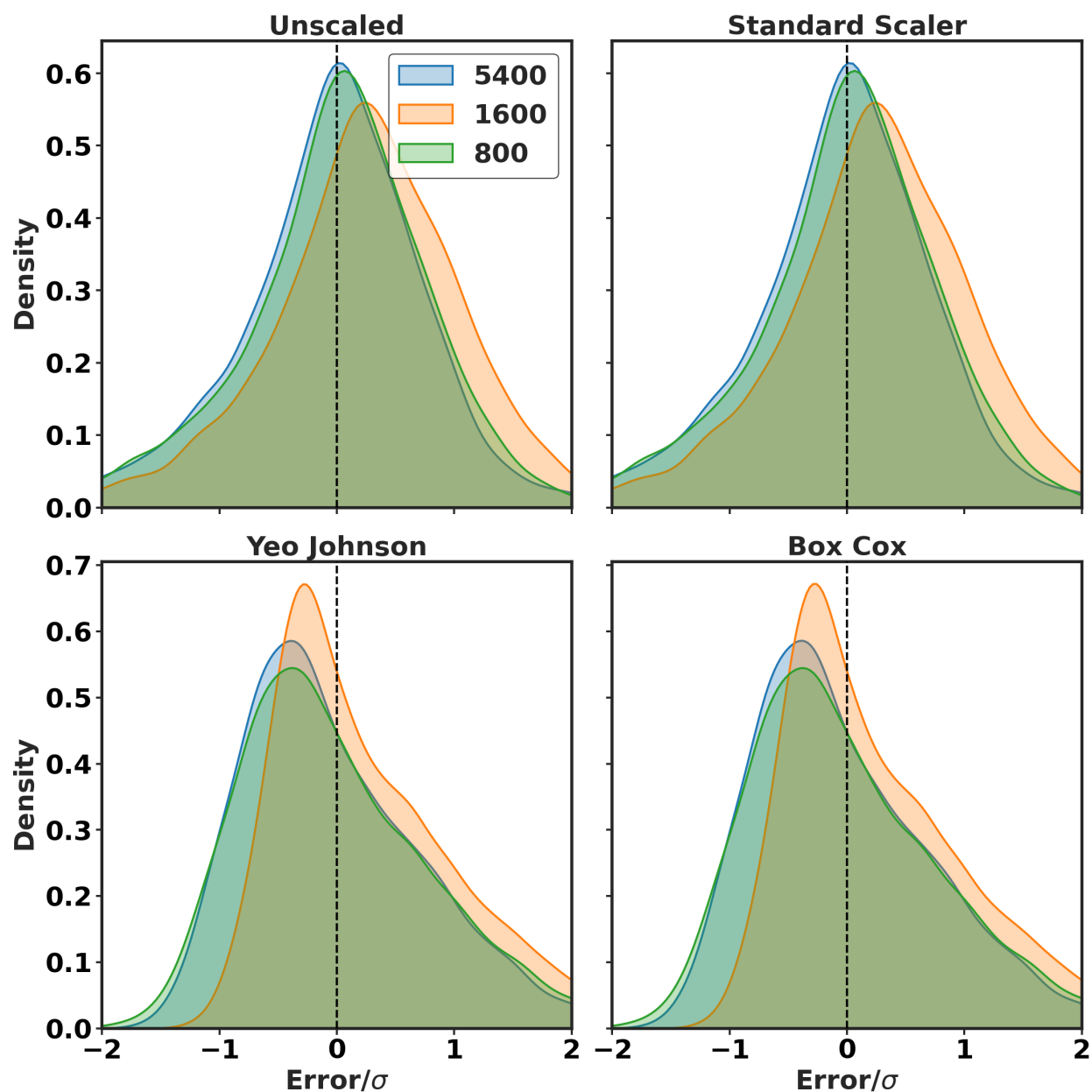


Figure 4.8: Effect of increasing data points on Bayesian Ridge Regressor for the small box. This plot compares the density plot between the training set of 800, 1600, and 5400 for unscaled, SC, YJ, and BC transformation as a function of Error/σ . The σ is calculated from the jackknife covariance matrix.

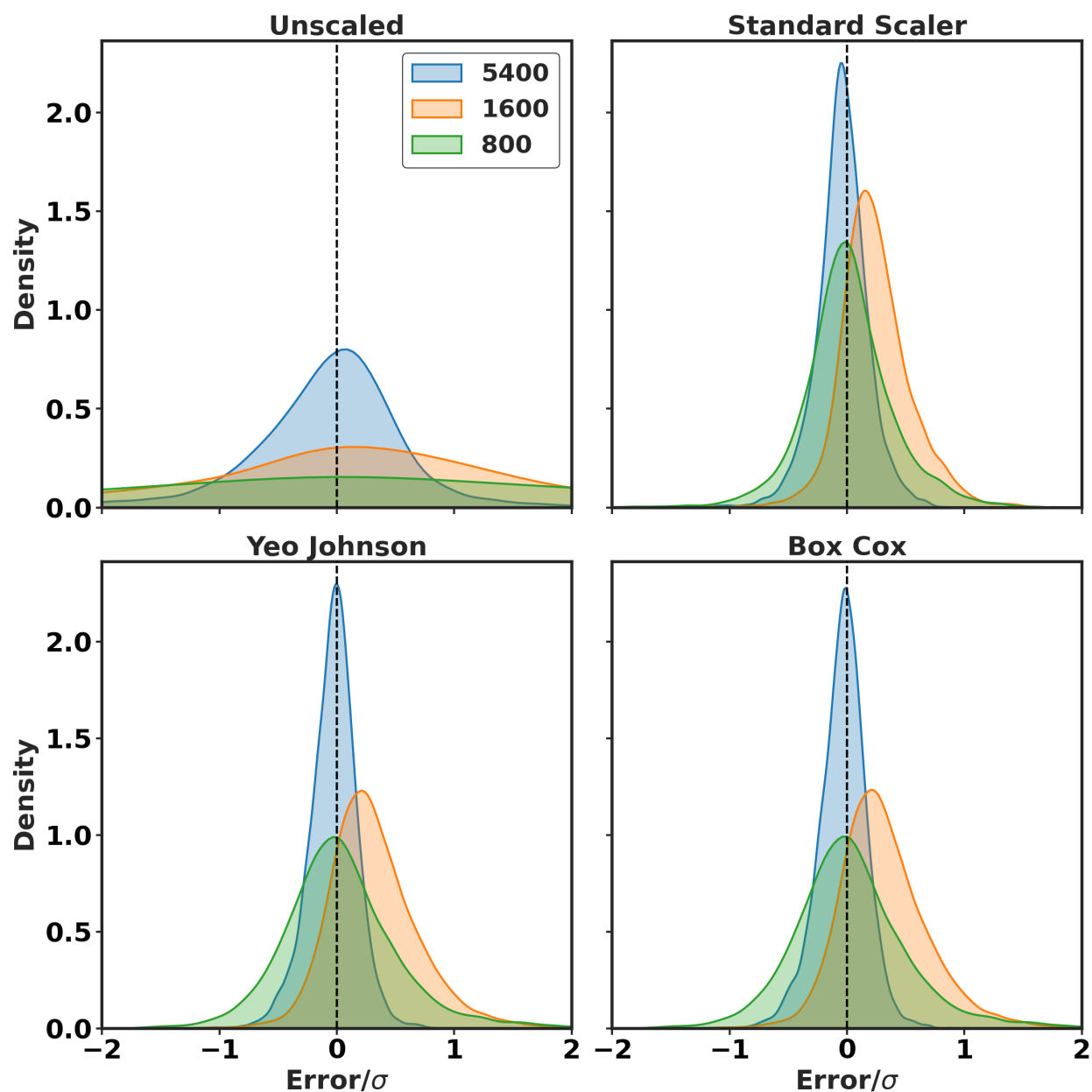


Figure 4.9: Effect of increasing data points on ANN for the small box. This plot compares the density plot between the training set of 800, 1600, and 5400 for unscaled, SC, YJ, and BC transformation as a function of Error/σ . The σ is calculated from the jackknife covariance matrix.

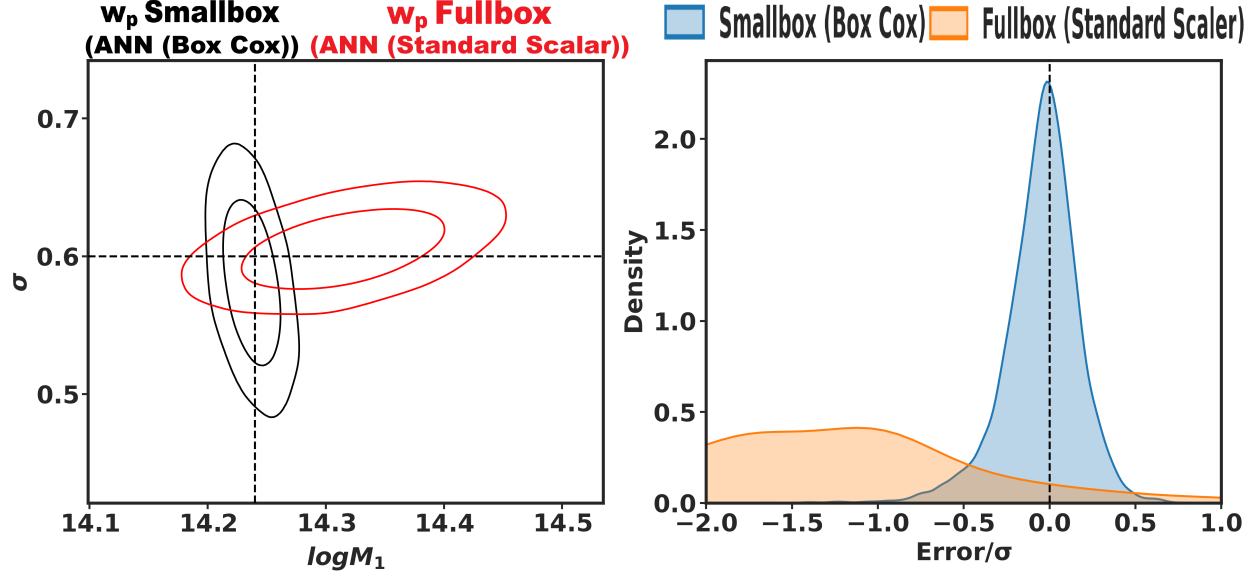


Figure 4.10: Error Comparison Between full Box and small box. Left: Equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . This plot shows the convergence between different box sizes. Right: Density plot between small box and full box as a function of Error/σ . The σ is calculated from the jackknife covariance matrix.

ing data is clearly visible. The four panels in Figure 4.9 shows the evolution of the error distribution as the sample size first doubles and then increases by a factor of seven. The improvement in this distribution, however, is not as drastic as one would hope.

These findings suggest that for the BR method, increasing the sample size does not yield significant improvement, while for ANN, increasing the sample size can improve performance.

4.6.4 Prior Parameter Ranges

Figure 4.10 shows the effect of increasing the prior range of HOD parameter space that the ML algorithm has to deal with. The black and red contours on the left panel correspond to the small box and full box respectively as defined in Section 4.2. The contours were derived with the ANN algorithm using the BC data transformation for the small box and SC for the full box, the method that was performing the best on a smaller box as described in Section 4.6.2. From Figure 4.10 it is clear that even the best-performing algorithm fails to accurately reproduce the likelihood surfaces when the parameter range gets too large. In the limit of a very large number of training samples, we still expect the ANN to reproduce the

Observables	1 σ error				
	Small Box				
	$\log M_{\text{cut}}$	$\log M_1$	σ	κ	α
w_p	$12.9988^{+0.011}_{-0.0111}$	$14.2414^{+0.0099}_{-0.0103}$	$0.5955^{+0.0136}_{-0.0138}$	$0.9859^{+0.023}_{-0.0226}$	$0.4005^{+0.0101}_{-0.0106}$
$w_p + VPF$	$12.9989^{+0.0103}_{-0.011}$	$14.2409^{+0.0097}_{-0.0096}$	$0.5962^{+0.0125}_{-0.013}$	$0.9856^{+0.0206}_{-0.0214}$	$0.4001^{+0.0088}_{-0.0091}$
$w_p + \xi_0$	$12.9984^{+0.0069}_{-0.007}$	$14.2416^{+0.008}_{-0.008}$	$0.5954^{+0.0121}_{-0.0121}$	$0.9848^{+0.0194}_{-0.0191}$	$0.4007^{+0.0089}_{-0.0086}$
$w_p + \xi_0 + \xi_2$	$12.9988^{+0.0064}_{-0.0066}$	$14.2416^{+0.0075}_{-0.0073}$	$0.5956^{+0.0116}_{-0.0117}$	$0.985^{+0.0182}_{-0.0185}$	$0.4001^{+0.0085}_{-0.0085}$
$w_p + \xi_0 + \xi_2 + \xi_4$	$12.9983^{+0.0058}_{-0.0059}$	$14.2411^{+0.0069}_{-0.0073}$	$0.5953^{+0.0113}_{-0.0113}$	$0.9841^{+0.0179}_{-0.0178}$	$0.3998^{+0.0081}_{-0.0081}$
ξ_0	$12.9975^{+0.0132}_{-0.0135}$	$14.2426^{+0.0113}_{-0.0113}$	$0.5941^{+0.0326}_{-0.0337}$	$0.9825^{+0.0583}_{-0.0558}$	$0.4005^{+0.0267}_{-0.0262}$
$\xi_0 + \xi_2$	$12.9986^{+0.0117}_{-0.0122}$	$14.2418^{+0.0101}_{-0.0099}$	$0.5965^{+0.0276}_{-0.0263}$	$0.9769^{+0.0508}_{-0.051}$	$0.3984^{+0.0224}_{-0.0222}$
$\xi_0 + \xi_2 + \xi_4$	$12.9986^{+0.0111}_{-0.0113}$	$14.2419^{+0.009}_{-0.0096}$	$0.5984^{+0.0267}_{-0.0273}$	$0.9707^{+0.0433}_{-0.0424}$	$0.396^{+0.0177}_{-0.0178}$
$w_p + VPF + \xi_0 + \xi_2 + \xi_4$	$12.9984^{+0.0056}_{-0.0056}$	$14.2409^{+0.0067}_{-0.0068}$	$0.596^{+0.0104}_{-0.0105}$	$0.9826^{+0.0158}_{-0.0166}$	$0.3999^{+0.0072}_{-0.0074}$

Table 4.7: 1 σ error for the small box. The values are calculated by running the MCMC chain and using the ANN models for w_p , VPF, ξ_0 , ξ_2 , and ξ_4 .

true likelihood faithfully. This would, however, require a very large training sample and the computational cost of training the network will start approaching that of a straightforward MCMC chain.

4.7 Results and Conclusions

Small-scale clustering of galaxies potentially contains useful information about cosmological parameters. To harness this information, however, one needs to first accurately model the nonlinear process of galaxy formation. The most important aspect of this, for the clustering measurements relevant to cosmology, is determining which dark matter halos certain types of galaxies occupy and exactly how they are distributed within these halos. The HOD-

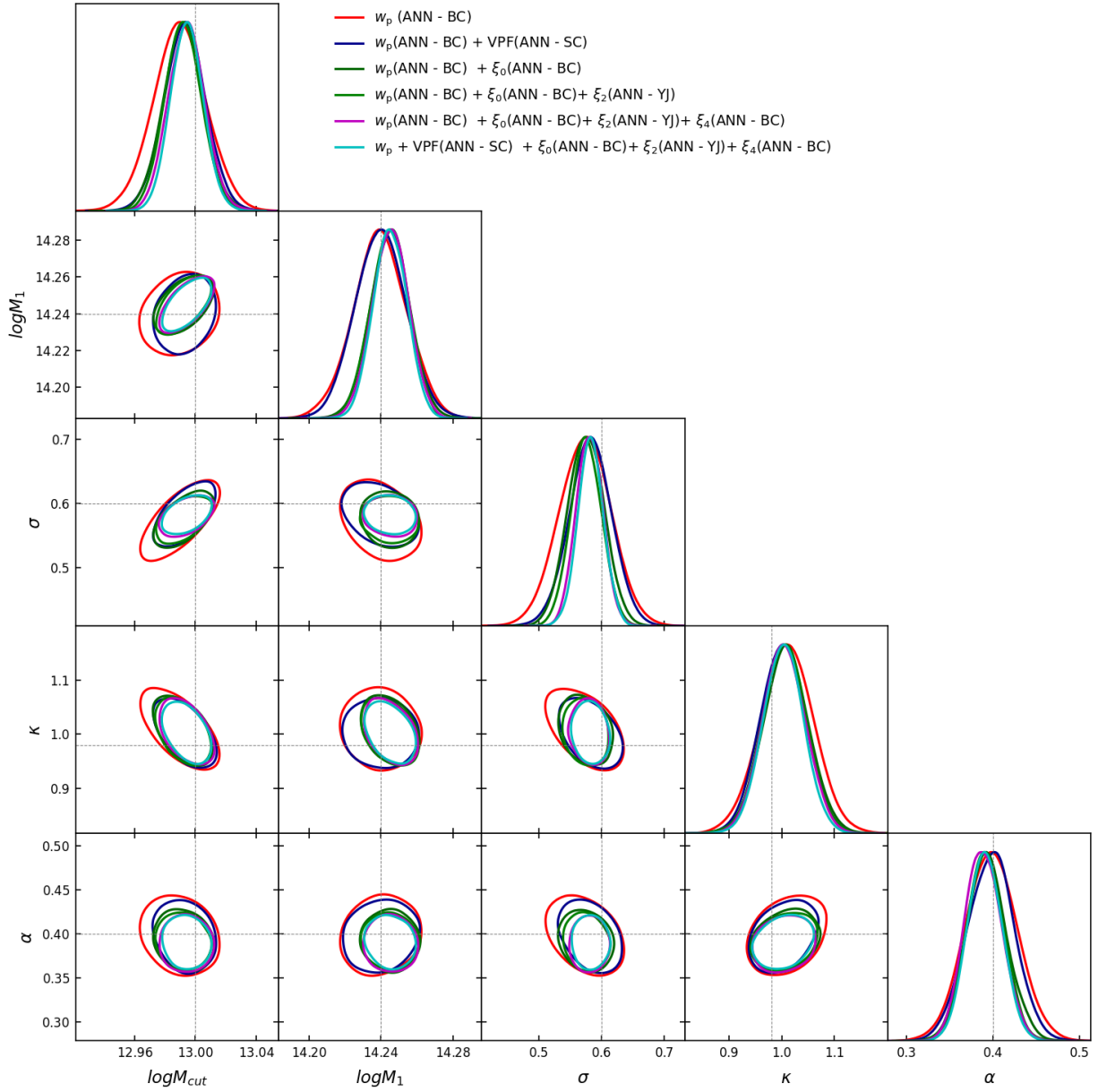


Figure 4.11: Equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$. The ANN models are trained on the small box for w_p , VPF, ξ_0 , ξ_2 , and ξ_4 . The dashed lines denote the expected values of the HOD parameters. Contours are generated using the best-performing combinations of the ML models.

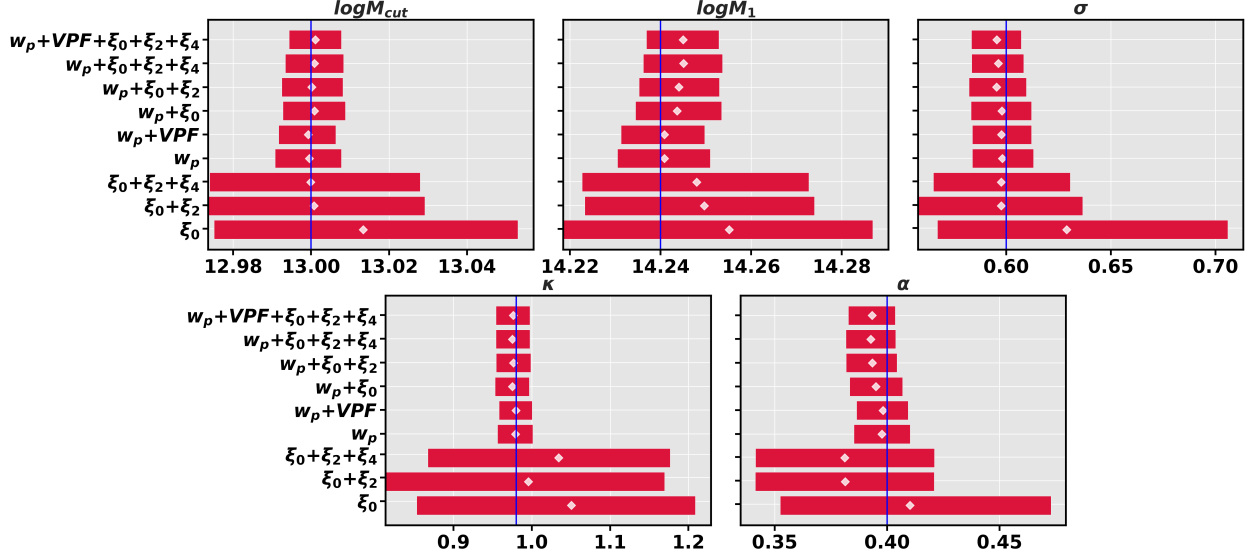


Figure 4.12: Effect of 1σ constraint on HOD parameters for a different combination of statistics. The blue lines denote the expected values of the HOD parameters. The white marks at the center of the red bars represent the mean of the MCMC samples after discarding 200 steps as burn-in.

based family of models is one of the most widely used methods for modeling this galaxy-halo connection. Since generating HOD model predictions requires computationally intensive calculations from N-body simulations, machine learning (ML)-based approaches that can significantly accelerate this computation have been increasingly used.

We found that great care must be taken when modeling the dependence of small-scale clustering on HOD parameters. Transforming the input data, the choice of ML approach, and the size of the HOD space under exploration have a significant impact on the robustness of the predictions. We found that, very often, ML methods report acceptable formal performance scores, but when inaccuracies in ML modeling are propagated to quantities important for cosmology, these small inaccuracies lead to unacceptably large biases. For example, the posterior likelihood in the HOD space can be strongly biased even when the ML pipeline reports a good performance score.

For the types of galaxy tracers and clustering measurements we studied, we found that artificial neural networks (ANN) provided the best results among the models tested. However, even for the ANN, the results were robust only when the HOD space was confined

to the small box with a relatively narrow range of allowed parameters; the method failed when we considered larger parameter spaces. The selection of the small box range is largely driven by M_{cut} , as this parameter is particularly sensitive to the filtering method we employed. In contrast, the other HOD parameters could potentially span larger ranges, as their distributions are less affected by this filtering process. The required restriction on the HOD parameter space was significant—much smaller than typical HOD parameter bounds allowed by currently available data. The ANN results become more robust with a larger set of training data as expected, but we found that the convergence to the true result with increasing training data size is not as fast as one would hope. The other two ML-based methods that we tried were Bayesian Ridge and Random Forest. These methods resulted in very large biases in the likelihood values even for the small box restriction.

Once the ML models achieve the desired robustness, the constraints on HOD parameters can be obtained with very little computational time. We checked how different combinations of w_p , VPF, and ξ_ℓ enhanced the constraining power of small-scale clustering on the HOD parameter space. The results of this exercise are shown in Table 4.7, and the 1σ posterior contours are shown in Figure 4.11. The improvement in the HOD parameter constraints is shown in Figure 4.12. We find that the addition of VPF to w_p resulted in a slight improvement, but the combination of $w_p + \xi_0$ worked better than $w_p + \text{VPF}$. These results are qualitatively in agreement with previous findings, but we want to warn the reader to treat the actual numbers (error bars, etc.) with caution. They should not be interpreted as actual forecasts for this type of analysis. The actual errors are very likely underestimated due to the usage of Jackknife covariance matrices and the prior effects from the small box restriction of allowable HOD space.

Acknowledgement

This work is supported by DOE grants DE-SC0021165 and DE-SC0011840.

This research used resources from 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.
AJ thanks Andrew Hearin and Hanyu Zhang for their useful discussions.

Data Availability

The data products related to this study will be shared upon request. The AbacusSummit simulations are publicly available to download from <https://abacusnbody.org/>

Appendix

4.A Machine Learning Overview

We used supervised machine-learning techniques for regression to model the dependence of small-scale clustering measurements on HOD parameters. Specifically, we tested and compared three different regression algorithms - ANN, RF, and BR regression.

4.A.1 Artificial Neural Networks

Artificial neural networks (ANNs) are computing systems inspired by biological neural networks, designed to identify complex patterns in data. They consist of an input layer, one or more hidden layers, and an output layer, each containing neuron-like processing units connected by synaptic weights²⁰⁹ By tuning these weights through backpropagation, ANNs can approximate complex functions.^{210;211}

We implement ANNs using MLPRegressor in Scikit-Learn, a fully-connected feedforward network. The HOD parameters were input features and clustering measurements were the output. Using Hyperparameter tuning we determine the optimal parameters like the number of layers/nodes, activation function, solver, and regularization.²¹² Activation functions like “ReLU” and “tanh” were considered. To prevent overfitting, we used early stopping based on a validation set and dropout regularization.^{213;214}

4.A.2 Random Forest Regression

Random forest (RF) aggregates predictions from decision trees trained on randomized data subsets.²¹⁵ This introduction of randomness improves generalization and reduces overfitting.²¹⁶ RF regression averages predictions over all trees:

$$\hat{y}_{\text{RF}} = \frac{1}{B} \sum_{b=1}^B \hat{y}_b \quad (4.15)$$

where $\hat{y}_b$ is the b^{th} tree's prediction. We use RandomForestRegressor in Scikit-Learn with hyperparameter tuning.²¹⁷ It is known for its scalability, ability to handle high-dimensional data, and resistance to noise and outliers. While capable of handling large datasets, RF may become computationally expensive as the number of trees or features increases.

4.A.3 Ridge Regression (BR)

Ridge regression regularizes linear regression by penalizing the residual sum of squares:

$$L(\beta) = \sum_{i=1}^n (y_i - \mathbf{x}_i\beta)^2 + \lambda|\beta|^2 \quad (4.16)$$

The λ hyperparameter controls regularization strength. This shrinks coefficients to prevent overfitting. We used BayesianRidge in Scikit-Learn, which estimates the regularization hyperparameter (λ) in a Bayesian framework.²¹⁸ Ridge regression is similar to LASSO but uses L2 regularization.

Ridge regression is commonly fitted using coordinate descent optimization algorithms. This involves iteratively minimizing the objective function along each coordinate direction while holding other coefficients fixed. The coordinate direction that leads to the largest improvement in the objective is selected in each iteration. This continues until convergence when changes along any coordinate cease to decrease the objective further.

4.B Evaluating Machine Learning Model

We evaluated our machine learning regression models using several numerical measures that quantify the difference between predicted ($\hat{y}_i$) and actual (y_i) values. These accuracy metrics provided insight into model performance and aided in selecting the best approach.

4.B.1 Mean Squared Error

The mean squared error (MSE) is computed by taking the average of the squared differences between the predicted and true values:

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (4.17)$$

where n is the number of samples.

Squaring the errors has the effect of heavily penalizing large deviations compared to smaller ones. However, this disproportionate penalty also makes MSE very sensitive to outliers that have large errors.²¹⁹ MSE ranges from 0 to ∞ , with lower values indicating better model fit.

4.B.2 Mean Absolute Error

The mean absolute error (MAE) calculates the average magnitude of errors without squaring

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (4.18)$$

Taking the absolute value of the differences avoids excessive penalties for large errors. This makes MAE more robust to outliers compared to MSE.^{220;221} MAE ranges from 0 to ∞ , with 0 being a perfect fit.

4.B.3 R-squared

The R-squared or coefficient of determination is:

$$R^2 = 1 - \frac{RSS}{TSS}, \quad (4.19)$$

where RSS is the residual sum of squares:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (4.20)$$

And TSS is the total sum of squares:

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2. \quad (4.21)$$

Here for i^{th} observation, y_i is the actual value, $\hat{y}_i$ is the predicted value and $\bar{y}$ represents the mean of the actual values. R^2 compares the accuracy of the model's predictions to the accuracy of simply predicting the mean in all cases. It provides an indication of the explanatory power of the model. R^2 values range from 0 to 1, with higher values indicating more variance explained by the model.

4.B.4 Adjusted R-Squared

The R-squared metric can be adjusted to account for the number of predictors in the model. The adjusted R-squared is defined as:

$$R_{adj}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1} \quad (4.22)$$

where n is the sample size and p is the number of explanatory variables in the model.

Unlike regular R^2 , the adjusted R^2 will penalize models for having too many unnecessary predictors. It will always be less than or equal to the unadjusted R^2 . Generally, adding predictors improves the basic R^2 but can worsen the adjusted R^2 if they are redundant. This helps prevent overfitting the training data.²²²

For comparing regression models, the adjusted R^2 provides a more conservative estimate that accounts for model complexity. However, it has limitations when the sample size is small or the number of predictors is large relative to the observations.

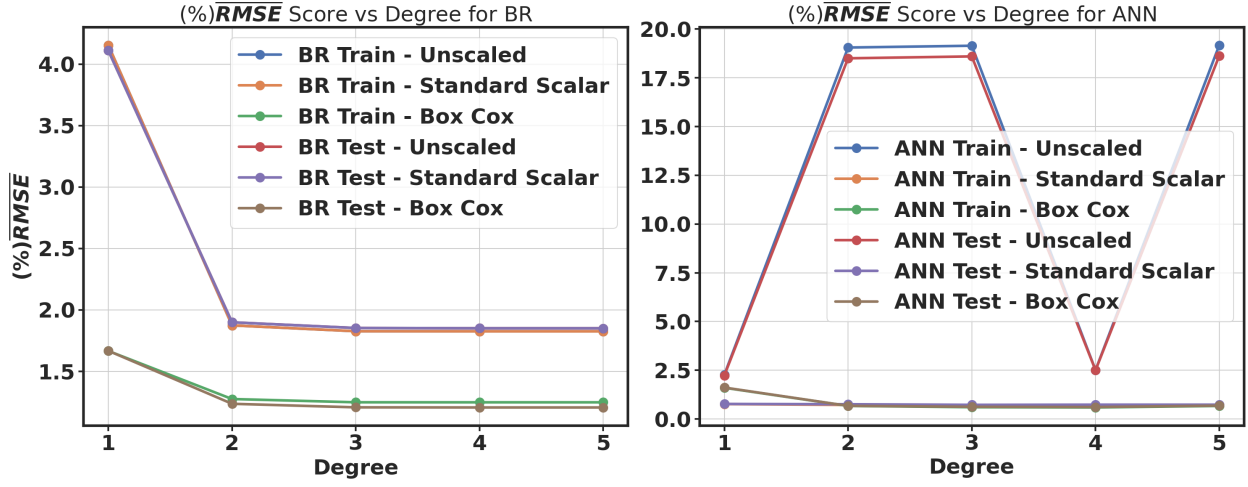


Figure 4.B.1: Comparison of average percent RMSE Scores between unscaled, SC, BC, and YJ on BR and ANN as a function of polynomial features on w_p for the small box train and test set. We calculate the average percent RMSE(see Section 4.5.6 for definition) as a function of degree polynomial. The left plot shows the average percent RMSE of the BR model. The left plot shows the average percent RMSE of the ANN model.

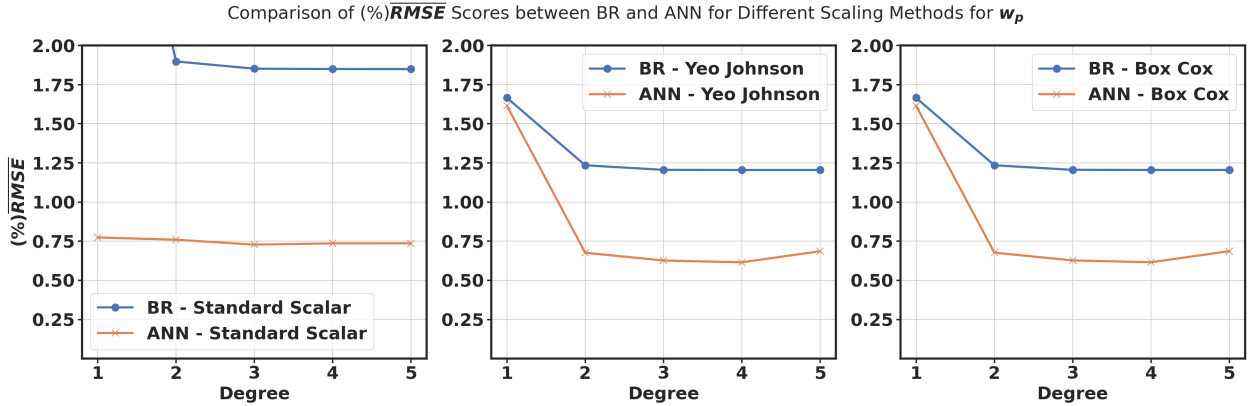


Figure 4.B.2: Comparison of average percent RMSE Scores between BR and ANN for SC, YJ, and BC on w_p test set. The left, middle, and right plot shows the average percent RMSE of SC, YJ, and BC transformations respectively.

4.B.5 Choice of Polynomial Features

Using polynomial features we can increase the number of features, capturing non-linear relationships in the data. Although a flexible model like ANN requires minimal preprocessing, incorporating polynomial features can reduce model complexity, leading to faster convergence. The selection of the polynomial degree is pivotal, as escalating it may result in model overfitting.

Figure 4.B.1 shows the relationship between $(\%) \overline{RMSE}$ and the degree of polynomial on BR and ANN for w_p on the small box. $(\%) \overline{RMSE}$ score decreases with increasing the model complexity but decreases slowly after 2^{nd} order. We opt for a degree of 2, given marginal performance improvements beyond. Interestingly, ANN exhibits inferior performance with raw target data.

Figure 4.B.2 illustrates the comparison between ANN and BR. This finding aligns with Table 4.4, demonstrating that ANN outperforms BR at the same degree of polynomial.

Chapter 5

Extended Analysis of Galaxy-Halo Connection with Machine Learning

5.1 Introduction

In this chapter, we provide additional details that complement the findings from the main analysis of Chapter 4. We will also explain in detail the N-body simulation used for the study. The focus is on the machine learning pipeline used for HOD parameters and further analysis of results from the full box simulations. We will elaborately discuss results from VPF and ξ_ℓ , which were not fully addressed in the previous chapter.

5.2 N-body Simulations

N-body simulations have been used to investigate various aspects of cosmology, including the impact of massive neutrinos on the matter power spectrum²²³, the anisotropic signal of higher-order statistics in galaxy surveys²²⁴, the dissociation of dark matter and gas in cosmic large-scale structure²²⁵, and the effects of decaying cold dark matter on structure formation²²⁶. These simulations have also been combined with machine learning techniques to classify dark matter particles and reconstruct the halo mass function²²⁷, and with hydro-

dynamical simulations to create emulators for predicting the Lyman-alpha forest flux power spectrum²²⁸.

There are several N-body simulations available that can be used to study the connection between galaxies and their dark matter halos in a cosmological context. These simulations follow the evolution of dark matter particles under the influence of gravity, providing a detailed picture of the growth of structure in the universe. Some notable examples include:

Dark Matter Only Simulations

1. Millennium Simulation: One of the largest and most widely used dark matter-only simulations, following over 10 billion particles in a $(500h^{-1} \text{ Mpc})^3$ volume.²²⁹
2. Aquarius Simulations: A suite of high-resolution zoom-in simulations of Milky Way-sized dark matter halos, designed to study the properties of substructure.²³⁰
3. Copernicus Complexio (CoCo) Simulations: High-resolution simulations comparing the small-scale structure in cold dark matter (CDM) and warm dark matter (WDM) cosmologies.²³¹
4. AbacusSummit: A suite of high-resolution dark matter only simulations designed for validating the DESI analysis pipelines, with a volume of $200h^{-3} \text{ Gpc}^3$ (10 times the DESI volume) for the base cosmology.²

Hydrodynamical Simulations

1. EAGLE Simulations: A suite of state-of-the-art hydrodynamical simulations that include detailed models for gas cooling, star formation, and feedback processes.³⁴
2. Illustris and IllustrisTNG: Large-scale hydrodynamical simulations that include a comprehensive model for galaxy formation physics.^{232;233}
3. APOSTLE Simulations: A set of zoom-in hydrodynamical simulations of Local Group-like environments, used to study the properties of satellite galaxies.²³⁴

4. Magneticum Simulations: A suite of hydrodynamical simulations that include the effects of magnetic fields and AGN feedback on galaxy cluster scales.²³⁵

The AbacusSummit simulations are particularly noteworthy for their large volume and high resolution, making them well-suited for validating the DESI analysis pipelines and studying the galaxy-halo connection on large scales. The suite covers a volume 10 times larger than the DESI survey, providing a robust statistical sample for cosmological studies. We used AbacusSummit simulations for our analysis which will be discussed in detail in the following section.

5.2.1 AbacusSummit Simulation Details

The AbacusSummit suite is a massive set of high-accuracy, high-resolution N-body simulations, totaling roughly 60 trillion particles. The core of the suite consists of 139 simulations with a particle mass of $2 \times 10^9 h^{-1} M_{\odot}$ in a box size of $2h^{-1}$ Gpc (Figure 5.2.1). The Abacus N-body code achieves a median fractional force error of $O(10^{-5})$ at impressive speeds, calculating 70M particle updates per second per node at early times and 45M particle updates per second per node at late times. The suite spans 97 cosmological models, including Planck 2018, previous flagship simulation cosmologies, and a linear derivative and cosmic emulator grid. Additionally, a sub-suite of 1883 boxes of size $500h^{-1}$ Mpc is available for covariance estimation. The AbacusSummit data products span 33 epochs from $z = 8$ to 0.1 and include lightcones, full particle snapshots, halo catalogs, and particle subsets sampled consistently across redshift, making it the largest high-accuracy cosmological N-body data set produced to date.

Several studies have utilized the AbacusSummit simulations to constrain the HOD of various galaxy populations. For example, Maksimova et al.² introduced the AbacusHOD model, a highly efficient and customizable HOD framework designed to take advantage of the volume and precision offered by the AbacusSummit simulations. They demonstrated the capabilities of AbacusHOD through an analysis of the full-shape redshift-space clustering of BOSS galaxies, assessing the necessity of introducing secondary galaxy biases such as

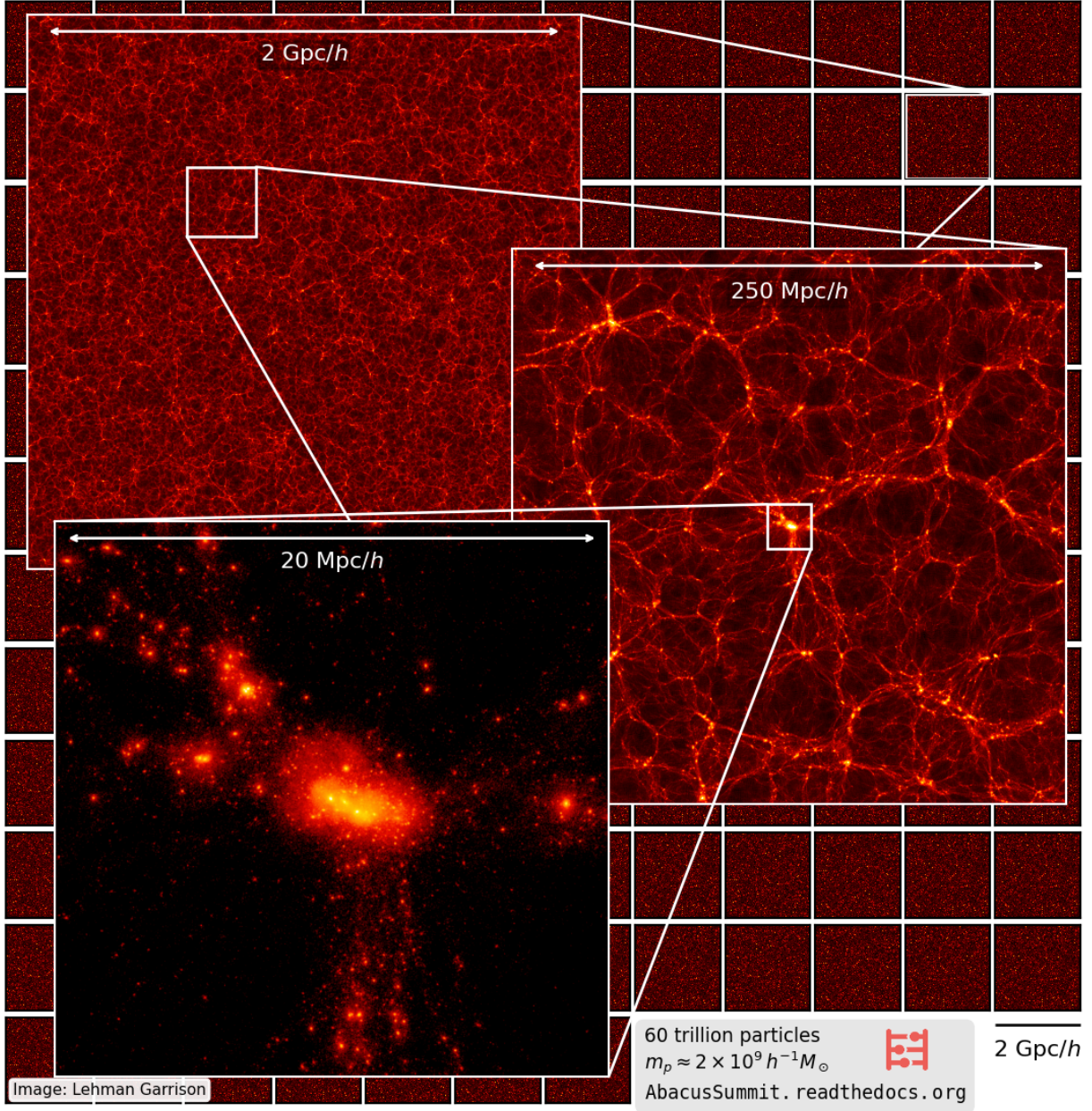


Figure 5.2.1: A visualization of the AbacusSummit base-resolution boxes taken from Maksimova et al.², showing progressive zoom-ins from the full box ($2h^{-1}$ Gpc) down to the cluster scale ($20h^{-1}$ Mpc). The 139 boxes that comprise the base-resolution suite are shown as tiles in the background. The renderings display the AbacusSummit base c000_ph000 simulation at $z = 0.1$, and the projections are $10h^{-1}$ Mpc deep. These state-of-the-art simulations, totaling roughly 60 trillion particles with particle mass $m_p \approx 2 \times 10^9 h^{-1} M_\odot$, span 97 cosmological models and provide a powerful tool for studying the evolution of the Universe and the formation of galaxies and clusters.

assembly bias.

The AbacusSummit simulations have also been used to develop novel techniques for constraining cosmological parameters and the HOD simultaneously. Yuan et al.²³⁶ presented a hybrid emulator approach to the BOSS CMASS data, that maximally derives cosmological and HOD information from non-linear galaxy clustering, leveraging the AbacusSummit simulations to achieve emulator errors comparable to the expected DESI Year 1 LRG sample variance.

In another study, Yuan et al.²³⁷ used the AbacusSummit simulations and the AbacusHOD package to present the first comprehensive HOD analysis of the DESI One-Percent survey Luminous Red Galaxy (LRG) and Quasi-Stellar Object (QSO) samples. They constrained the HOD of each sample by fitting the redshift-space galaxy 2-point correlation functions and generated a large suite of high-fidelity galaxy mocks based on their fits.

The information provided here complements the details discussed in Section 4.2.1 on page 52, offering additional context on the applications of the AbacusSummit simulations in HOD modeling and their significance for large-scale structure surveys such as DESI.

5.3 Machine Learning Pipeline for HOD Parameter Prediction

This section presents the machine learning (ML) pipeline, introduced in Section 4.1 on page 50, designed to predict clustering statistics based on HOD parameters. As illustrated in Figure 5.3.1, the left side of the flowchart (enclosed in the red box) represents the traditional brute-force approach for constraining HOD parameters. This method is computationally expensive and time-consuming, as it requires the generation of mock catalogs at each step of the Markov Chain Monte Carlo (MCMC) process to compute observational statistics. The complexity and resource demands increase significantly when incorporating multiple statistical measures.

To address these limitations, the ML emulator (right side of Figure 5.3.1) offers a more

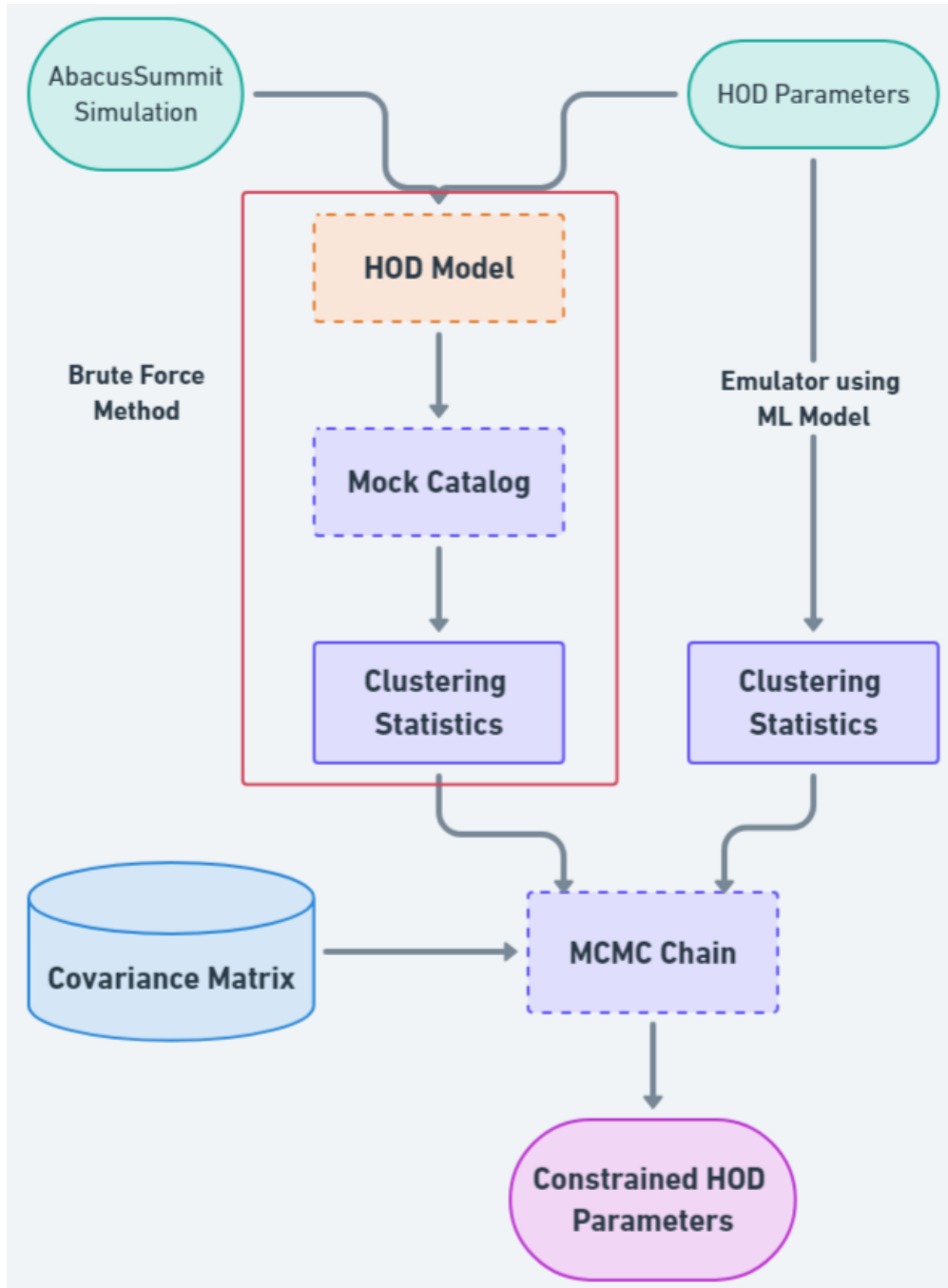


Figure 5.3.1: Flowchart illustrating the process of constraining HOD parameters using the AbacusSummit simulations and ML models. Left shows the brute force method of constraining HOD parameters by generating mock catalogs while running the MCMC chain. On the right, an ML model predicts clustering statistics directly from the HOD parameters, bypassing the need for generating mock catalogs and speeding up the parameter inference process.

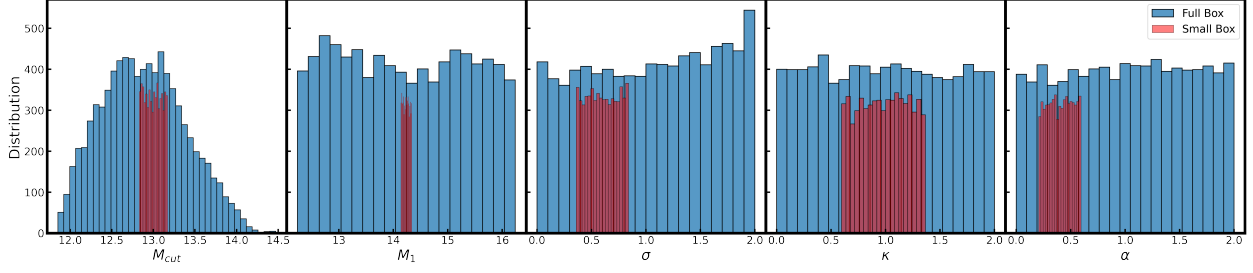


Figure 5.3.2: Input parameter distribution on full and small boxes (see Section 4.2.3). HOD parameters are drawn uniformly for both the full box and small box, but after applying filtering to the full box, the distribution of M_{cut} becomes nearly Gaussian.

efficient solution. By training machine learning models to predict observational statistics before running the MCMC chains, the need for repeated mock catalog generation is eliminated. Once the models are trained, the MCMC chains can achieve convergence in a matter of minutes, greatly reducing computational cost and enabling more efficient exploration of the parameter space.

5.3.1 Overview of the ML Methods

The overview of machine learning methods used in this research is discussed in Section 4.5 on page 60. Here we will mention some of the extended results to help understand the overall process.

Check the Distribution of the Data As mentioned in Section 4.5.1 on page 61, the HOD parameter space was initially uniform and became nearly Gaussian for M_{cut} in the full box after applying the galaxy count filter (Section 4.2.3 on page 54). The rest of the parameters still remain uniform, which proves that HOD heavily depends on M_{cut} . For the small box, the distribution remains uniform. The distribution of all the HOD parameters is shown in Figure 5.3.2, where blue bars represent the full box and red bars represent the small box.

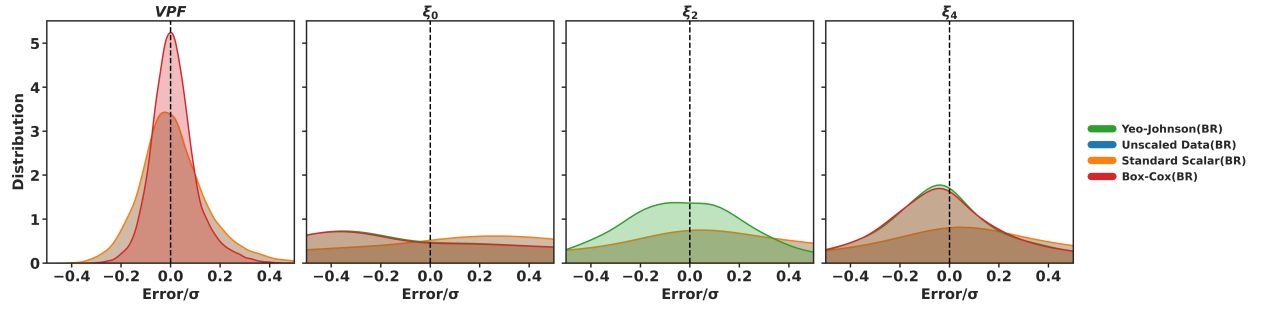


Figure 5.3.3: Density plot of VPF , ξ_0 , ξ_2 , ξ_4 between Transformed vs. Non-transformed Data as a function of $Error/\sigma$. The σ is calculated from the jackknife covariance matrix. The plot shows the impact of data transformation on the BR Method within the small Box.

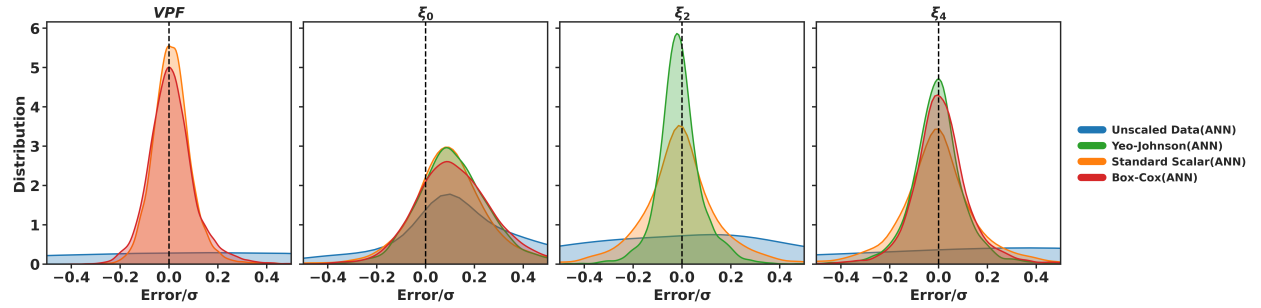


Figure 5.3.4: Density plot of VPF , ξ_0 , ξ_2 , ξ_4 between Transformed vs. Non-transformed Data as a function of $Error/\sigma$. The σ is calculated from the jackknife covariance matrix. The plot shows the impact of data transformation on the ANN Method within the small Box.

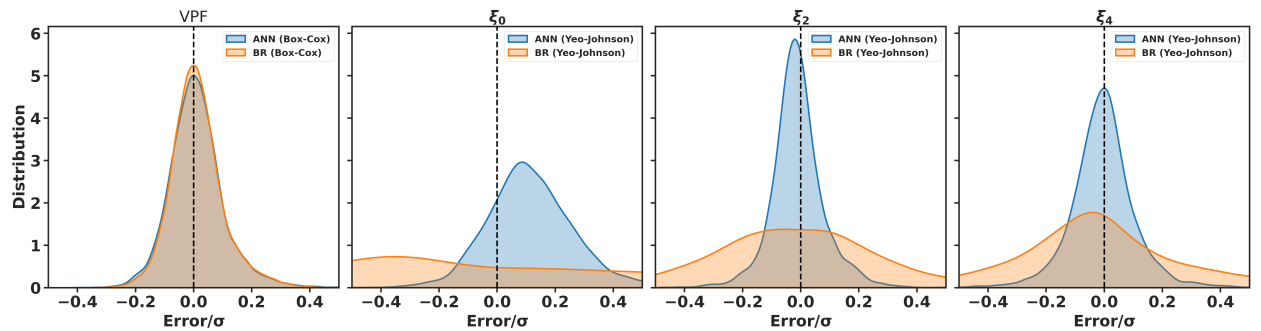


Figure 5.3.5: Density plot of VPF , ξ_0 , ξ_2 , ξ_4 between the BR and ANN models as a function of $Error/\sigma$. The σ is calculated from the jackknife covariance matrix. The plot shows the impact of different ML methods within the small box.

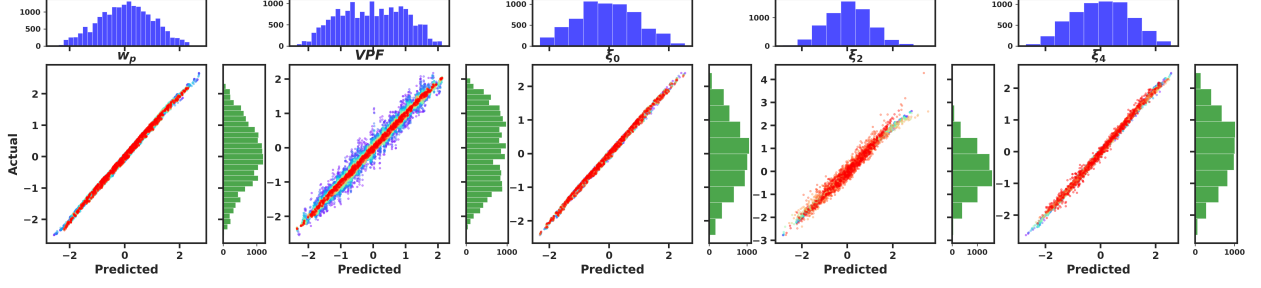


Figure 5.3.6: Comparison of predicted versus actual values for different observables (w_p , VPF , ξ_0 , ξ_2 , and ξ_4) using the BR model. Each scatter plot displays the predicted values on the x-axis and the actual values on the y-axis, with colors representing different data bins. Histograms above the scatter plots represent the distribution of predicted values, while histograms on the right represent the distribution of actual values. We show the best-performing models for each statistic.

5.3.2 Performance Evaluation and Results

Following the discussion of **Algorithm Choice** in Section 4.6.2 on page 69, we only showed various density plot comparisons of w_p (Figures 4.5, 4.6, and 4.7). In this section, we include the other statistics (VPF , ξ_ℓ).

Figure 5.3.3 shows the density plot for the Bayesian Ridge (BR) model for three different transformations: Standard Scaler (SC), Yeo-Johnson (YJ), and Box-Cox (BC), along with unscaled data. We see that even though the BR model works better for VPF with YJ transformation, it fails to provide reliable results for ξ_0 , ξ_2 , and ξ_4 . The Artificial Neural Network (ANN) in Figure 5.3.4 improves the results significantly for ξ_2 and ξ_4 , but it still does not perform well on ξ_0 . Finally, Figure 5.3.5 shows the comparison between the best-performing combinations of the statistics (VPF - BC, ξ_0 - YJ, ξ_2 - YJ, and ξ_4 - YJ). We see that ANN performs similarly for VPF and outperforms BR for ξ_ℓ .

Another way of viewing the results is shown in Figures 5.3.6 and 5.3.7. Each plot displays the true vs. predicted data points in a scatter plot (in the middle) and their overall distributions in histograms. The multiple colors on the scatter plot represent different bins. The blue histograms above the scatter plot and the green histograms beside it show the overall distributions of the true and predicted data, respectively. Here, we see that even though some of the plots like ξ_0 and ξ_2 look acceptable for the ANN and BR models, the density

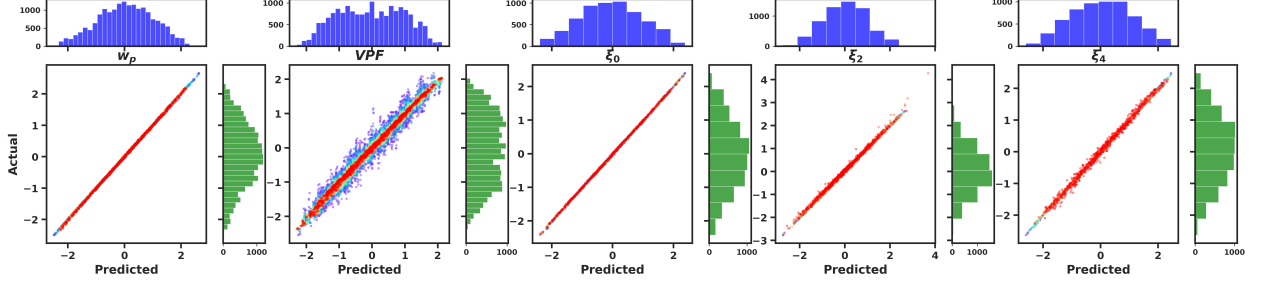


Figure 5.3.7: Comparison of predicted versus actual values for different observables (w_p , VPF , ξ_0 , ξ_2 , and ξ_4) using the ANN model. Each scatter plot displays the predicted values on the x-axis and the actual values on the y-axis, with colors representing different data bins. Histograms above the scatter plots represent the distribution of predicted values, while histograms on the right represent the distribution of actual values. We show the best-performing models for each statistic.

plots shown previously tell a different story. Therefore, we need to be careful while analyzing these models.

5.3.3 Likelihood Contours and Results

In Section 4.6.2 on page 69, we compared the likelihood contour plots for w_p . In particular, we showed the contour plot of $\log M_1$ vs σ . Here, we'll show the full five-parameter HOD triangle plot to help the reader understand the overall relationship between the parameters. We only show w_p as this is the main statistic we are analyzing. The effect of adding additional statistics to w_p is discussed in Section 4.7 on page 75.

Figures 5.3.8, 5.3.9, 5.3.10, and 5.3.11 show the full contour plots corresponding to Figures 4.5, 4.6, 4.7, and 4.10. From Figures 5.3.8, 5.3.9, and 5.3.10, we see that the contours of the BR model are biased with data transformation, whereas the ANN model is consistent with data transformation. Figure 5.3.11 shows that additional care must be taken to constrain the full box, as the constraints are inconsistent.

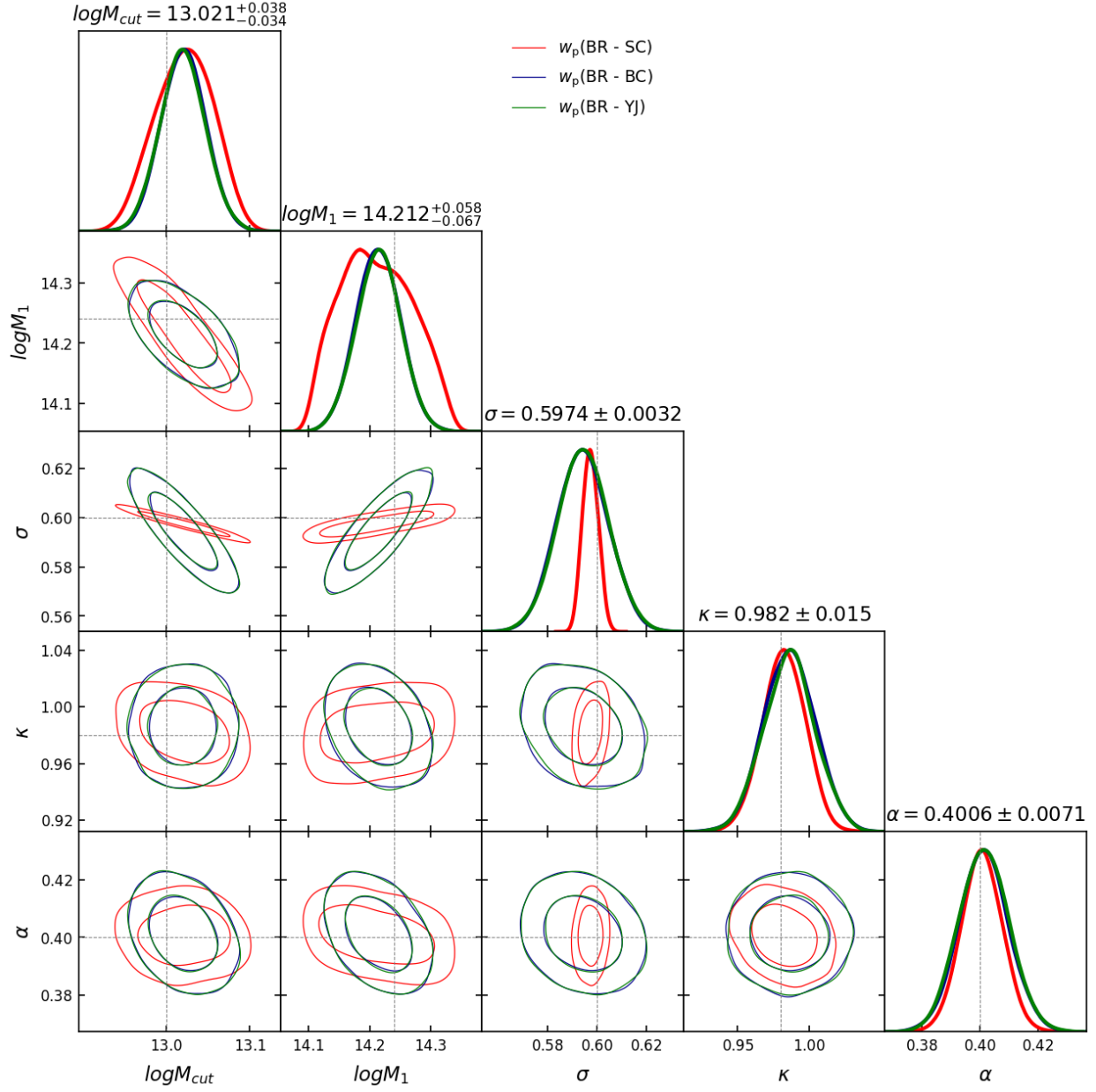


Figure 5.3.8: Exploring the impact of data transformation on BR Method within the small Box. Figure shows equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . The plot shows the convergence between different transformed data. The contours represent the 68% and 95% confidence intervals.

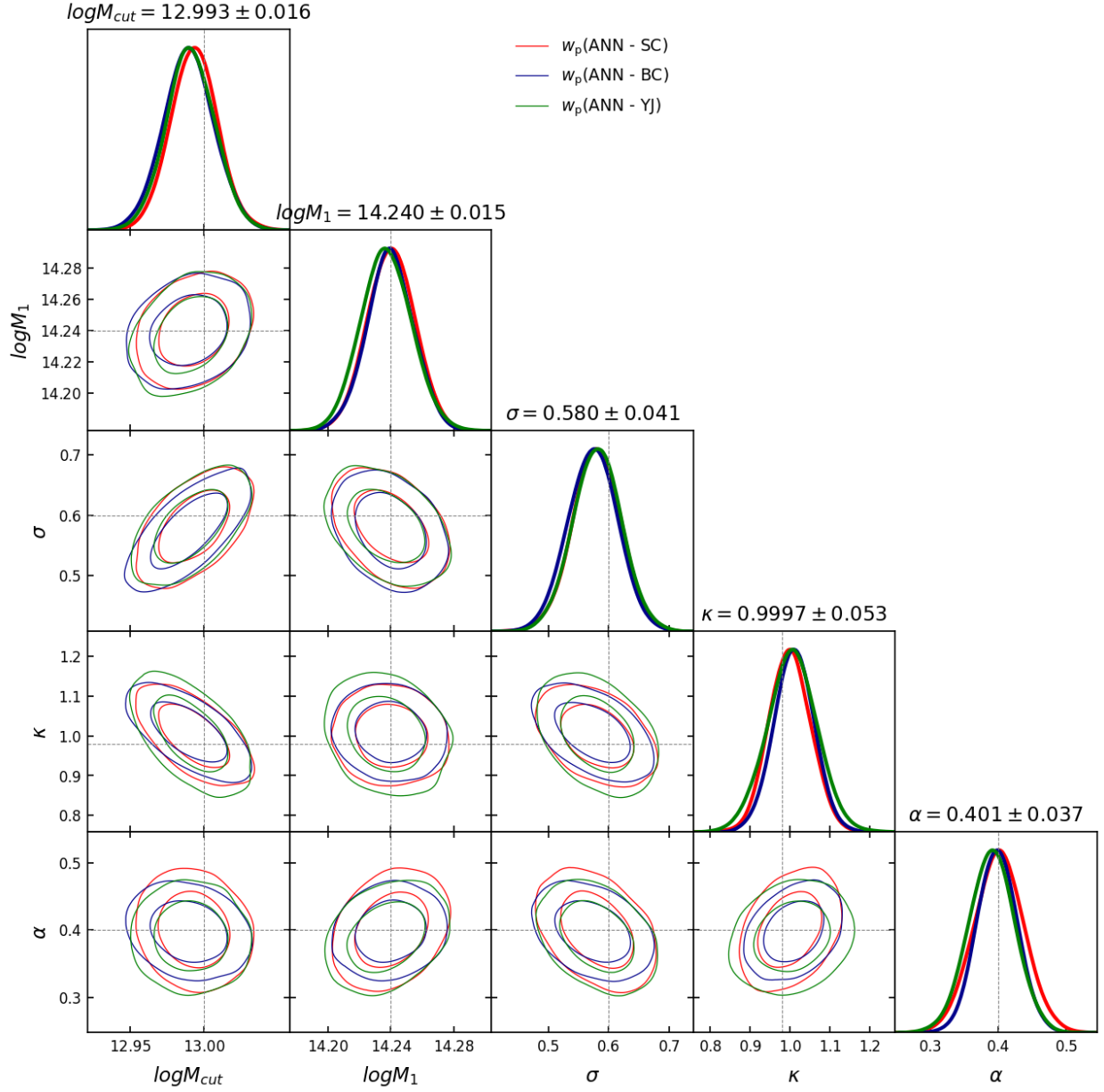


Figure 5.3.9: Exploring the impact of data transformation on ANN Method within the small Box. Figure shows equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . The plot shows the convergence between different transformed data. The contours represent the 68% and 95% confidence intervals.

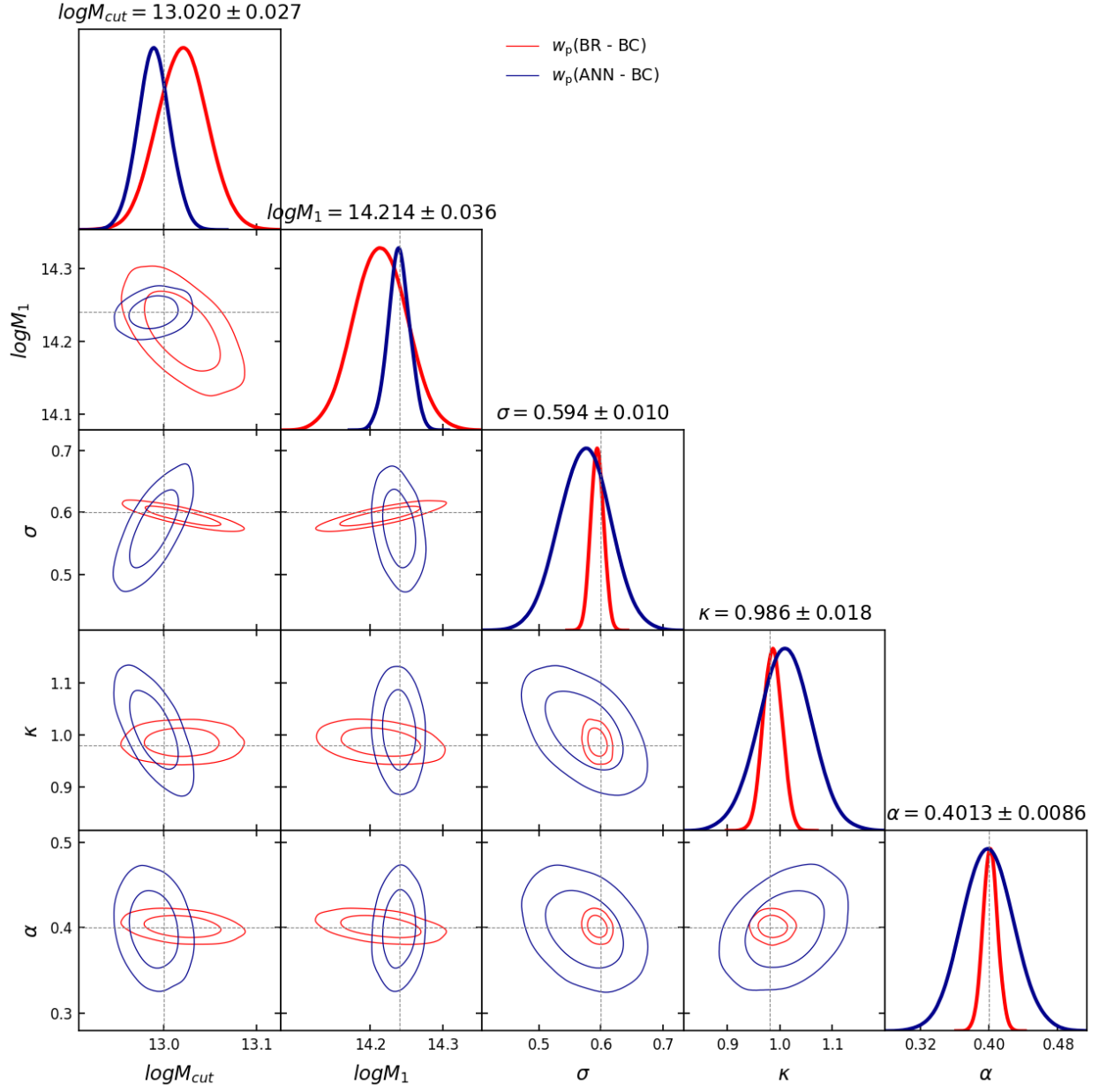
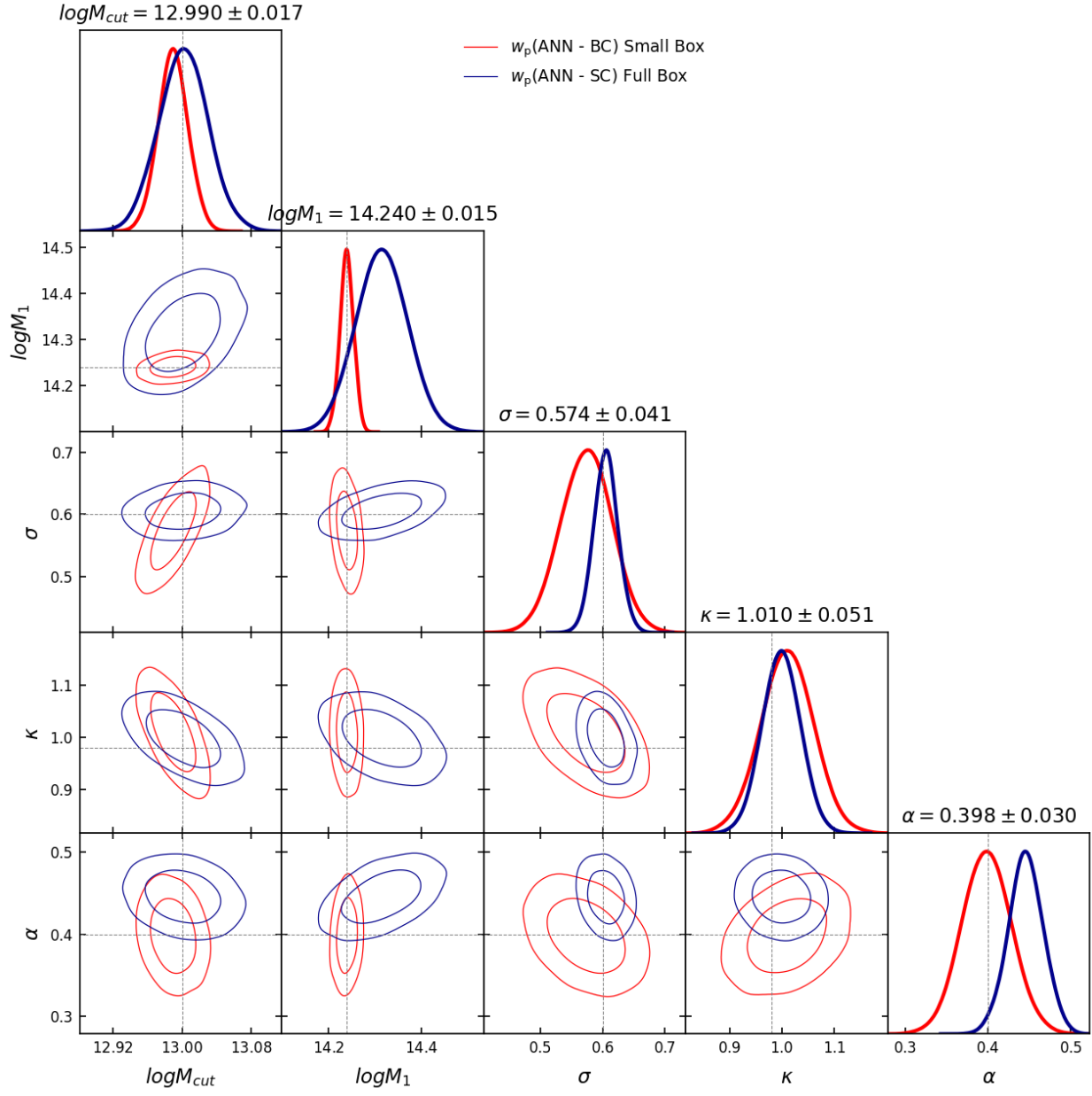


Figure 5.3.10: Exploring the impact of BR and MLP methods within the small box. The figure shows equipotential contours of the MCMC posterior from fitting the HOD model at $z = 1.1$ using w_p . The plot compares the convergence between the BR and MLP Box-Cox methods. The contours represent the 68% and 95% confidence intervals.



Full Box Hyperparameters for Bayesian Ridge Regressor					
Statistics	w_p	VPF	ξ_0	ξ_2	ξ_4
lambda_init	0.01	0.01	0.1	0.01	0.1

Table 5.4.1: Hyperparameters of BR for full box.

Full Box Hyperparameters for Neural Network					
Statistics	w_p	VPF	ξ_0	ξ_2	ξ_4
hidden_layer_sizes	500	50	50	50	50
alpha	0.0001	0.001	0.001	0.0001	0.0001
activation	tanh	tanh	tanh	tanh	tanh
learning_rate_init	0.005	0.005	0.005	0.005	0.005
learning_rate	constant	adaptive	adaptive	adaptive	adaptive
solver	adam	adam	adam	adam	adam
random_state	1	1	1	1	1
validation_fraction	0.1	0.1	0.1	0.1	0.1
early_stopping	True	True	True	True	True
max_iter	2000	5000	1000	1000	1000

Table 5.4.2: Hyperparameters for ANN on full box

5.4 Additional Results: Full Box Analysis

This section presents additional results from the full box simulations. These include the projected two-point correlation function ($w_p(r_p)$), the void probability function (VPF), and the multipoles (ξ_0 , ξ_2 , ξ_4). A comparison with the small box results is provided to highlight the consistency of the machine-learning pipeline across different simulation volumes. We want to warn readers that currently, the machine learning analysis does not work properly for the full box, so the results should not be taken seriously.

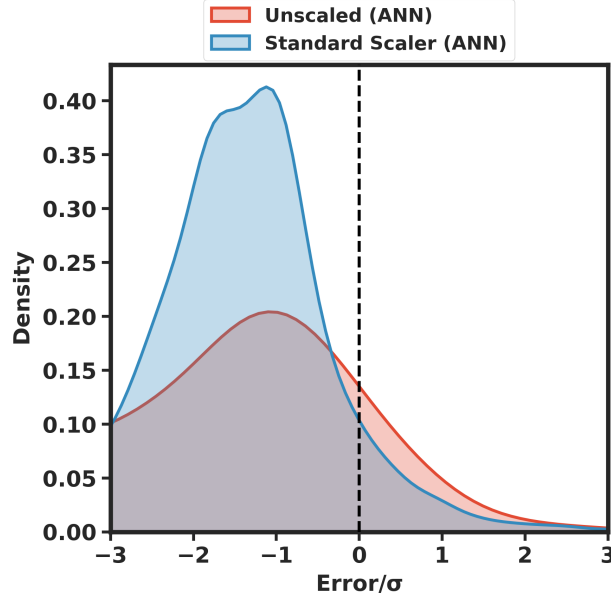


Figure 5.4.1: Exploring the impact of data transformation on ANN Method within the full Box. Density plot between Transformed vs. Non-transformed Data as a function of Error/σ . The σ is calculated from the jackknife covariance matrix.

We used the Scikit-Learn library to perform hyperparameter tuning for two models. We tried the BR and ANN models for our analysis, and the hyperparameters for these models are given in Tables 5.4.1 and 5.4.2, respectively. We used the BayesianRidge module for the BR model. The modified hyperparameters are shown in Table 5.4.1, while all other parameters were kept at their default values. For the ANN model, we used the MLPRegressor module, and the corresponding hyperparameters are listed in Table 5.4.2. We encourage readers to review Section 4.5.5 on page 63 for a refresher.

Similar to Figure 4.5, we checked the effect of data transformation on the full box in Figure 5.4.1. In this density plot, we evaluated our model on the same test set as the small box and found that even though there is an improvement with data transformation, the model fails to predict accurately.

We investigated the impact of various combinations of w_p , VPF , and ξ_ℓ on enhancing the constraining power of small-scale clustering for the HOD parameter space. Table 5.4.3 presents the results of this analysis, while Figure 5.4.2 illustrates the improvements in HOD parameter constraints.

Observables	1 σ error				
	Full Box				
	$\log M_{\text{cut}}$	$\log M_1$	σ	κ	α
w_p	$13.0166^{+0.014}_{-0.0151}$	$14.3038^{+0.0368}_{-0.036}$	$0.6271^{+0.0377}_{-0.037}$	$0.9718^{+0.0325}_{-0.0325}$	$0.3939^{+0.0485}_{-0.0466}$
$w_p + VPF$	$13.0133^{+0.0108}_{-0.0105}$	$14.2929^{+0.0288}_{-0.0293}$	$0.6183^{+0.0261}_{-0.0261}$	$0.9786^{+0.027}_{-0.0266}$	$0.3978^{+0.0338}_{-0.0324}$
$w_p + \xi_0$	$13.0115^{+0.0108}_{-0.0107}$	$14.463^{+0.0179}_{-0.0185}$	$0.7372^{+0.0215}_{-0.0212}$	$0.7564^{+0.0199}_{-0.0201}$	$0.5312^{+0.0236}_{-0.025}$
$w_p + \xi_0 + \xi_2$	$13.0465^{+0.0107}_{-0.0108}$	$14.4926^{+0.0177}_{-0.0178}$	$0.7964^{+0.0108}_{-0.0102}$	$0.7466^{+0.0156}_{-0.0162}$	$0.4677^{+0.0202}_{-0.0211}$
$w_p + \xi_0 + \xi_2 + \xi_4$	$13.1456^{+0.0091}_{-0.0081}$	$14.6517^{+0.0144}_{-0.0136}$	$0.7812^{+0.0097}_{-0.0091}$	$0.8481^{+0.0134}_{-0.0133}$	$0.2722^{+0.0124}_{-0.0127}$
ξ_0	$13.1341^{+0.0302}_{-0.0298}$	$14.5169^{+0.0395}_{-0.0376}$	$0.6237^{+0.0388}_{-0.0396}$	$0.7682^{+0.0597}_{-0.0558}$	$0.2983^{+0.0594}_{-0.0531}$
$\xi_0 + \xi_2$	$13.1752^{+0.012}_{-0.0169}$	$14.4574^{+0.0281}_{-0.0268}$	$0.7596^{+0.018}_{-0.0175}$	$0.7191^{+0.023}_{-0.0224}$	$0.2247^{+0.0345}_{-0.0177}$
$\xi_0 + \xi_2 + \xi_4$	$13.2217^{+0.008}_{-0.0079}$	$14.6573^{+0.0202}_{-0.0196}$	$0.6815^{+0.0124}_{-0.0123}$	$0.9004^{+0.0162}_{-0.0173}$	$0.2018^{+0.003}_{-0.0013}$
$w_p + VPF + \xi_0 + \xi_2 + \xi_4$	$13.142^{+0.0077}_{-0.0077}$	$14.6517^{+0.0131}_{-0.0127}$	$0.7809^{+0.0091}_{-0.009}$	$0.8538^{+0.0118}_{-0.0118}$	$0.2833^{+0.0107}_{-0.0117}$

Table 5.4.3: 1 σ error for full box. The values are calculated by running the MCMC chain and using the ANN models for w_p , VPF, ξ_0 , ξ_2 , and ξ_4 .

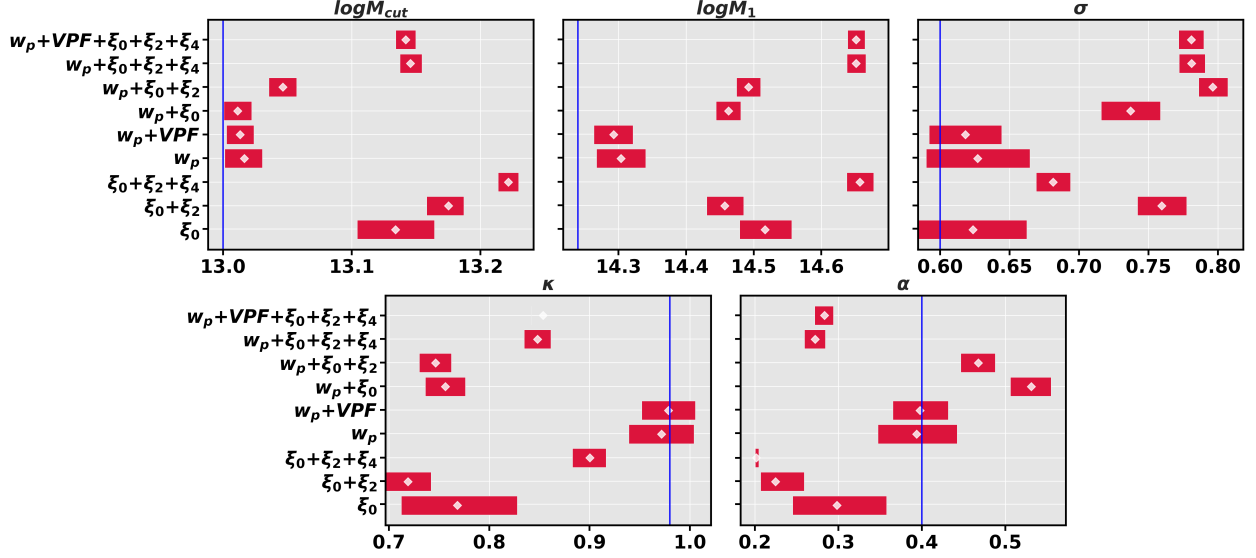


Figure 5.4.2: 1σ constraint on HOD parameters for a different combination of statistics on the full box. The blue lines denote the expected values of the HOD parameters. The white marks at the center of the red bars represent the mean of the MCMC samples after discarding 200 steps as burn-in.

Our findings reveal that incorporating VPF alongside w_p yielded a modest improvement in constraining power. However, other combinations involving w_p and ξ_ℓ proved ineffective. This outcome underscores the need for a more thorough and nuanced examination of ξ_ℓ in the context of the full box analysis.

5.5 Conclusion

This chapter expands on the main analysis results (Chapter 4), providing a detailed discussion of the machine learning pipeline for HOD parameter prediction and its application to small and full box simulations. Our exploration of machine learning techniques for predicting Halo Occupation Distribution (HOD) parameters from small-scale clustering statistics reveals both potential and challenges. The analysis, spanning small and full box simulations, highlights critical factors influencing the robustness and accuracy of HOD parameter predictions:

- **Data Preparation:** Careful examination and preparation of input data are crucial.

HOD parameter distribution and data transformation techniques significantly affect model performance and prediction reliability.

- **Model Selection:** Machine learning models (e.g., Bayesian Ridge, Artificial Neural Networks) show varying effectiveness depending on specific HOD parameters and clustering statistics. Comprehensive model comparison using multiple accuracy metrics and visualization techniques is essential for selecting the most appropriate approach.
- **Clustering Statistics Combinations:** Combining different clustering statistics (w_p , VPF , ξ_0 , ξ_2 , ξ_4) generally improves HOD parameter constraints. However, effectiveness varies across parameters, suggesting the need for tailored approaches for optimal results.
- **Scale of Analysis:** The transition from small to full box simulations introduces additional complexities, necessitating refined techniques for larger-scale cosmic structures.
- **Validation Beyond Standard Metrics:** Conventional performance metrics (RMSE, MAE, R^2 -score) provide valuable insights but may not capture all aspects of model performance relevant to cosmological applications. Small inaccuracies can lead to significant biases in cosmologically relevant quantities, emphasizing the need for thorough validation processes, including likelihood analysis in the HOD parameter space.
- **Likelihood Analysis:** Critical evaluation of machine learning results is essential, even when formal performance scores appear satisfactory. Likelihood analysis is a crucial step for further model validation in HOD analysis.

In conclusion, machine learning techniques show promise for constraining HOD parameters from clustering statistics, but require a nuanced, multifaceted approach. Future work should focus on refining these methods, particularly for full box simulations, and developing robust validation techniques aligned with cosmological objectives. This research advances sophisticated analyses of the galaxy-halo connection, potentially enhancing our understanding of large-scale structure formation and evolution.

Chapter 6

Other Project I: Using Deep Learning in Fingerprint Verification

This chapter appeared in literature as

Abhishek Jana et al. - Neural Fuzzy Extractors: A Secure Way to Use Artificial Neural Networks for Biometric User Authentication

Abstract

With advancements in sensor technology, artificial intelligence, and the decreasing cost of computing, biometric authentication has become increasingly widespread, especially with the prevalence of mobile devices. Modern biometric systems, which rely on advanced machine learning techniques, still need to store classifier data or specific biometric details, making user credentials vulnerable to potential falsification. This paper presents a secure approach for managing user information in biometric authentication systems using artificial neural networks. The proposed solution called a Neural Fuzzy Extractor (NFE), integrates existing classifiers with fuzzy extractors via an artificial neural network buffer, known as an expander, without sacrificing performance. The NFE combines the accuracy of modern deep learning classifiers with the security provided by traditional fuzzy extractors. The paper also

showcases how NFE can be applied to several classic neural networks in basic biometric authentication scenarios.

6.1 Introduction

Secure architectures for password-based authentication systems avoid storing actual passwords for each username. Instead, they store cryptographic hash digests of these passwords, often combined with salt, pepper, or other random auxiliary data. This approach is widely accepted because it ensures that even a superuser, with full access to the system’s file structure, cannot impersonate other users by accessing their passwords.

In biometric authentication systems, similar security can be achieved using fuzzy extractors²³⁸. However, despite the security advantages they offer, fuzzy extractor-based architectures are rarely implemented in real-world systems. For example, in fingerprint-based authentication, algorithms that rely on similarity scores tend to perform better than those using fuzzy extractors. A study²³⁹ reported a 15% equal error rate for fuzzy extractors on the FVC2000 database²⁴⁰, while the top five competitors in the FVC2000 challenge achieved error rates ranging from 0.7% to 5%²⁴¹.

Although similarity-score algorithms are more effective, they require storing biometric data such as fingerprints, which can be vulnerable to breaches. High-profile incidents, such as the 2015 Office of Personnel Management data breach, which exposed 5.6 million fingerprint records²⁴², and the 2019 Suprema Biostar breach, which compromised over a million people’s biometric data²⁴³, highlight this risk. Therefore, an authentication method that combines the security of fuzzy extractors with the accuracy of similarity-score algorithms would be ideal.

The underperformance of fuzzy extractors may stem from the nature of their functionality, where the first step involves a channel decoding process within a vector space. Channel coding typically assumes near-spherical decoding regions, but classifiers such as support vector machines (SVMs) excel because they can handle more complex decision regions. Similarly, artificial neural networks (ANNs) outperform SVMs when abundant training data is available

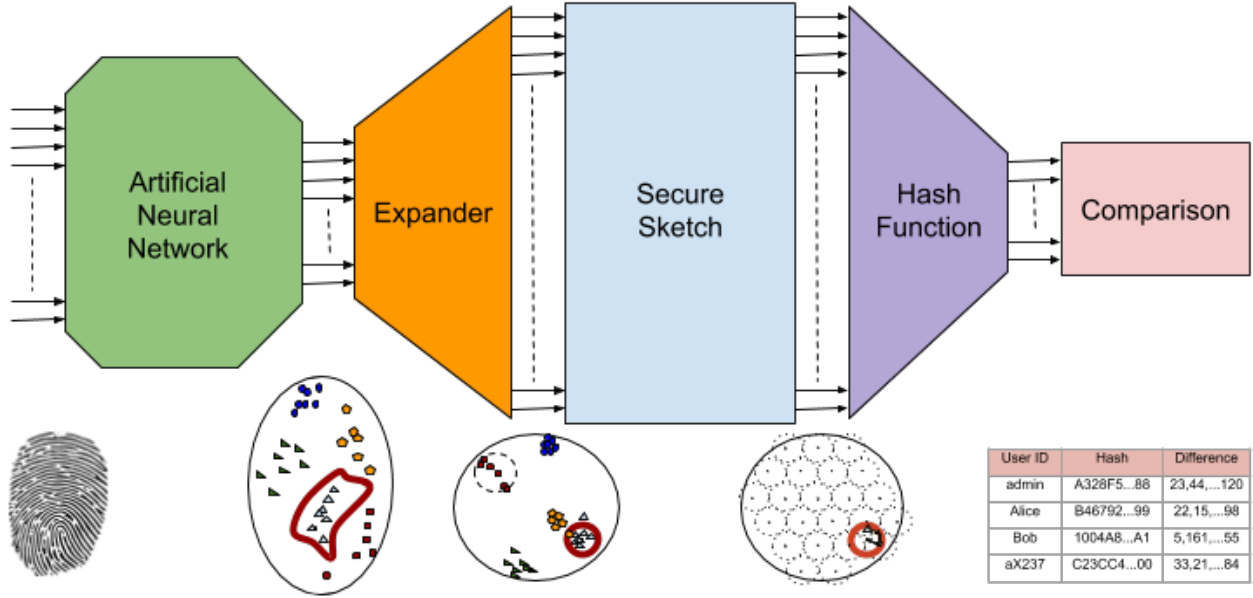


Figure 6.1.1: Neural Fuzzy Extractor architecture.

due to their ability to learn highly irregular classification boundaries.

However, both SVMs and ANNs, along with other classifiers like k-nearest neighbors (KNN), decision trees, and random forests, rely on learned structures that must be stored in non-volatile memory, similar to password files. If a malicious user gains access to this data, they could potentially exploit it to generate inputs that would pass authentication by reverse-engineering the model.

The key challenge is how to protect biometric authentication data in a way similar to password protection while avoiding the limitations of fuzzy extractors. In this paper, we propose a new architecture, which we call neural fuzzy extractors (NFEs), to address this issue.

Neural Fuzzy Extractors (NFEs) are formed by combining a classifier, like an artificial neural network (ANN), with a fuzzy extractor, as depicted in Figure 6.1.1. Since most ANNs are designed to output class labels or scores within a particular vector space, and the aim is to make NFE integration as seamless as possible with existing classification models, a modification is required to adapt NFEs to standard ANNs. This is done by creating an interface, termed an *expander*, between the ANN and the fuzzy extractor's decoding process. An expander typically consists of an additional neural network, usually with a few layers,

added to the output of an existing ANN. Its purpose is to reshape the output embeddings of the original ANN into a space where class-specific embeddings cluster into sphere-like groups. The expander can be trained either separately using labeled embeddings from the original ANN or jointly with the ANN.

In some cases, certain ANN architectures may naturally be compatible with fuzzy extractors, meaning their output embeddings already form clusters in a suitable vector space. In such cases, an expander may not be necessary.

Moreover, while the vector spaces discussed in this paper are defined over $\mathbb{R}^n$, they could also be defined over other fields, such as $GF(2^n)$. This choice can simplify the secure sketch process due to the availability of error correction codes (ECCs) optimized for binary channels. In contrast, ECCs designed for additive white Gaussian noise (AWGN) channels are less common and involve more complex decoding methods. However, converting the expander’s output into binary vectors may result in unacceptable performance reductions, which is why binary-ECC-based NFEs are left for future exploration.

The key contributions of this paper include the following:

1. Introduction of NFEs, a secure architecture for integrating ANNs with biometric user authentication systems.
2. Demonstration that NFEs can be adapted to work with many existing ANN-based biometric authentication models.
3. Application of the NFE framework to three established fingerprint-based authentication systems, showing minimal impact on performance while adding a manageable increase in training and running time.

The paper is organized as follows: Section 6.1 introduces the NFE architecture and provides some general implementation insights. Section 6.3 describes the design choices made for our NFE instantiation, including classifier selection, expander architecture, and error correction codes. Section 6.4 outlines the evaluation framework, including datasets and experimental setups, and presents performance results. The conclusions are summarized in Section 6.5.

6.2 Preliminary Concepts: The Architecture of a Neural Fuzzy Extractor

Most classifiers, whether they are neural networks or based on vector spaces, generate vector representations of input data. In certain cases, these output vectors are already well-suited for direct input into the secure sketch (see Figure 6.1.1). For this to work, the classifier must ensure that the outputs corresponding to each class form clusters in roughly spherical regions within the vector space. This requirement arises because the secure sketch employs error correction codes (ECCs) that are designed for either additive white Gaussian noise (AWGN) channels or binary symmetric channels (BSCs), where the decoding regions are inherently spherical. If the classifier’s decision regions are not already spherical, imposing spherical regions on top of them can degrade the classification performance.

However, many classifiers are designed solely to optimize accuracy, without considering the spherical shape of their decision regions. To address this issue, we propose the use of an *expander* (as shown in Figure 6.1.1), which reshapes the classifier’s decision regions into spherical clusters. We call this process *retrofitting the NFE to pre-existing classifiers*. Ideally, we want to maintain the classifier’s accuracy while reshaping its output, which means the expander should project the classifier’s vector output into a higher-dimensional space where spherical decision regions can be formed without significantly impacting accuracy. In the case of neural networks, where the final layer often reduces the dimensionality to match the number of relevant classes, we suggest removing this final layer before adding the expander. If the classifier has a wide output, the expander may act as a compression mechanism to reduce the dimensionality to a size manageable by the secure sketch, whose decoding process can be computationally intensive, although it runs in linear time relative to code length.²⁴⁴

6.2.1 The Classifier

As mentioned, when retrofitting the NFE to existing classifiers, the primary requirement is for the classifier to transform users’ biometric readings into vectors within a vector space

where classification can be done with reasonable accuracy. The NFE retrofit does not aim to improve the classifier’s accuracy but rather to preserve it while enhancing the system’s security. Therefore, our NFE approach is compatible with a variety of existing classifiers, such as K-nearest neighbors (KNN), support vector machines (SVMs), or neural networks. The term “*neural*” in Neural Fuzzy Extractor refers not to the classifier itself, but to the expander, which is implemented as an artificial neural network.

6.2.2 The Expander

The expander is implemented as a neural network that takes the output of the original classifier (whether modified or intact) as input. It is trained using a cost function that penalizes deviations from spherical clustering. When the classifier is also a neural network, the expander can be trained concurrently with the classifier, effectively adding a constraint on the shape of the decision regions. In cases where the NFE is retrofitted to an already-trained neural network, the expander can be trained separately by using the labeled outputs of the original network and feeding them into a new or existing training dataset.

6.2.3 The Secure Sketch

The secure sketch, as introduced in Dodis et al.²³⁸, consists of two components: (1) a mechanism that maps a user’s fuzzy biometric data to a fixed point in the vector space of the biometric representation, and (2) a secure storage method for this identifying information. If the user’s biometric data is positioned within a roughly spherical region of the vector space, an error correction code can be defined over the vector space, aligning the user’s decision sphere with one of the decoding regions of the code.

This construction is illustrated in Figure 6.2.1. The left section of the figure shows the authentic user’s (AU’s) biometric data (represented as light blue triangles) mostly falling within the bottom-right red sphere, which represents the user’s *decision region*. The radius of this sphere is user-specific and must balance false positives and false negatives. The larger circle in the figure represents the hyper-sphere, encompassing all possible embeddings. This

corresponds to the support of the expander’s output, essentially the region where one would expect the biometric embeddings of any user to be located.

Registration phase In this phase, a codebook is defined over the vector space, restricted to the support of the expander’s output. Various methods can define this code, but an optimal choice would be a capacity-achieving code for AWGN channels if the vector space is $\mathbb{R}^n$, or for binary symmetric channels if the space is $GF(2^n)$. In our example, where embeddings are defined over $\mathbb{R}^{128}$, we choose a low-density lattice code (LDLC)²⁴⁵ in 128 dimensions. The codebook contains codewords, represented as black dots in the center of Figure 6.2.1, each surrounded by its decoding sphere (or *Voronoi* region). Any vector falling within a codeword’s Voronoi region can be decoded to that codeword.

Next, the center of the AU’s decision region, $\mathbf{r}_i$, is decoded to the closest codeword in the codebook, denoted $\mathbf{c}_i$. The difference vector (DV) $\mathbf{d}_i = \mathbf{r}_i - \mathbf{c}_i$ is then calculated and stored as part of the AU’s authentication record, along with the hash $h(\mathbf{r}_i)$ of the center of the AU’s decision region.

Verification phase For each registered user, the verifier has access to their tuple $(\mathbf{d}_i, h(\mathbf{r}_i))$. When verifying a user claim, the current biometric reading $\mathbf{b}$ is mapped into the vector space by the expander-enhanced classifier. The DV is subtracted from it, giving $\mathbf{f}_i = \mathbf{b} - \mathbf{d}_i$. In Figure 6.2.1, $\mathbf{b}$ is shown as a light blue triangle, while $\mathbf{f}_i$ is the tip of the yellow arrow. This vector is decoded to the closest codeword $\mathbf{c}_j$. If $\mathbf{b}$ matches the biometric of user i , we should have $\mathbf{c}_j = \mathbf{c}_i$. Finally, the DV is added to the codeword $\mathbf{c}_j$, and the hash $h(\mathbf{c}_j + \mathbf{d}_i)$ is compared to $h(\mathbf{r}_i)$ in the user’s record. Authentication is successful if the two hashes match.

Note 1: Instead of decoding the AU’s decision center to the nearest codeword, a random codeword could be selected, with the DV stored between the center and the chosen codeword. However, this introduces the risk of leakage, which requires further steps to mitigate. For example, if the selected codeword is far from the decision center, an attacker might infer the general location of the AU’s decision region. To prevent this, a random vector can be added to the DV, neutralized when wrapped around the larger sphere.

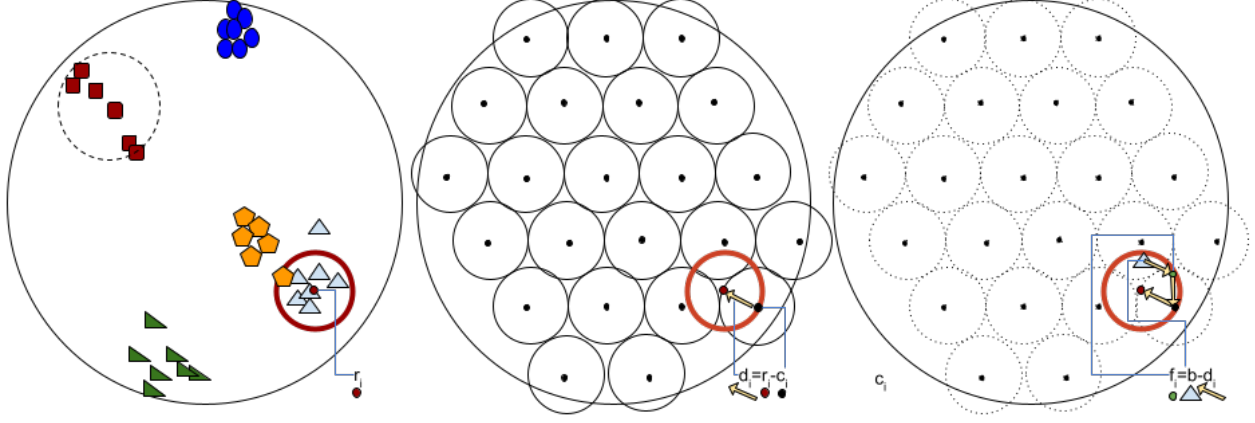


Figure 6.2.1: Secure sketch construction. Left: The large circle contains all available embeddings from all users; the small circle for the authentic user (AU) is chosen to optimize the trade-off between false positives and false negatives (different AUs may have different radii). Middle: A codebook is built with a Voronoi region congruent to the AU’s decision region; the center of the AU’s decision region (r_i) is decoded to the nearest codeword (c_i), and the difference between the center and this codeword (d_i) is saved in the AU’s record, along with the hash of the center of the decision region. Right: The AU provides a new sample for authentication (b); subtracting the difference vector (d_i) and decoding to the nearest codeword (c_i) retrieves the previously identified codeword. Adding the difference to this codeword (d_i) yields the center of the AU’s decision region (r_i), which is hashed and compared to the AU’s stored hash.

Note 2: Hashing the nearest or randomly chosen codeword directly might seem like a simpler approach. However, this would require ensuring that different users are assigned distinct codewords, adding complexity to the system. Storing the hash of the AU’s decision region center, as described, naturally avoids this issue.

6.2.4 The Hash

The hash component within the NFE architecture is implemented using a simple cryptographic hash function, selected based on the latest recommendations from NIST. Currently, hash functions from the SHA-2 and SHA-3 families are suitable options. The authentication database stores the following identifying components for each user, along with their username: (1) codebook parameters or just a decoding algorithm, (2) the difference vector (DV) as described in Section 6.2.3, (3) any non-secret randomness such as salt and pepper used in the hashing process, and (4) the hash value (with salt, pepper, etc.) of the center of the

AU’s decision region, as outlined in Section 6.2.3.

6.2.5 Security Considerations

Threat Model We assume a biometric-based authentication system that utilizes an artificial neural network (ANN) hosted on an authentication server (or, in some cases, a local machine that users attempt to log into). Legitimate users undergo a registration phase where multiple biometric readings are captured to train the system. The trained authentication system, along with users’ authentication records, is securely stored (and possibly encrypted) on the authentication server. We assume no malicious interference or eavesdropping during the registration phase, which takes place in a secure environment, similar to setting up new accounts and passwords in a password-based system.

Additionally, we assume that the attacker does not have real-time access to the authentication process or the biometric information provided by the user. However, such access may be possible if the attacker eavesdrops on the connection between the biometric sensor and the server or has administrative privileges to monitor memory during the authentication process. These assumptions are common in biometric systems, and further strengthening the system to address such risks is compatible with our approach but beyond the scope of this work.

The threat model involves an attacker who gains access to the weights of the ANN-based system and the biometric records of all users after the registration phase. The attacker cannot access the server beyond this point, except potentially as a regular user. Additionally, the attacker can produce a single biometric reading that is indistinguishable from a real one. This restriction prevents brute-force attacks, which depend primarily on the entropy of the biometric data rather than the security mechanism itself.

The attacker’s goal is to successfully login as any legitimate user, excluding any they already control or have compromised. Success is achieved if the attacker can log in as a legitimate user with a probability significantly higher than a random individual attempting to authenticate with their biometric data.

Security Evaluation of the NFE To successfully impersonate a legitimate user, the attacker would need to input the user’s name (which is available from the authentication records) and generate a biometric reading that, when processed by the authentication system, matches the stored output for that user.

In a typical ANN-based system, this task is possible because each user’s output is represented as a defined region in the output vector space. The attacker could select a vector within this region and reverse-engineer the corresponding input to the ANN classifier, producing a valid biometric reading. Under our threat model, the attacker can force the system to process this chosen biometric reading.

However, with an NFE-enhanced system, the attacker’s challenge is significantly greater. Each user’s authentication record consists of a username, a difference vector, and a hash value. The difference vector provides no information about the pre-image of the hash value, and finding the pre-image from the hash is infeasible due to the hash function’s pre-image resistance. Thus, the attacker has no way of determining what the expander’s output should be for a given user. Their best option is to select a random codeword and add the difference vector from the user’s record, hoping that the result matches the center of that user’s decision region. The attacker can then reverse the process through the ANN and expander to generate a biometric reading corresponding to this result. However, the number of codewords is designed to match the number of distinguishable user profiles, so selecting a random codeword is effectively equivalent to picking a random person and asking for their biometric data. Consequently, the attacker’s chances of success are no better than those of a random individual attempting to authenticate.

Thus, under our threat model and assuming a pre-image-resistant hash function, the attacker cannot successfully compromise the system.

NFEs, Codebooks, Sphere Packings, and Biometric Entropy The goal of the NFE is to ensure that embeddings of different users’ biometric data form quasi-spherical clusters in a specific vector space. In Figure 6.2.1, we depict multiple quasi-spherical clusters, where each cluster (denoted by different markers) corresponds to the embeddings of a particular

user’s biometric data. The expander’s output is assumed to be supported by a sphere that encloses all biometric data points. Choosing the smallest sphere may artificially reduce the perceived security of the biometric system by yielding a smaller codebook, so a larger sphere with a radius 10% bigger than the smallest is typically chosen.

The radius of the user-specific decoding region (the small sphere representing a user in Figure 6.2.1) is chosen to achieve an optimal false-positive and false-negative balance. For simplicity, we select the same radius (i.e., the same error correction code) for all users.

For example, using the VGG-16 architecture retrofitted with the NFE and the FVC2006 database, the outer sphere has a radius of 1.0153, and the small sphere has a radius of 0.7 (selected to achieve a 95.36% accuracy on the training set). This configuration allows for a maximum of $(1.0153/0.7)^{128} \approx 4.7 \times 10^{20}$ small spheres to be packed inside the large 128-dimensional sphere. Thus, the maximum number of distinct individuals whose biometric embeddings fit within non-overlapping small spheres is approximately $4.7 \times 10^{20} \approx 2^{68.67}$, meaning that the entropy of the secret biometric information in this space is around 68.67 bits.

In other words, the NFE approach not only improves security but also offers a way to estimate the authentication potential of different types of biometric data. This methodology allows us to assess whether, for example, fingerprint-based biometrics are more or less secure than retina-scan biometrics. The evaluation depends on the classifier’s accuracy across different AUs but cannot be easily derived without the expander.

Previous studies have approached biometric entropy quantification in various ways. For instance, Yankov et al.²⁴⁶ provides empirical distributions of fingerprint data, calculating entropy per pixel for several fingerprint databases, including FVC2002 and FVC2004, which are comparable to our FVC2006 database. Their estimated entropy per pixel is slightly above 0.25 bits, leading to a total entropy of approximately 50,000 bits per image. However, this reflects the entropy of the fingerprint image, not the biometric modality itself. To approximate the latter, Yankov et al.²⁴⁶ uses mutual information between fingerprint images and estimates a maximum of 10^{28} distinct individual representations, which makes our estimate of 4.7×10^{20} conservative. Other approaches to evaluating biometric entropy can be found

in Lim and Yuen²⁴⁷ and Sutcu et al.²⁴⁸

6.3 An Instantiation of NFE Retrofit

To implement the NFE retrofit architecture described earlier, several design decisions were required. First, we selected three pre-existing classifier architectures, detailed in Section 6.3.1, as the foundation for the retrofit. Each of these architectures underwent minor modifications. Additionally, we used an expander architecture composed of multiple fully connected layers, with progressively smaller sizes, to reduce the original classifier output to a fixed size of 128 neurons. This was necessary to ensure a manageable decoding process. For the error correction, we selected a 128-dimensional low-density lattice code, as outlined in Sommer et al.²⁴⁵. The following sections provide further details on these design decisions.

6.3.1 Classifier and Expander Architecture

VGG16 The VGG16 classifier^{249;250} was used with weights pre-trained on ImageNet via Keras. To adapt it for fingerprint-based authentication, we removed the final softmax layer, resulting in an output size of 4608. This vector space size (4608) is too large for efficient decoding. To retrofit the NFE, we introduced an expander by adding three fully connected layers with 512, 256, and 128 neurons, respectively.

ResNet50 The ResNet50 classifier²⁵¹, also pre-trained on ImageNet using Keras, had its final softmax layer removed, leaving an output size of 2048, achieved by average pooling the last layer. To incorporate the NFE, we added an expander consisting of four fully connected layers, with sizes 1024, 512, 256, and 128 neurons, respectively.

MobileNet We used a modified MobileNet v1 architecture²⁵², pre-trained on ImageNet through Keras. After removing the final softmax layer, the output was 1024, achieved by average pooling the last layer. The expander for this classifier consisted of three fully connected layers with 512, 256, and 128 neurons, respectively.

One Shot Classification and Siamese Network In traditional identification methods, a set of images is input into an artificial neural network (ANN) to produce an output probability for each class. For instance, if we want to distinguish between a cat and a dog, we typically collect a large number of images (usually over 500 per class) to improve model accuracy. However, this approach presents two challenges in fingerprint identification: **first**, it is often impractical to gather a large number of images, and **second**, introducing a new user into the database would require retraining the entire model to recognize that user. To address these challenges, we adopt a Siamese network configuration for training our classifier, as discussed below.

It is important to note that, in this specific scenario, our “expander” does not expand the ANN’s output but instead contracts it. This reduces the complexity of the secure sketch’s decoding process. Nevertheless, the same principles apply in cases where the expander actually increases the output size.

A Siamese network, sometimes referred to as a twin neural network, is an ANN designed to learn the similarity between two inputs rather than performing classification. It processes two input images through the same network in parallel, producing two vector embeddings of the images. These embeddings are compared using a logistic loss function to calculate a similarity score between the two images²⁵³. This approach is beneficial because it requires fewer data points to train the model. For training, only one reference image belonging to the legitimate user needs to be stored, and the similarity score is calculated for each new image presented to the network.

In our implementation, we use a triplet loss function with Siamese networks. The advantage of combining the triplet loss function (explained in the next subsection) with a Siamese network is twofold:²⁵⁴

1. It extracts more features by learning to maximize the similarity between two similar images (Anchor-Positive) while maximizing the distance between two different images (Anchor-Negative).
2. It generates more training samples than the logistic loss. If there are P similar pairs and

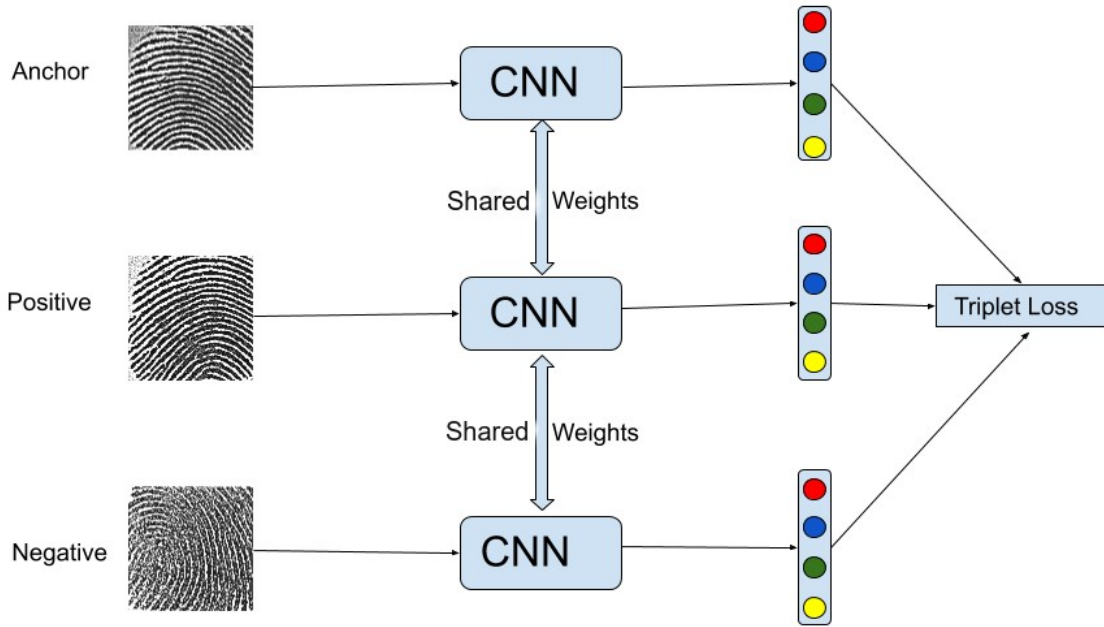


Figure 6.3.1: Triplet Loss architecture with the Siamese network: Three images “Anchor”, “Positive”, and “Negative”) are processed through the same CNN simultaneously, generating a 128-dimensional vector in the final layer. These vectors are then passed through the triplet loss function, which minimizes the distance between “Anchor” and “Positive” while maximizing the distance between “Anchor” and “Negative”.

N dissimilar pairs, the logistic loss would result in $P + N$ training samples. In contrast, triplet loss generates PN triplets for training, which improves model accuracy.

Figure 6.3.1 shows the architecture of our Siamese network. An interesting property of training Siamese networks is that the embeddings naturally cluster into spheres, as the loss function forces embeddings of the same class to be close together and those of different classes to be far apart in terms of Euclidean distance. However, in situations with more training data, alternative expander architectures can be used, as long as their loss functions incorporate similar distance-based penalties.

Triplet Loss Function Triplet loss functions are commonly used in computer vision tasks, such as face recognition²⁵⁵, person re-identification²⁵⁶, and image retrieval²⁵⁷. Drawing inspiration from these applications, we implemented triplet loss for fingerprint verification. Below, we explain how the triplet loss function works.

Using a convolutional neural network (CNN), we can map an image x to a d -dimensional Euclidean space, where the embedding is denoted $f(x) \in \mathbb{R}^d$, with f being the function computed by the CNN. For training, we use triplets of fingerprint images, as depicted in Figure 6.3.2:

- **A** is an “anchor” image — a fingerprint image of the user.
- **P** is a “positive” image — a fingerprint of the same user.
- **N** is a “negative” image — a fingerprint of a different user.

Each triplet is represented as $(A^{(i)}, P^{(i)}, N^{(i)})$, where i denotes the i^{th} training example. We aim to ensure that the positive pair (A, P) is closer than the negative pair (A, N). Mathematically, we want to satisfy:

$$\|f(A^{(i)}) - f(P^{(i)})\|_2^2 + \alpha < \|f(A^{(i)}) - f(N^{(i)})\|_2^2 \quad (6.1)$$

for all triplets $\{f(A^{(i)}), f(P^{(i)}), f(N^{(i)})\} \in T$, where T is the set of all possible triplets. The margin α ensures that the positive pair has at least a margin of difference over the negative pair.

Thus, the triplet cost function for the CNN becomes:

$$\sum_{i=1}^n [\|f(A^{(i)}) - f(P^{(i)})\|_2^2 - \|f(A^{(i)}) - f(N^{(i)})\|_2^2 + \alpha]_+. \quad (6.2)$$

Generating all possible triplets during training can lead to slower convergence, so it is essential to use a combination of “hard” and “easy” batches to improve model performance.

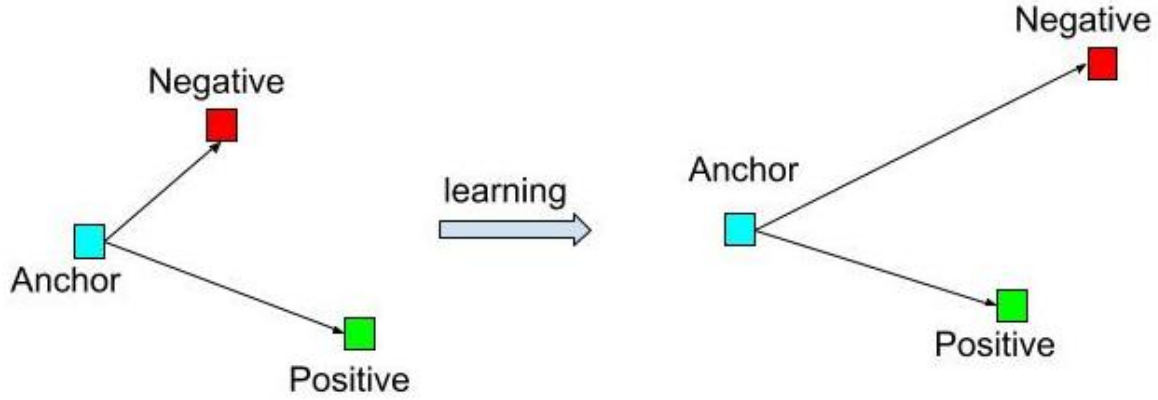


Figure 6.3.2: Triplet Loss architecture: This architecture minimizes the distance between the “Anchor” and “Positive” while maximizing the distance between the “Anchor” and “Negative”.

6.3.2 The Error-Correction Code

As discussed in Section 6.2.3, we utilized a low-density lattice code (LDLC), directly taken from Sommer et al.²⁴⁵, in $n = 128$ dimensions for our error-correction code. According to Sommer et al.²⁴⁵, the LDLC is defined by its sparse parity-check matrix H . We employed the *Latin Square* construction described in Sommer et al.²⁴⁵, where each row and column of H contains the same d nonzero values. The nonzero values are taken from the sequence $\{1/2.31, 1/3.17, 1/5.11, 1/7.33, 1/11.71\}$. They are distributed across the rows and columns of H following the algorithms provided in Appendix VII of Sommer et al.²⁴⁵ We experimented with $d = 3$ and $d = 5$. The error-correction performance of both codes is shown in Figure 6.3.3. Each point in the figure represents the result of applying the decoding algorithm 20 times to a randomly chosen codeword distorted by additive white Gaussian noise (AWGN) with a standard deviation σ , shown on the horizontal axis. The results indicate that the $d = 3$ code performs slightly better, both in artificial settings and in practical NFE performance.

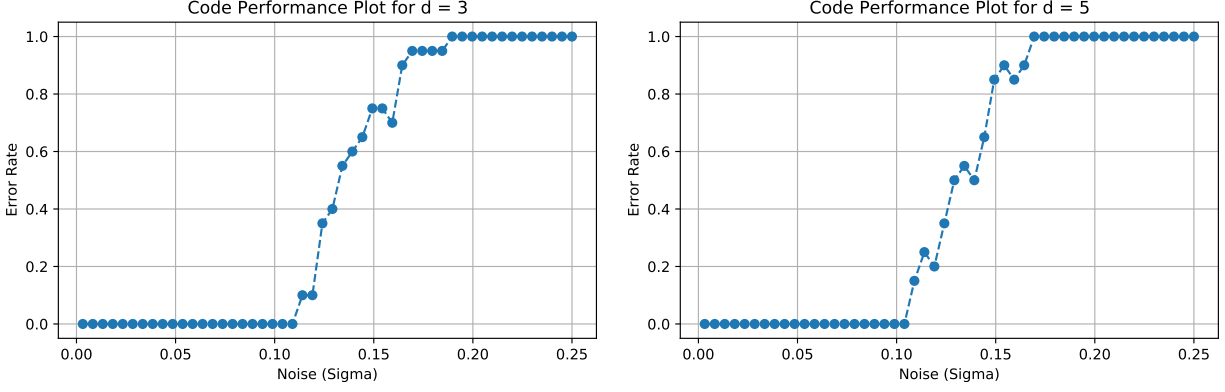


Figure 6.3.3: Error rate vs. noise level for the Latin-Square LDLC with row/column degree d .

6.3.3 System Configuration Guidelines

Before evaluating the performance and cost of our NFE implementation, we outline key configuration aspects for setting up NFE-retrofitted fingerprint authentication.

Number of users There is no practical limit to the number of users. Our conservative estimate of $4.7 \cdot 10^{20}$ differentiable users means that, even with a large number of registered users, the probability of false identification remains low, especially when authentication also requires a username.

Original classifier preparation When retrofitting a neural network-based classifier for NFE, we recommend removing the final softmax layer (if present). The output size should be large enough to avoid excessive compression and loss of biometric information.

Codeword size The appropriate codeword size n depends on the efficiency of the decoding algorithm and the acceptable computation overhead. We recommend choosing n between 100 and 500 based on current consumer-grade hardware capabilities. For more detailed results, refer to Wang and Mow²⁵⁸.

FPR-FNR tradeoff The tradeoff between false positive rate (FPR) and false negative rate (FNR) for LDLC codes is controlled by scaling the NFE’s output with a scalar factor γ . We recommend experimenting with multiple values of γ to build a receiver operating characteristic (ROC) curve and select an appropriate operating point.

6.4 NFE Evaluation

In evaluating our NFE retrofit, several experimental design choices were made. First, we selected two fingerprint databases of varying sizes. We then conducted four distinct experiments, allowing comparisons between different retrofitted architectures and two baseline architectures, applied across both databases and all three classifiers. Finally, we explored various methods to fine-tune the pre-trained classifiers (which were slightly modified as detailed in Section 6.3) and selected the best-performing configurations. The remainder of this section provides further details on these choices and the resulting experimental outcomes.

6.4.1 Datasets

For the evaluation, we used two datasets: the 2006 Fingerprint Verification Competition (FVC2006) Database²⁵⁹ and the PolyU fingerprint database provided by the Hong Kong Polytechnic University.²⁶⁰

The FVC2006 Database This dataset includes four subsets: DB1, DB2, DB3, and DB4, each containing 150 fingers and 12 impressions per finger. Each subset is divided into “set A” with 140x12 images and “set B” with 10x12 images. In this study, only DB1 was used, which consists of 96x96-sized images. These fingerprints were collected as part of the European Project BioSec from anonymous volunteers and were publicly released during the 2006 Fingerprint Verification Competition, organized by several renowned institutions, including the Biometric System Laboratory (University of Bologna) and others.²⁵⁹ At the time of writing, we have no reason to question the ethical nature of the data collection.

The PolyU Database This database contains 2016 fingerprint images from 336 users, with 6 impressions per user. The original image size was 328x356, but we augmented the dataset to ensure consistency with our initial dataset by applying random noise, rotations, and other techniques. This increased the dataset to 12 impressions per user, and the image size was standardized to 224x224. The PolyU database was released in 2017 by the Hong Kong Polytechnic University for public use, and, like the FVC2006 dataset, we have no ethical concerns at the time of writing.

Training Data For each database, training data consisted of 10 impressions per finger (140x10 images for FVC2006 and 336x10 images for PolyU). From these, we generated triplet pairs, with 50% being “hard” pairs and 50% “easy” pairs.

Testing Data The testing data included 2 impressions per finger for both databases.

6.4.2 Experimental Setup

We performed four experiments using ResNet50, MobileNet, and VGG16 architectures with the two datasets: FVC2006 and PolyU.

Baseline Classifier (BC) The first experiment was designed to provide a baseline for evaluating expander-enhanced architectures. We trained the three classifiers—ResNet50, MobileNet, and VGG16—individually. The classification model’s top layer was a softmax layer with 140 nodes for FVC2006 and 336 nodes for PolyU. We calculated the false positive rate (FPR) and false negative rate (FNR) by establishing score thresholds for each user and conducting tests using 20 additional fingerprint images for FPR and 20 augmented images for FNR.

Classifier Trained in a Siamese-Network Configuration (CSN) In the second experiment, we trained each classifier using a triplet-based Siamese network. No expander was applied, and the size of the final embeddings varied: ResNet50 (2048), MobileNet (1024),

and VGG-16 (4608). This experiment was aimed at providing a fair comparison with the expander-enhanced architectures.

Classifier Trained in a Siamese-Network Configuration Plus Expander (CSN+ESN)

For the third experiment, we used the previously trained classifiers from the second experiment and retrofitted them with expanders. Only the expanders were trained while keeping the classifier’s weights fixed. This mimics a scenario where a pre-trained classifier is enhanced with an expander for efficiency.

Jointly Trained Classifier and Expander ((C+E)SN) In the final experiment, we trained both the classifier and the expander from scratch, using the Siamese network configuration. Initially, we expected this joint training to yield the best results, but the outcomes were dependent on the classifier architecture.

For both the CSN+ESN and (C+E)SN experiments, we conducted two types of error-correction decoding: (a) fixed-distance-based decoding, and (b) LDLC decoding, with results reported for both.

6.4.3 Training of the Classifiers

VGG-16 For VGG-16, all layers were trained in both the baseline and Siamese configurations to achieve better generalization.

ResNet50 In the Siamese configuration, the first 143 layers were frozen, while the remaining 32 layers were trained to implement transfer learning. For the baseline model, the entire architecture was trained.

MobileNet For MobileNet, in the Siamese configuration, only the last 23 layers were trained. In the baseline model, all layers were trained.

Equal Error Rates (EER) and AUROC for Various Architectures						
Experiment	ResNet		MobileNet		VGG-16	
	EER	AUROC	EER	AUROC	EER	AUROC
BC	0.032	0.972	0.022	0.969	0.040	0.940
CSN	0.065	0.982	0.055	0.989	0.014	0.999
CSN+ESN Distance	0.054	0.986	0.044	0.994	0.013	0.990
CSN+ESN LDLC	0.053	0.987	0.025	0.995	0.013	0.999
(C+E)SN Distance	0.066	0.984	0.042	0.993	0.020	0.997
(C+E)SN LDLC	0.065	0.984	0.039	0.993	0.020	0.995

Table 6.4.1: Equal error rates (EER) and areas under the receiver operating characteristics (ROC) curve (AUROC) for our experiments using the FVC2006 database, with different neural network architectures.

6.4.4 Results

We identified two potential sources of performance degradation in the NFE retrofit: the inclusion of the expander and the use of an error correction code instead of distance-based decoding. We systematically analyzed these sources.

The Effect of the Expander We compared the results of the BC and CSN experiments with the NFE-retrofitted architectures (CSN+ESN and (C+E)SN) using distance-based decoding. For both datasets, the results are shown in Tables 6.4.1 and 6.4.2. While some performance degradation occurred in specific configurations, the results indicate that training the expander separately generally performed better than joint training.

The Effect of the Decoder We compared the performance between distance-based decoding and LDLC decoding for both CSN+ESN and (C+E)SN experiments. Overall, the LDLC decoder slightly outperformed the distance-based decoder, with significant improvements observed for MobileNet on the FVC2006 dataset.

Equal Error Rates (EER) and AUROC for Various Architectures						
Experiment	ResNet		MobileNet		VGG-16	
	EER	AUROC	EER	AUROC	EER	AUROC
BC	0.031	0.958	0.017	0.979	0.062	0.972
CSN	0.035	0.995	0.025	0.996	0.013	0.999
CSN+ESN Distance	0.037	0.992	0.026	0.995	0.007	0.999
CSN+ESN LDLC	0.037	0.992	0.025	0.995	0.007	0.999
(C+E)SN Distance	0.059	0.985	0.054	0.986	0.033	0.993
(C+E)SN LDLC	0.058	0.985	0.054	0.981	0.034	0.981

Table 6.4.2: Equal error rates (EER) and areas under the receiver operating characteristics (ROC) curve (AUROC) for our experiments using the PolyU database, with different neural network architectures.

Overall Effect of the NFE Retrofit Comparing the best-performing results from the BC and CSN experiments with those from CSN+ESN LDLC and (C+E)SN LDLC, we observed minimal performance degradation due to the NFE. This encouraging result supports further exploration of NFEs in other biometric authentication architectures.

Training and Running Times The training and running times for the experiments on both datasets are provided in Tables 6.4.3 and 6.4.4. We noted that adding the expander introduced minimal overhead, but switching from baseline to Siamese networks increased both training and runtime significantly. The decoding times were not included, but improvements in LDLC decoding efficiency are noted for future work.

Comparison with Plain Fuzzy Extractors In terms of security and cost, plain fuzzy extractors are comparable to NFEs, though NFEs involve a more complex decoding process. However, the security guarantees and storage requirements for both systems are similar, with NFEs offering a more secure and efficient alternative for biometric authentication systems.

Experiment		ResNet50	MobileNet	VGG-16
BC	Training	7m11s (60 ep)	7m14s (80 ep)	8m48s (70 ep)
	Runtime	70ms	48ms	55ms
CSN	Training	1h53m22s (5000 ep)	1h06m12s (5000 ep)	1h05m39s (5000 ep)
	Runtime	529ms	264ms	173ms
CSN+ESN	Training	1h50m20s (5000 ep)	1h08m13s (5000 ep)	1h00m41s (5000 ep)
	Runtime	645ms	343ms	134ms
(C+E)SN	Training	1h57m00s (5000 ep)	1h06m30s (5000 ep)	1h08m48s (5000 ep)
	Runtime	589ms	325ms	189ms

Table 6.4.3: Training and running times for our four experiments, under each of the three architectures, when using the FVC2006 database.

Experiment		ResNet50	MobileNet	VGG-16
BC	Training	27m50s (50 ep)	21m38s (40 ep)	22m35s (50 ep)
	Runtime	81ms	55ms	63ms
CSN	Training	3h12m21s (4000 ep)	2h10m08s (4000 ep)	3h26m43s (4000 ep)
	Runtime	579ms	214ms	151ms
CSN+ESN	Training	3h17m23s (4000 ep)	2h08m06s (4000 ep)	3h09m32s (4000 ep)
	Runtime	459ms	227ms	160ms
(C+E)SN	Training	3h12m45s (4000 ep)	2h08m46s (4000 ep)	3h22m00s (4000 ep)
	Runtime	755ms	240ms	156ms

Table 6.4.4: Training and running times for our four experiments, under each of the three architectures, when using the PolyU database.

6.5 Conclusions

In this paper, we introduced a secure architecture for biometric user authentication (or identification) that minimizes the risk of storing sensitive information, such as neural network weights or biometric data, which could potentially be exploited by malicious actors to create synthetic biometric inputs capable of bypassing authentication checks. The proposed architecture follows the modern approach to securing user passwords. This system, referred to as a *neural fuzzy extractor* (NFE), integrates a classifier with a fuzzy extractor using an additional neural network component called an *expander*. The expander can be trained either concurrently with the classifier or afterward. NFEs can be retrofitted onto any well-established classifier, combining the robust classification abilities of artificial neural networks with the security features of fuzzy extractors.

Our implementation of the NFE for fingerprint-based authentication demonstrates how an existing classifier can be modified for NFE compatibility, and how two distinct training approaches can achieve output-space clustering in spherical regions required by NFEs. The experimental results reveal that the integration of the NFE introduces minimal performance degradation, as measured by equal error rates and areas under the receiver operating characteristic (ROC) curves. However, the system does experience higher overhead in terms of training and runtime, largely due to the Siamese network configuration. Despite this, we believe that the additional 1 to 3 hours required for training is justified by the enhanced security provided by NFEs. Moreover, the increase of 100 to 700 milliseconds in runtime remains within an acceptable range for practical applications.

In the future, our work will explore (1) the application of NFEs to other biometric modalities and (2) the development and training of expanders designed for coding over binary extension fields.

6.6 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1619201.

Chapter 7

Other Project II: Few-Shot Classification in Bioacoustics

This chapter is under review in literature as

Abhishek Jana et al. - An Automated Pipeline for Few-Shot Bird Call Classification, a Case Study with the Tooth-Billed Pigeon.

7.1 Introduction

Amidst a global crisis of biodiversity loss, this research presents a novel approach to wildlife monitoring, focusing on the critically endangered Tooth-billed Pigeon (*Didunculus stri-girostris*) of Samoa. A few-shot classification pipeline, utilizing machine learning, is proposed to detect and categorize bird vocalizations, addressing the issues of data scarcity and imbalanced classes that conventional models face.

This research integrates advanced audio processing, feature extraction, and clustering analysis to identify bird species with minimal training data. The study assesses a variety of deep learning architectures, such as convolutional neural networks and transformers, to generate resilient feature embeddings. Furthermore, dimensionality reduction techniques and

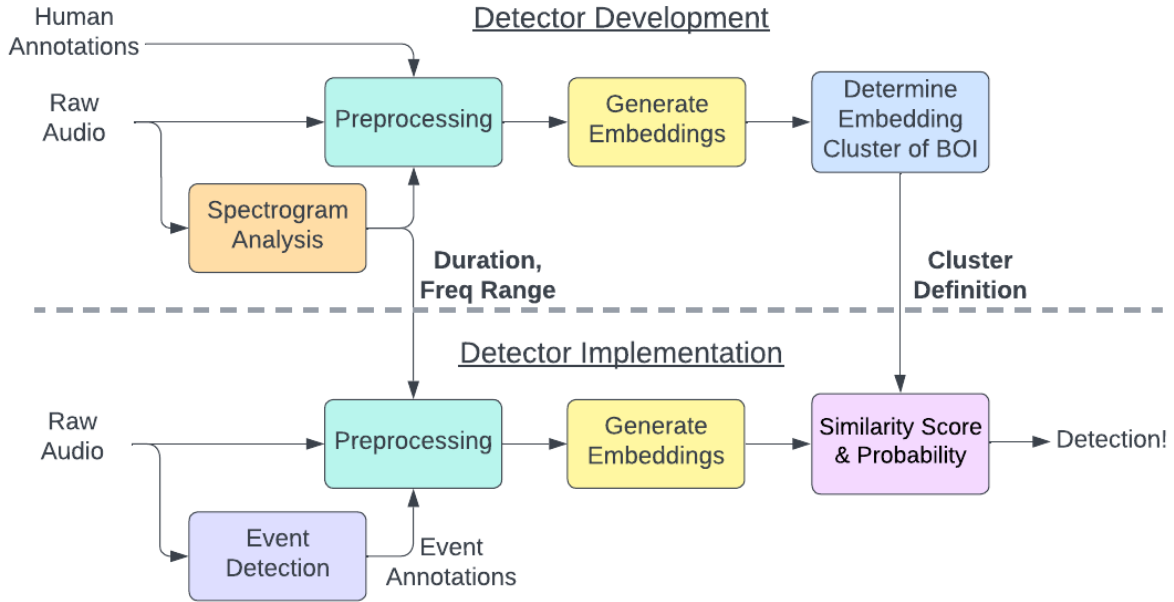


Figure 7.2.1: Overview of the classification pipeline, showcasing the key steps from data acquisition to final bird call detection.

similarity-based classification methods are explored to optimize rare species detection.

By combining AI with conservation biology, this work aims to demonstrate how interdisciplinary methods can address urgent environmental concerns. The findings offer practical solutions for improving wildlife monitoring techniques and set the stage for future applications of few-shot and few-shot learning in ecological and conservation sciences. The ultimate goal is to improve detection accuracy for rare species, aiding conservation efforts aimed at preserving the Tooth-billed Pigeon.

7.2 Methods

Our study developed a few-shot classification pipeline for rare bird call detection, focusing on the critically endangered Tooth-billed Pigeon (*Didunculus strigirostris*). Figure 7.2.1 presents an overview of the classification pipeline, outlining both the development process and the intended use of the final detector.

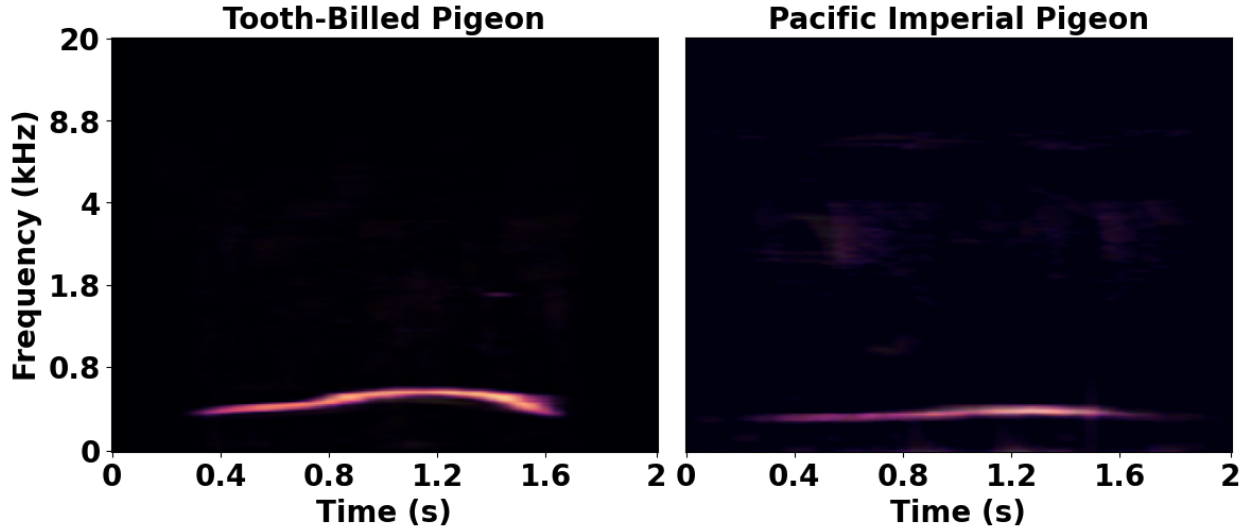


Figure 7.2.2: Average spectrograms of vocalizations from the TBP (left) and PIP (right). The TBP shows a distinct low-frequency call with a clear curvature, while the PIP displays a flatter frequency distribution, indicating a broader and less variable frequency range.

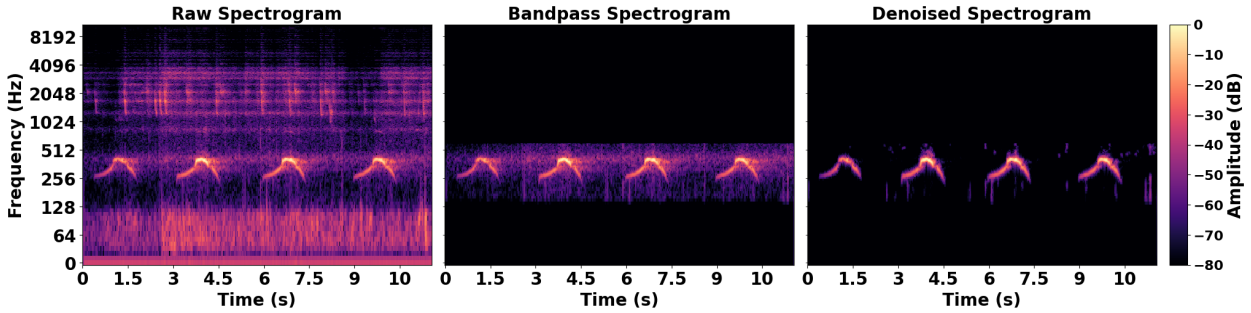


Figure 7.2.3: Comparison of raw, bandpass filtered, and denoised spectrograms of a TBP bird call. Spectrograms are shown on a (0 - 80)dB scale. The raw spectrogram (left) displays the full frequency range of the original recording. The bandpass spectrogram (center) isolates the target frequency band of 150 Hz - 600 Hz, removing noise from other frequencies. The denoised spectrogram (right) further reduces the background noise, highlighting the essential features of the bird call.

Data Collection: Audio recordings from the Samoan Conservation Society and Xencanto were utilized, focusing on six bird species: Crimson-crowned Fruit Dove (CCFD), Pacific Imperial Pigeon (PIP), White-throated Pigeon (WTP), Tooth-billed Pigeon (TBP), Polynesian Wattled Honeyeater (PWH), and Samoan Starling (SS).

Preprocessing: Audio files were resampled to 22.05 kHz and filtered using a bandpass range of 155-674 Hz for the Tooth-billed Pigeon. We applied noise reduction techniques

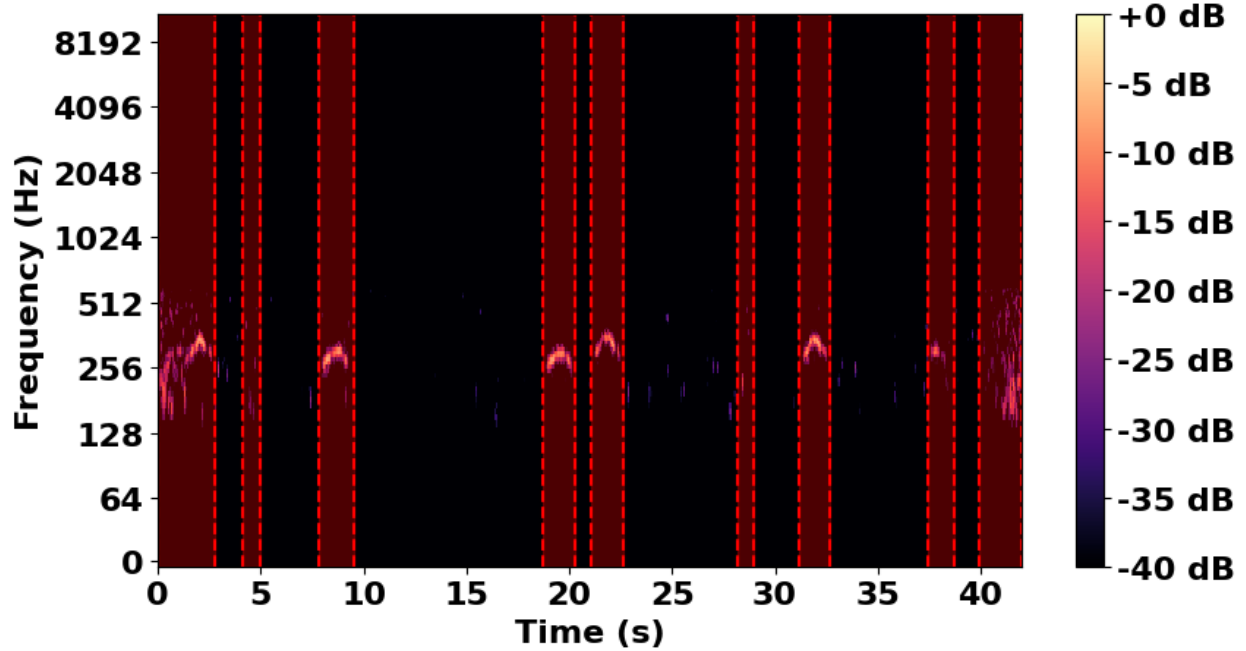


Figure 7.2.4: Spectrogram of the uncertain audio samples after applying a bandpass filter and noise reduction. The red-highlighted regions indicate the call’s detected vocalization events.

and segmented calls into 2-second clips. Figure 7.2.2 shows the average spectrograms of vocalizations from the TBP and PIP, illustrating the distinct frequency patterns of each species. Mel-spectrograms were generated with 227 mel bands and converted to a decibel scale with a -40 dB threshold. Figure 7.2.3 demonstrates the preprocessing steps, comparing raw, bandpass filtered, and denoised spectrograms of a TBP bird call.

Feature Embedding Extraction: Fourteen convolutional neural networks and two transformer models, including specialized bird call models such as Perch²⁶¹ and BirdNET²⁶², were evaluated. These models produced feature embeddings from the processed spectrograms.

Clustering Analysis: Principal Component Analysis (PCA) was used for dimensionality reduction, retaining 95% of the variance. The clustering performance was assessed using metrics like the Silhouette score, Davies-Bouldin index, and Calinski-Harabasz index.

Event Detection: A wavelet-based method identified vocalization events, utilizing an adaptive threshold at 10% of the mean feature value, with a minimum event duration of 0.5 seconds. Figure 7.2.4 displays a spectrogram of preprocessed audio samples with detected vocalization events highlighted in red.

Classification: A cosine similarity-based approach was used, comparing embeddings with prototypes from known samples. A softmax function transformed similarity scores into probabilities for species prediction.

Evaluation: Model performance was measured using precision, recall, and F1 score, with a primary focus on recall due to the rarity of the Tooth-billed Pigeon. The priority was to minimize false negatives.

This methodology provides an efficient solution for detecting rare species with limited data, overcoming challenges related to imbalanced datasets and scarce recordings.

7.3 Results

The comparison of various deep-learning models for bird call classification revealed substantial performance variations. The Perch model exhibited superior performance in detecting the Tooth-billed Pigeon, with the best clustering results and classification accuracy.

Table 7.3.1 summarizes the clustering performance metrics for each model. The Perch model achieved the highest overall score of 0.99, followed by BirdNET at 0.72. These scores were derived from normalized Silhouette, Davies-Bouldin, and Dunn indices.

Figure 7.3.1, a heatmap visualizing these normalized scores, illustrates the relative performance of each model across metrics. This visualization clearly shows the Perch model's superiority in all three clustering quality indices.

Scatter plot matrices of the top 5 PCA components for Perch and BirdNET models (Figures 7.3.2 and 7.3.3, respectively) revealed distinct clustering patterns corresponding to different bird species. These visualizations offer insight into how well each model separates

Perch	1	0.96	1
BirdNET	0.8	0.85	0.52
ResNet18	0.73	1	0.04
EfficientNet-B0	0.64	0.87	0.23
ResNet50	0.53	0.85	0.24
EfficientNet-B2	0.6	0.82	0.18
DenseNet	0.33	0.78	0.47
EfficientNet-B1	0.62	0.94	0
VGG	0.57	0.91	0.039
ResNet34	0.39	0.83	0.088
MobileNet	0.36	0.76	0.17
EfficientNet-B3	0.39	0.76	0.1
ResNeXt	0.18	0.78	0.16
ViT	0.048	0	0.95
CVIT	0	0.19	0.37
	Silh	DB	Dunn

Figure 7.3.1: Normalized Scores Heatmap of Clustering Quality Indices for various Deep Learning Models. The heatmap presents the performance of models across three key metrics: Silhouette Score (Silh), Davies-Bouldin Index (DB), and Dunn Index (Dunn). The values have been normalized for comparability.

Performance Metrics of Various Models Using Clustering Quality Indices					
Model	Overall Score	Silhouette Score	Davies-Bouldin Index	Dunn Index	N Components
Perch	0.99	0.21	1.87	0.41	142
BirdNET	0.72	0.19	2.04	0.32	130
ResNet18	0.59	0.18	1.81	0.23	62
EfficientNet-B0	0.58	0.18	2.02	0.27	82
ResNet50	0.54	0.16	2.05	0.27	92
EfficientNet-B2	0.53	0.17	2.11	0.26	95
DenseNet	0.53	0.14	2.16	0.31	76
EfficientNet-B1	0.52	0.17	1.90	0.23	76
VGG	0.51	0.17	1.95	0.23	87
ResNet34	0.44	0.15	2.08	0.24	69
MobileNet	0.43	0.15	2.19	0.26	98
EfficientNet-B3	0.42	0.15	2.20	0.25	102
ResNeXt	0.37	0.13	2.17	0.26	116
ViT	0.33	0.11	3.43	0.40	236
CVIT	0.19	0.11	3.12	0.30	141

Table 7.3.1: Performance Metrics of Various Models Using Clustering Quality Indices. The table summarizes the overall clustering performance for different deep-learning models. 'N Components' is the number of components used in PCA.

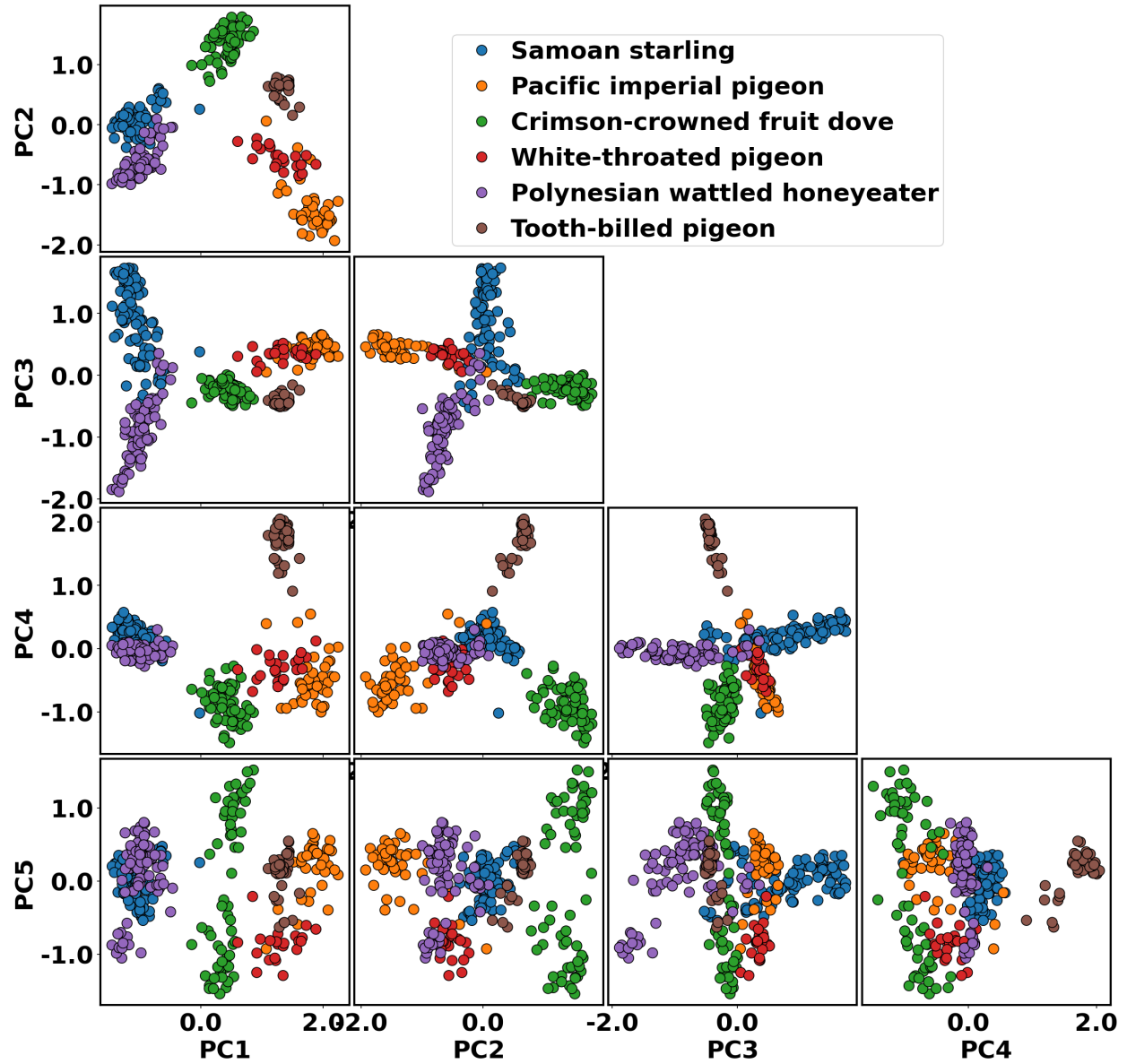


Figure 7.3.2: Scatter plot matrix of the top 5 PCA components of Perch model

Evaluation Metrics for Tooth-billed Pigeon (TBP)				
Model	F1 Score	Precision	Recall	Accuracy
Perch	0.6	0.5	0.75	0.86
BirdNET	0.44	0.4	0.5	0.82

Table 7.3.2: Evaluation Metrics for Tooth-billed Pigeon (TBP) Using BirdNET and Perch Models.

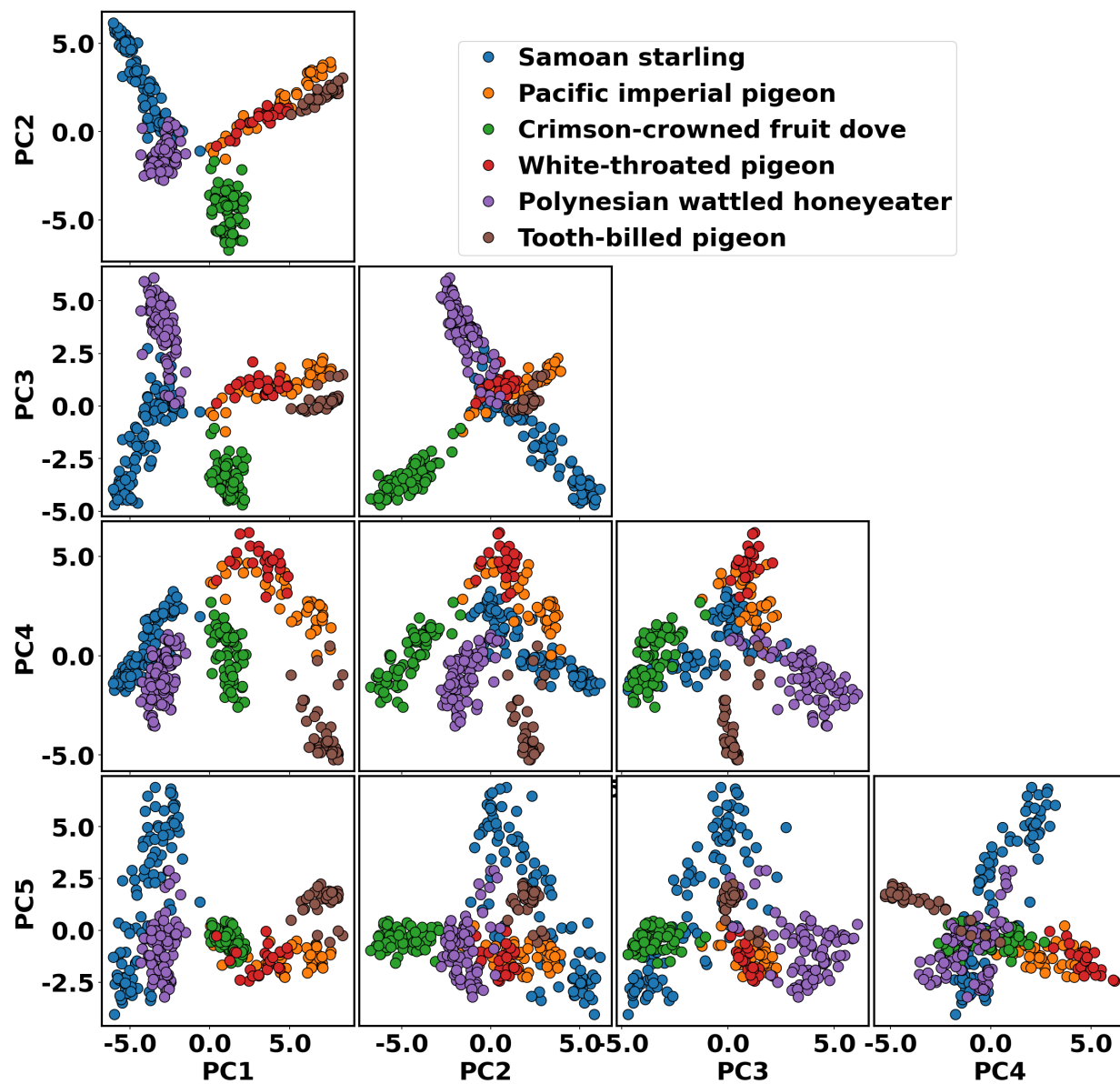


Figure 7.3.3: Scatter plot matrix of the top 5 PCA components of BirdNET model

different bird calls in the embedding space.

For TBP detection, we implemented a classification confidence threshold of 45% above equal probability distribution (approximately 0.24 for six species). Table 7.3.2 presents the evaluation metrics for the Perch and BirdNET models in identifying TBP. The Perch model exhibited superior performance with a recall of 0.75, precision of 0.50, F1 score of 0.60, and overall accuracy of 0.86. BirdNET showed lower but still notable performance with a recall of 0.50, precision of 0.40, F1 score of 0.44, and accuracy of 0.82.

These results underscore the Perch model’s effectiveness in minimizing false negatives, a crucial factor in rare species detection while maintaining acceptable precision levels for conservation applications.

7.4 Conclusion

This thesis presents an automated few-shot classification pipeline for detecting bird calls, focusing on the critically endangered Tooth-billed Pigeon. The Perch model showed notable success in detecting rare bird vocalizations with limited data.

Key outcomes include:

1. The Perch model’s recall of 0.75 for Tooth-billed Pigeon detection prioritized the reduction of false negatives.
2. Preprocessing techniques like bandpass filtering and noise reduction substantially improved the quality of audio inputs.
3. Lowering the classification threshold increased the detection rate of rare species, demonstrating the need for a balanced approach between precision and recall in conservation work.
4. PCA-based dimensionality reduction effectively preserved crucial features while improving model efficiency.

This work provides a valuable tool for conservationists and lays the groundwork for future research in few-shot learning for wildlife monitoring. Further research should investigate alternative classification techniques and distance metrics to enhance the accuracy of species detection, contributing to global efforts in biodiversity conservation.

Chapter 8

Conclusion

The journey through this dissertation has explored a range of applications of machine learning techniques, from their integration with traditional statistical methods in the galaxy-halo connection to innovative uses in other interdisciplinary fields. Beginning with a deep dive into the fundamental clustering statistics in Chapter 2, the groundwork was laid for understanding how galaxies populate dark matter halos. The progression into machine learning methods in Chapter 3 opened new doors, demonstrating the powerful potential of these techniques in modeling complex cosmic relationships that would be difficult to capture with conventional methods alone.

Chapter 4 showed how the Halo Occupation Distribution (HOD) model, a cornerstone in galaxy-halo studies, can be enhanced with machine learning to constrain galaxy clustering parameters more efficiently. The results of these studies, combined with the extensive analysis of the AbacusSummit simulations in Chapter 5, highlighted not only the accuracy but also the versatility of these methods in handling large-scale cosmological datasets. These advancements reflect the growing role that computational tools play in cosmology, offering new insights into the structure and evolution of the universe.

Chapters 6 and 7 took a departure from the cosmology-centric focus, delving into unrelated fields: deep learning for fingerprint verification and few-shot classification in bioacoustics. Though these topics might seem disconnected from the galaxy-halo connection, there

is a broader lesson to be drawn. The deep learning methods applied to secure fingerprint verification and the few-shot classification of bird species illustrate how adaptable machine learning techniques are across diverse fields. These projects can serve as a source of inspiration for cosmologists, encouraging the exploration of similar approaches in their research. For instance, the methods used for few-shot learning in bioacoustics could inspire new strategies for handling sparse or unlabeled data in cosmological surveys, where efficient classification of galaxies or other cosmic structures could face similar challenges.

As this dissertation concludes, it is clear that the fusion of machine learning with cosmological analysis has set the stage for future discoveries. The techniques developed and applied here can be expanded and refined, not only within the field of galaxy-halo studies but also across a wider spectrum of scientific disciplines. The final chapters serve as a reminder that innovation often comes from unexpected places, and the adaptability of machine learning offers endless possibilities for researchers aiming to push the boundaries of what is known.

Bibliography

- [1] Risa H. Wechsler and Jeremy L. Tinker. The Connection Between Galaxies and Their Dark Matter Halos. *Annual Review of Astron and Astrophys*, 56:435–487, September 2018. doi: 10.1146/annurev-astro-081817-051756.
- [2] Nina A. Maksimova, Lehman H. Garrison, Daniel J. Eisenstein, Boryana Hadzhiyska, Sownak Bose, and Thomas P. Satterthwaite. ABACUSSUMMIT: a massive set of high-accuracy, high-resolution N-body simulations. *Monthly Notices of the RAS*, 508 (3):4017–4037, December 2021. doi: 10.1093/mnras/stab2484.
- [3] 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. Combet, 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. Girard, 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, G. Prézeau, S. Prunet, J. L. Puget, J. P. Rachen, W. T. Reach, R. Rebolo, M. Reinecke, M. Remazeilles, C. Renault, A. Renzi, I. Ristorcelli, G. Rocha, C. Rosset, M. Rossetti, G. Roudier, B. Rouillé d'Orfeuil, M. Rowan-Robinson, J. A. Rubiño-Martín, B. Rusholme, N. Said, V. Salvatelli, L. Salvati, M. Sandri, D. Santos, M. Savelainen, G. Savini, D. Scott, M. D. Seiffert, P. Serra, E. P. S. Shellard, L. D. Spencer, M. Spinelli, V. Stolyarov, R. Stompor, R. Sudiwala, R. Sunyaev, D. Sutton, A. S. Suur-Uski, J. F. Sygnet, J. A. Tauber, L. Terenzi, L. Toffolatti, M. Tomasi, M. Tristram, T. Trombetti, M. Tucci, J. Tuovinen, M. Türlér, G. Umana, L. Valenziano, J. Valiviita, F. Van Tent, P. Vielva, F. Villa, 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 2015 results. XIII. Cosmological parameters. *Astronomy and Astrophysics*, 594:A13, September 2016. doi: 10.1051/0004-6361/201525830.

- [4] Phillip James Edwin Peebles. *The large-scale structure of the universe*, volume 96.

Princeton university press, 1980.

- [5] M. Davis and P. J. E. Peebles. A survey of galaxy redshifts. V. The two-point position and velocity correlations. *Astrophysical Journal*, 267:465–482, April 1983. doi: 10.1086/160884.
- [6] N. A. Bahcall and R. M. Soneira. The spatial correlation function of rich clusters of galaxies. *Astrophysical Journal*, 270:20–38, July 1983. doi: 10.1086/161094.
- [7] A. A. Klypin and A. I. Kopylov. The Spatial Covariance Function for Rich Clusters of Galaxies. *Soviet Astronomy Letters*, 9:41–44, February 1983.
- [8] N. Kaiser. On the spatial correlations of Abell clusters. *Astrophysical Journal, Letters*, 284:L9–L12, September 1984. doi: 10.1086/184341.
- [9] J. M. Bardeen, J. R. Bond, N. Kaiser, and A. S. Szalay. The Statistics of Peaks of Gaussian Random Fields. *Astrophysical Journal*, 304:15, May 1986. doi: 10.1086/164143.
- [10] H. J. Mo and S. D. M. White. An analytic model for the spatial clustering of dark matter haloes. *Monthly Notices of the RAS*, 282(2):347–361, September 1996. doi: 10.1093/mnras/282.2.347.
- [11] Anatoly Klypin, Joel Primack, and Jon Holtzman. Small-Scale Power Spectrum and Correlations in Lambda + Cold Dark Matter Models. *Astrophysical Journal*, 466:13, July 1996. doi: 10.1086/177489.
- [12] A. Jenkins, C. S. Frenk, F. R. Pearce, P. A. Thomas, J. M. Colberg, S. D. M. White, H. M. P. Couchman, J. A. Peacock, G. Efstathiou, and A. H. Nelson. Evolution of Structure in Cold Dark Matter Universes. *Astrophysical Journal*, 499(1):20–40, May 1998. doi: 10.1086/305615.
- [13] C. M. Baugh. The real-space correlation function measured from the APM Galaxy

- Survey. *Monthly Notices of the RAS*, 280(1):267–275, May 1996. doi: 10.1093/mnras/280.1.267.
- [14] Andrey V. Kravtsov and Anatoly A. Klypin. The Origin and Evolution of Halo Bias in Linear and Nonlinear Regimes. *Astrophysical Journal*, 520(2):437–453, August 1999. doi: 10.1086/307495.
 - [15] Pedro Colín, Anatoly A. Klypin, Andrey V. Kravtsov, and Alexei M. Khokhlov. Evolution of Bias in Different Cosmological Models. *Astrophysical Journal*, 523(1):32–53, September 1999. doi: 10.1086/307710.
 - [16] Matthew Colless, Gavin Dalton, Steve Maddox, Will Sutherland, Peder Norberg, Shaun Cole, Joss Bland-Hawthorn, Terry Bridges, Russell Cannon, Chris Collins, Warrick Couch, Nicholas Cross, Kathryn Deeley, Roberto De Propris, Simon P. Driver, George Efstathiou, Richard S. Ellis, Carlos S. Frenk, Karl Glazebrook, Carole Jackson, Ofer Lahav, Ian Lewis, Stuart Lumsden, Darren Madgwick, John A. Peacock, Bruce A. Peterson, Ian Price, Mark Seaborne, and Keith Taylor. The 2dF Galaxy Redshift Survey: spectra and redshifts. *Monthly Notices of the RAS*, 328(4):1039–1063, December 2001. doi: 10.1046/j.1365-8711.2001.04902.x.
 - [17] Donald G. York, J. Adelman, Jr. Anderson, John E., 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 Ruthmanskorf, 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. *Astronomical Journal*, 120(3):1579–1587, September 2000. doi: 10.1086/301513.

- [18] Kurt L. Adelberger, Charles C. Steidel, Mauro Giavalisco, Mark Dickinson, Max Pettini, and Melinda Kellogg. A Counts-in-Cells Analysis Of Lyman-break Galaxies At Redshift $Z \sim 3$. *Astrophysical Journal*, 505(1):18–24, September 1998. doi: 10.1086/306162.
- [19] DESI Collaboration, B. Abareshi, J. Aguilar, S. Ahlen, Shadab Alam, David M. Alexander, R. Alfarsy, L. Allen, C. Allende Prieto, O. Alves, J. Ameer, E. Armengaud, J. Asorey, Alejandro Aviles, S. Bailey, A. Balaguera-Antolínez, O. Ballester, C. Baltay, A. Bault, S. F. Beltran, B. Benavides, S. BenZvi, A. Berti, R. Besuner, Florian Beutler,

D. Bianchi, C. Blake, P. Blanc, R. Blum, A. Bolton, S. Bose, D. Bramall, S. Brieden, A. Brodzeller, D. Brooks, C. Brownell, E. Buckley-Geer, R. N. Cahn, Z. Cai, R. Canning, R. Capasso, A. Carnero Rosell, P. Carton, R. Casas, F. J. Castander, J. L. Cervantes-Cota, S. Chabanier, E. Chaussidon, C. Chuang, C. Circosta, S. Cole, A. P. Cooper, L. da Costa, M. C. Cousinou, A. Cuceu, T. M. Davis, K. Dawson, R. de la Cruz-Noriega, A. de la Macorra, A. de Mattia, J. Della Costa, P. Demmer, M. Derwent, A. Dey, B. Dey, G. Dhungana, Z. Ding, C. Dobson, P. Doel, J. Donald-McCann, J. Donaldson, K. Douglass, Y. Duan, P. Dunlop, J. Edelstein, S. Eftekharzadeh, D. J. Eisenstein, M. Enriquez-Vargas, S. Escoffier, M. Evatt, P. Fagrelus, X. Fan, K. Fanning, V. A. Fawcett, S. Ferraro, J. Ereza, B. Flaugh, A. Font-Ribera, J. E. Forero-Romero, C. S. Frenk, S. Fromenteau, B. T. Gänsicke, C. Garcia-Quintero, L. Garrison, E. Gaztañaga, F. Gerardi, H. Gil-Marín, S. Gontcho A Gontcho, Alma X. Gonzalez-Morales, G. Gonzalez-de-Rivera, V. Gonzalez-Perez, C. Gordon, O. Graur, D. Green, C. Grove, D. Gruen, G. Gutierrez, J. Guy, C. Hahn, S. Harris, D. Herrera, Hiram K. Herrera-Alcantar, K. Honscheid, C. Howlett, D. Huterer, V. Iršič, M. Ishak, P. Jelsky, L. Jiang, J. Jimenez, Y. P. Jing, R. Joyce, E. Jullo, S. Juneau, N. G. Karaçaylı, M. Karamanis, A. Karcher, T. Karim, R. Kehoe, S. Kent, D. Kirkby, T. Kisner, F. Kitaura, S. E. Koposov, A. Kovács, A. Kremin, Alex Krolewski, B. L’Huillier, O. Lahav, A. Lambert, C. Lamman, Ting-Wen Lan, M. Landriau, S. Lane, D. Lang, J. U. Lange, J. Lasker, L. Le Guillou, A. Leauthaud, A. Le Van Suu, Michael E. Levi, T. S. Li, C. Magneville, M. Manera, Christopher J. Manser, B. Marshall, Paul Martini, W. McCollam, P. McDonald, Aaron M. Meisner, J. Mena-Fernández, J. Meneses-Rizo, M. Mezcua, T. Miller, R. Miquel, P. Montero-Camacho, J. Moon, J. Moustakas, E. Mueller, Andrea Muñoz-Gutiérrez, Adam D. Myers, S. Nadathur, J. Najita, L. Napolitano, E. Neilsen, Jeffrey A. Newman, J. D. Nie, Y. Ning, G. Niz, P. Norberg, Hernán E. Noriega, T. O’Brien, A. Obuljen, N. Palanque-Delabrouille, A. Palmese, P. Zhiwei, D. Pappalardo, X. PENG, W. J. Percival, S. Perruchot, R. Pogge, C. Poppett, A. Porredon, F. Prada, J. Prochaska, R. Pucha, A. Pérez-Fernández, I. Pérez-Ràfols, D. Rabinowitz, A. Raichoor, S. Ramirez-Solano, César

- Ramírez-Pérez, C. Ravoux, K. Reil, M. Rezaie, A. Rocher, C. Rockosi, N. A. Roe, A. Roodman, A. J. Ross, G. Rossi, R. Ruggeri, V. Ruhlmann-Kleider, C. G. Sabiu, S. Gaines, K. Said, A. Saintonge, Javier Salas Catonga, L. Samushia, E. Sanchez, C. Saulder, E. Schaan, E. Schlafly, D. Schlegel, J. Schmoll, D. Scholte, M. Schubnell, A. Secroun, H. Seo, S. Serrano, Ray M. Sharples, Michael J. Sholl, Joseph Harry Silber, D. R. Silva, M. Sirk, M. Siudek, A. Smith, D. Sprayberry, R. Staten, B. Stupak, T. Tan, Gregory Tarlé, Suk Sien Tie, R. Tojeiro, L. A. Ureña-López, F. Valdes, O. Valenzuela, M. Valluri, M. Vargas-Magaña, L. Verde, M. Walther, B. Wang, M. S. Wang, B. A. Weaver, C. Weaverdyck, R. Wechsler, Michael J. Wilson, J. Yang, Y. Yu, S. Yuan, Christophe Yèche, H. Zhang, K. Zhang, Cheng Zhao, Rongpu Zhou, Zhimin Zhou, H. Zou, J. Zou, S. Zou, and Y. Zu. Overview of the Instrumentation for the Dark Energy Spectroscopic Instrument. *Astronomical Journal*, 164(5):207, November 2022. doi: 10.3847/1538-3881/ac882b.
- [20] Chung-Pei Ma and J. N. Fry. Deriving the Nonlinear Cosmological Power Spectrum and Bispectrum from Analytic Dark Matter Halo Profiles and Mass Functions. *Astrophysical Journal*, 543(2):503–513, November 2000. doi: 10.1086/317146.
- [21] J. A. Peacock and R. E. Smith. Halo occupation numbers and galaxy bias. *Monthly Notices of the RAS*, 318(4):1144–1156, November 2000. doi: 10.1046/j.1365-8711.2000.03779.x.
- [22] Uroš Seljak. Analytic model for galaxy and dark matter clustering. *Monthly Notices of the RAS*, 318(1):203–213, October 2000. doi: 10.1046/j.1365-8711.2000.03715.x.
- [23] Asantha Cooray and Ravi Sheth. Halo models of large scale structure. *Physics Reports*, 372(1):1–129, December 2002. doi: 10.1016/S0370-1573(02)00276-4.
- [24] M. Betoule, R. Kessler, J. Guy, J. Mosher, D. Hardin, R. Biswas, P. Astier, P. El-Hage, M. König, 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. Pritchett, M. Richmond, A. G. Riess, V. Ruhlmann-Kleider, M. Sako, K. Schahmanche, 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. *Astronomy and Astrophysics*, 568:A22, August 2014. doi: 10.1051/0004-6361/201423413.
- [25] 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. Frieman, 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. Mar-

shall, 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. *Physical Review D*, 98(4):043526, August 2018. doi: 10.1103/PhysRevD.98.043526.

- [26] 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. *Monthly Notices of the RAS*, 470(3):2617–2652, September 2017. doi: 10.1093/mnras/stx721.
- [27] C. S. Frenk and S. D. M. White. Dark matter and cosmic structure. *Annalen der Physik*, 524(9-10):507–534, October 2012. doi: 10.1002/andp.201200212.
- [28] Michael Kuhlen, Mark Vogelsberger, and Raul Angulo. Numerical simulations of the dark universe: State of the art and the next decade. *Physics of the Dark Universe*, 1(1-2):50–93, November 2012. doi: 10.1016/j.dark.2012.10.002.
- [29] Rachel S. Somerville and Romeel Davé. Physical Models of Galaxy Formation in a Cosmological Framework. *Annual Review of Astron and Astrophys*, 53:51–113, August 2015. doi: 10.1146/annurev-astro-082812-140951.
- [30] Thorsten Naab and Jeremiah P. Ostriker. Theoretical Challenges in Galaxy Formation. *Annual Review of Astron and Astrophys*, 55(1):59–109, August 2017. doi: 10.1146/annurev-astro-081913-040019.
- [31] Andrey V. Kravtsov and Stefano Borgani. Formation of Galaxy Clusters. *Annual Review of Astron and Astrophys*, 50:353–409, September 2012. doi: 10.1146/annurev-astro-081811-125502.
- [32] David H. Weinberg, Michael J. Mortonson, Daniel J. Eisenstein, Christopher Hirata, Adam G. Riess, and Eduardo Rozo. Observational probes of cosmic acceleration. *Physics Reports*, 530(2):87–255, September 2013. doi: 10.1016/j.physrep.2013.05.001.
- [33] James S. Bullock and Michael Boylan-Kolchin. Small-Scale Challenges to the Λ CDM Paradigm. *Annual Review of Astron and Astrophys*, 55(1):343–387, August 2017. doi: 10.1146/annurev-astro-091916-055313.

- [34] 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. *Monthly Notices of the RAS*, 446(1): 521–554, January 2015. doi: 10.1093/mnras/stu2058.
- [35] 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. *Monthly Notices of the RAS*, 475(1):648–675, March 2018. doi: 10.1093/mnras/stx3112.
- [36] Philip F. Hopkins. Galaxies on FIRE: Stellar Feedback Explains Inefficient Star Formation. In *American Astronomical Society Meeting Abstracts #224*, volume 224 of *American Astronomical Society Meeting Abstracts*, page 215.06, June 2014.
- [37] 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. *Monthly Notices of the RAS*, 326(2):649–666, September 2001. doi: 10.1046/j.1365-8711.2001.04616.x.
- [38] Martin White, Lars Hernquist, and Volker Springel. The Halo Model and Numerical Simulations. *Astrophysical Journal, Letters*, 550(2):L129–L132, April 2001. doi: 10.1086/319644.
- [39] 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. *Astrophysical Journal*, 593(1):1–25, August 2003. doi: 10.1086/376517.

- [40] A. J. Benson, S. Cole, C. S. Frenk, C. M. Baugh, and C. G. Lacey. The nature of galaxy bias and clustering. *Monthly Notices of the RAS*, 311(4):793–808, February 2000. doi: 10.1046/j.1365-8711.2000.03101.x.
- [41] 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. *Astrophysical Journal*, 633(2):791–809, November 2005. doi: 10.1086/466510.
- [42] 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. *Monthly Notices of the RAS*, 423(4):3458–3473, July 2012. doi: 10.1111/j.1365-2966.2012.21142.x.
- [43] 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. *Monthly Notices of the RAS*, 460(3):3100–3118, August 2016. doi: 10.1093/mnras/stw1225.
- [44] Simon D. M. White and Carlos S. Frenk. Galaxy Formation through Hierarchical Clustering. *Astrophysical Journal*, 379:52, September 1991. doi: 10.1086/170483.
- [45] G. Kauffmann, S. D. M. White, and B. Guiderdoni. The formation and evolution of galaxies within merging dark matter haloes. *Monthly Notices of the RAS*, 264:201–218, September 1993. doi: 10.1093/mnras/264.1.201.
- [46] Rachel S. Somerville and Joel R. Primack. Semi-analytic modelling of galaxy formation: the local Universe. *Monthly Notices of the RAS*, 310(4):1087–1110, December 1999. doi: 10.1046/j.1365-8711.1999.03032.x.
- [47] Shaun Cole, Cedric G. Lacey, Carlton M. Baugh, and Carlos S. Frenk. Hierarchical

- galaxy formation. *Monthly Notices of the RAS*, 319(1):168–204, November 2000. doi: 10.1046/j.1365-8711.2000.03879.x.
- [48] 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. *Monthly Notices of the RAS*, 370(2):645–655, August 2006. doi: 10.1111/j.1365-2966.2006.10519.x.
- [49] Hong Guo, Idit Zehavi, Zheng Zheng, David H. Weinberg, Andreas A. Berlind, Michael Blanton, Yanmei Chen, Daniel J. Eisenstein, Shirley Ho, Eyal Kazin, Marc Manera, Claudia Maraston, Cameron K. McBride, Sebastián E. Nuza, Nikhil Padmanabhan, John K. Parejko, Will J. Percival, Ashley J. Ross, Nicholas P. Ross, Lado Samushia, Ariel G. Sánchez, David J. Schlegel, Donald P. Schneider, Ramin A. Skibba, Molly E. C. Swanson, Jeremy L. Tinker, Rita Tojeiro, David A. Wake, Martin White, Neta A. Bahcall, Dmitry Bizyaev, Howard Brewington, Kevin Bundy, Luiz N. A. da Costa, Garrett Ebelke, Elena Malanushenko, Viktor Malanushenko, Daniel Oravetz, Graziano Rossi, Audrey Simmons, Stephanie Snedden, Alina Streblyanska, and Daniel Thomas. The Clustering of Galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: Luminosity and Color Dependence and Redshift Evolution. *Astrophysical Journal*, 767(2):122, April 2013. doi: 10.1088/0004-637X/767/2/122.
- [50] Bruno M. B. Henriques, Peter A. Thomas, Seb Oliver, and Isaac Roseboom. Monte Carlo Markov Chain parameter estimation in semi-analytic models of galaxy formation. *Monthly Notices of the RAS*, 396(1):535–547, June 2009. doi: 10.1111/j.1365-2966.2009.14730.x.
- [51] Yu Lu, H. J. Mo, Martin D. Weinberg, and Neal Katz. A Bayesian approach to the semi-analytic model of galaxy formation: methodology. *Monthly Notices of the RAS*, 416(3):1949–1964, September 2011. doi: 10.1111/j.1365-2966.2011.19170.x.
- [52] Yu Lu, Risa H. Wechsler, Rachel S. Somerville, Darren Croton, Lauren Porter, Joel Primack, Peter S. Behroozi, Henry C. Ferguson, David C. Koo, Yicheng Guo, Mohammadtaher Safarzadeh, Kristian Finlator, Marco Castellano, Catherine E. White,

- Veronica Sommariva, and Chris Moody. Semi-analytic Models for the CANDELS Survey: Comparison of Predictions for Intrinsic Galaxy Properties. *Astrophysical Journal*, 795(2):123, November 2014. doi: 10.1088/0004-637X/795/2/123.
- [53] Rachel S. Somerville, Rudy C. Gilmore, Joel R. Primack, and Alberto Domínguez. Galaxy properties from the ultraviolet to the far-infrared: Λ cold dark matter models confront observations. *Monthly Notices of the RAS*, 423(3):1992–2015, July 2012. doi: 10.1111/j.1365-2966.2012.20490.x.
- [54] 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. *Monthly Notices of the RAS*, 451(3):2663–2680, August 2015. doi: 10.1093/mnras/stv705.
- [55] Darren J. Croton, Adam R. H. Stevens, Chiara Tonini, Thibault Garel, Maksym Bernyk, Antonio Bibiano, Luke Hodkinson, Simon J. Mutch, Gregory B. Poole, and Genevieve M. Shattow. Semi-Analytic Galaxy Evolution (SAGE): Model Calibration and Basic Results. *Astrophysical Journal, Supplement*, 222(2):22, February 2016. doi: 10.3847/0067-0049/222/2/22.
- [56] 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. *Astrophysical Journal*, 609(1):35–49, July 2004. doi: 10.1086/420959.
- [57] A. Vale and J. P. Ostriker. Linking halo mass to galaxy luminosity. *Monthly Notices of the RAS*, 353(1):189–200, September 2004. doi: 10.1111/j.1365-2966.2004.08059.x.
- [58] Argyro Tasitsiomi, Andrey V. Kravtsov, Risa H. Wechsler, and Joel R. Primack. Modeling Galaxy-Mass Correlations in Dissipationless Simulations. *Astrophysical Journal*, 614(2):533–546, October 2004. doi: 10.1086/423784.

- [59] Charlie Conroy, Risa H. Wechsler, and Andrey V. Kravtsov. Modeling Luminosity-dependent Galaxy Clustering through Cosmic Time. *Astrophysical Journal*, 647(1): 201–214, August 2006. doi: 10.1086/503602.
- [60] Peter S. Behroozi, Charlie Conroy, and Risa H. Wechsler. A Comprehensive Analysis of Uncertainties Affecting the Stellar Mass-Halo Mass Relation for $0 < z < 4$. *Astrophysical Journal*, 717(1):379–403, July 2010. doi: 10.1088/0004-637X/717/1/379.
- [61] Andreas A. Berlind and David H. Weinberg. The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass. *Astrophysical Journal*, 575(2):587–616, August 2002. doi: 10.1086/341469.
- [62] Risa H. Wechsler, Rachel S. Somerville, James S. Bullock, Tsafir S. Kolatt, Joel R. Primack, George R. Blumenthal, and Avishai Dekel. Galaxy Formation at $z \sim 3$: Constraints from Spatial Clustering. *Astrophysical Journal*, 554(1):85–103, June 2001. doi: 10.1086/321373.
- [63] Román Scoccimarro, Ravi K. Sheth, Lam Hui, and Bhuvnesh Jain. How Many Galaxies Fit in a Halo? Constraints on Galaxy Formation Efficiency from Spatial Clustering. *Astrophysical Journal*, 546(1):20–34, January 2001. doi: 10.1086/318261.
- [64] James S. Bullock, Risa H. Wechsler, and Rachel S. Somerville. Galaxy halo occupation at high redshift. *Monthly Notices of the RAS*, 329(1):246–256, January 2002. doi: 10.1046/j.1365-8711.2002.04959.x.
- [65] Andrew P. Hearin, Andrew R. Zentner, Frank C. van den Bosch, Duncan Campbell, and Erik Tollerud. Introducing decorated HODs: modelling assembly bias in the galaxy-halo connection. *Monthly Notices of the RAS*, 460(3):2552–2570, August 2016. doi: 10.1093/mnras/stw840.
- [66] Andrew P. Hearin and Douglas F. Watson. The dark side of galaxy colour. *Monthly Notices of the RAS*, 435(2):1313–1324, October 2013. doi: 10.1093/mnras/stt1374.

- [67] Ying Zu and Rachel Mandelbaum. Mapping stellar content to dark matter haloes - II. Halo mass is the main driver of galaxy quenching. *Monthly Notices of the RAS*, 457(4):4360–4383, April 2016. doi: 10.1093/mnras/stw221.
- [68] Ying Zu and Rachel Mandelbaum. Mapping stellar content to dark matter haloes - III. Environmental dependence and conformity of galaxy colours. *Monthly Notices of the RAS*, 476(2):1637–1653, May 2018. doi: 10.1093/mnras/sty279.
- [69] Shadab Alam, John A. Peacock, Katarina Kraljic, Ashley J. Ross, and Johan Comparat. Multitracer extension of the halo model: probing quenching and conformity in eBOSS. *Mon. Not. Roy. Astron. Soc.*, 497(1):581–595, 2020. doi: 10.1093/mnras/staa1956.
- [70] Idit Zehavi, Zheng Zheng, David H. Weinberg, Michael R. Blanton, Neta A. Bahcall, Andreas A. Berlind, Jon Brinkmann, Joshua A. Frieman, James E. Gunn, Robert H. Lupton, Robert C. Nichol, Will J. Percival, Donald P. Schneider, Ramin A. Skibba, Michael A. Strauss, Max Tegmark, and Donald G. York. Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity. *Astrophysical Journal*, 736(1):59, July 2011. doi: 10.1088/0004-637X/736/1/59.
- [71] Martin White, M. Blanton, A. Bolton, D. Schlegel, J. Tinker, A. Berlind, L. da Costa, E. Kazin, Y. T. Lin, M. Maia, C. K. McBride, N. Padmanabhan, J. Parejko, W. Percival, F. Prada, B. Ramos, E. Sheldon, F. de Simoni, R. Skibba, D. Thomas, D. Wake, I. Zehavi, Z. Zheng, R. Nichol, Donald P. Schneider, Michael A. Strauss, B. A. Weaver, and David H. Weinberg. The Clustering of Massive Galaxies at $z \sim 0.5$ from the First Semester of BOSS Data. *Astrophysical Journal*, 728(2):126, February 2011. doi: 10.1088/0004-637X/728/2/126.
- [72] Zhongxu Zhai, Jeremy L. Tinker, ChangHoon Hahn, Hee-Jong Seo, Michael R. Blanton, Rita Tojeiro, Hugo O. Camacho, Marcos Lima, Aurelio Carnero Rosell, Flavia Sobreira, Luiz N. da Costa, Julian E. Bautista, Joel R. Brownstein, Johan Comparat, Kyle Dawson, Jeffrey A. Newman, Abhishek Prakash, Alexandre Roman-Lopes,

- and Donald P. Schneider. The Clustering of Luminous Red Galaxies at $z \sim 0.7$ from EBOSS and BOSS Data. *Astrophysical Journal*, 848(2):76, October 2017. doi: 10.3847/1538-4357/aa8eee.
- [73] Shadab Alam, John A. Peacock, Katarina Kraljic, Ashley J. Ross, and Johan Comparat. Multitracer extension of the halo model: probing quenching and conformity in eBOSS. *Monthly Notices of the RAS*, 497(1):581–595, September 2020. doi: 10.1093/mnras/staa1956.
- [74] Shadab Alam, Marie Aubert, Santiago Avila, Christophe Balland, Julian E. Bautista, Matthew A. Bershad, 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, Héliion du Mas des Bourboux, Stephanie Escoffier, Thomas Etourneau, James Farr, Andreu Font-Ribera, Peter M. Frinchaboy, Sebastien 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. *Physical Review D*, 103(8):083533, April 2021. doi: 10.1103/PhysRevD.103.083533.
- [75] Rongpu Zhou, Jeffrey A. Newman, Yao-Yuan Mao, Aaron Meisner, John Moustakas, Adam D. Myers, Abhishek Prakash, Andrew R. Zentner, David Brooks, Yutong Duan, Martin Landriau, Michael E. Levi, Francisco Prada, and Gregory Tarle. The clustering of DESI-like luminous red galaxies using photometric redshifts. *Monthly Notices of the RAS*, 501(3):3309–3331, March 2021. doi: 10.1093/mnras/staa3764.
- [76] Benjamin D. Wibking, Andrés N. Salcedo, David H. Weinberg, Lehman H. Garrison, Douglas Ferrer, Jeremy Tinker, Daniel Eisenstein, Marc Metchnik, and Philip Pinto. Emulating galaxy clustering and galaxy-galaxy lensing into the deeply non-linear regime: methodology, information, and forecasts. *Monthly Notices of the RAS*, 484(1):989–1006, March 2019. doi: 10.1093/mnras/sty2258.
- [77] Edo van Uitert, Marcello Cacciato, Henk Hoekstra, Margot Brouwer, Cristóbal Sifón, Massimo Viola, Ivan Baldry, Joss Bland-Hawthorn, Sarah Brough, M. J. I. Brown, Ami Choi, Simon P. Driver, Thomas Erben, Catherine Heymans, Hendrik Hildebrandt, Benjamin Joachimi, Konrad Kuijken, Jochen Liske, Jon Loveday, John McFarland, Lance Miller, Reiko Nakajima, John Peacock, Mario Radovich, A. S. G. Robotham, Peter Schneider, Gert Sikkema, Edward N. Taylor, and Gijs Verdoes Kleijn. The stellar-to-halo mass relation of GAMA galaxies from 100 deg² of KiDS weak lensing data. *Monthly Notices of the RAS*, 459(3):3251–3270, July 2016. doi: 10.1093/mnras/stw747.
- [78] Chiaki Hikage, Masamune Oguri, Takashi Hamana, Surhud More, Rachel Mandelbaum, Masahiro Takada, Fabian Köhlinger, Hironao Miyatake, Atsushi J. Nishizawa, Hiroaki Aihara, Robert Armstrong, James Bosch, Jean Coupon, Anne Ducout, Paul Ho, Bau-Ching Hsieh, Yutaka Komiyama, François Lanusse, Alexie Leauthaud, Robert H. Lupton, Elinor Medezinski, Sogo Mineo, Shoken Miyama, Satoshi Miyazaki,

- Ryoma Murata, Hitoshi Murayama, Masato Shirasaki, Cristóbal Sifón, Melanie Simet, Joshua Speagle, David N. Spergel, Michael A. Strauss, Naoshi Sugiyama, Masayuki Tanaka, Yousuke Utsumi, Shiang-Yu Wang, and Yoshihiko Yamada. Cosmology from cosmic shear power spectra with Subaru Hyper Suprime-Cam first-year data. *Publications of the ASJ*, 71(2):43, April 2019. doi: 10.1093/pasj/psz010.
- [79] S. Avila, V. Gonzalez-Perez, F. G. Mohammad, A. de Mattia, C. Zhao, A. Raichoor, A. Tamone, S. Alam, J. Bautista, D. Bianchi, E. Burtin, M. J. Chapman, C. H. Chuang, J. Comparat, K. Dawson, T. Divers, H. du Mas des Bourboux, H. Gil-Marin, E. M. Mueller, S. Habib, K. Heitmann, V. Ruhlmann-Kleider, N. Padilla, W. J. Percival, A. J. Ross, H. J. Seo, D. P. Schneider, and G. Zhao. The Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: exploring the halo occupation distribution model for emission line galaxies. *Monthly Notices of the RAS*, 499(4): 5486–5507, December 2020. doi: 10.1093/mnras/staa2951.
- [80] Jonathan Richardson, Zheng Zheng, Suchetana Chatterjee, Daisuke Nagai, and Yue Shen. The Halo Occupation Distribution of SDSS Quasars. *Astrophysical Journal*, 755(1):30, August 2012. doi: 10.1088/0004-637X/755/1/30.
- [81] 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. *Monthly Notices of the RAS*, 511(4):5492–5524, April 2022. doi: 10.1093/mnras/stac390.
- [82] Andrew R. Zentner, Andrew Hearin, Frank C. van den Bosch, Johannes U. Lange, and Antonia Sierra Villarreal. Constraints on assembly bias from galaxy clustering. *Monthly Notices of the RAS*, 485(1):1196–1209, May 2019. doi: 10.1093/mnras/stz470.
- [83] Idit Zehavi, Sergio Contreras, Nelson Padilla, Nicholas J. Smith, Carlton M. Baugh,

- and Peder Norberg. The Impact of Assembly Bias on the Galaxy Content of Dark Matter Halos. *Astrophysical Journal*, 853(1):84, January 2018. doi: 10.3847/1538-4357/aaa54a.
- [84] Risa H. Wechsler, Andrew R. Zentner, James S. Bullock, Andrey V. Kravtsov, and Brandon Allgood. The Dependence of Halo Clustering on Halo Formation History, Concentration, and Occupation. *Astrophysical Journal*, 652(1):71–84, November 2006. doi: 10.1086/507120.
- [85] Sownak Bose, Carlos S. Frenk, Adrian Jenkins, Azadeh Fattahi, Facundo A. Gómez, Robert J. J. Grand, Federico Marinacci, Julio F. Navarro, Kyle A. Oman, Rüdiger Pakmor, Joop Schaye, Christine M. Simpson, and Volker Springel. No cores in dark matter-dominated dwarf galaxies with bursty star formation histories. *Monthly Notices of the RAS*, 486(4):4790–4804, July 2019. doi: 10.1093/mnras/stz1168.
- [86] Hong Guo, Zheng Zheng, Idit Zehavi, Peter S. Behroozi, Chia-Hsun Chuang, Johan Comparat, Ginevra Favole, Stefan Gottloeber, Anatoly Klypin, Francisco Prada, David H. Weinberg, and Gustavo Yepes. Redshift-space clustering of SDSS galaxies - luminosity dependence, halo occupation distribution, and velocity bias. *Monthly Notices of the RAS*, 453(4):4368–4383, November 2015. doi: 10.1093/mnras/stv1966.
- [87] Thomas J. Armitage, David J. Barnes, Scott T. Kay, Yannick M. Bahé, Claudio Dalla Vecchia, Robert A. Crain, and Tom Theuns. The Cluster-EAGLE project: velocity bias and the velocity dispersion-mass relation of cluster galaxies. *Monthly Notices of the RAS*, 474(3):3746–3759, March 2018. doi: 10.1093/mnras/stx3020.
- [88] Guinevere Kauffmann, Cheng Li, Wei Zhang, and Simone Weinmann. A re-examination of galactic conformity and a comparison with semi-analytic models of galaxy formation. *Monthly Notices of the RAS*, 430(2):1447–1456, April 2013. doi: 10.1093/mnras/stt007.
- [89] Jeremy L. Tinker, ChangHoon Hahn, Yao-Yuan Mao, Andrew R. Wetzel, and Char-

- lie Conroy. Halo histories versus galaxy properties at $z = 0$ II: large-scale galactic conformity. *Monthly Notices of the RAS*, 477(1):935–945, June 2018. doi: 10.1093/mnras/sty666.
- [90] Victor F. Calderon, Andreas A. Berlind, and Manodeep Sinha. Small- and large-scale galactic conformity in SDSS DR7. *Monthly Notices of the RAS*, 480(2):2031–2045, October 2018. doi: 10.1093/mnras/sty2000.
- [91] Martin Kerscher, István Szapudi, and Alexander S. Szalay. A Comparison of Estimators for the Two-Point Correlation Function. *Astrophysical Journal, Letters*, 535(1): L13–L16, May 2000. doi: 10.1086/312702.
- [92] M. Ryleigh Davis and P J E Peebles. A survey of galaxy redshifts. v. the two-point position and velocity correlations. *The Astrophysical Journal*, 267:465–482, 1983. URL <https://api.semanticscholar.org/CorpusID:122320623>.
- [93] P. C. Hewett. The estimation of galaxy angular correlation functions. *Monthly Notices of the RAS*, 201:867–883, December 1982. doi: 10.1093/mnras/201.4.867.
- [94] A. J. S. Hamilton. Toward Better Ways to Measure the Galaxy Correlation Function. *Astrophysical Journal*, 417:19, November 1993. doi: 10.1086/173288.
- [95] Stephen D. Landy and Alexander S. Szalay. Bias and Variance of Angular Correlation Functions. *Astrophysical Journal*, 412:64, July 1993. doi: 10.1086/172900.
- [96] István Szapudi and Alexander S. Szalay. A New Class of Estimators for the N-Point Correlations. *Astrophysical Journal, Letters*, 494(1):L41–L44, February 1998. doi: 10.1086/311146.
- [97] E. Keihänen, V. Lindholm, P. Monaco, L. Blot, C. Carbone, K. Kiiveri, A. G. Sánchez, A. Viitanen, J. Valiviita, A. Amara, N. Auricchio, M. Baldi, D. Bonino, E. Branchini, M. Brescia, J. Brinchmann, S. Camera, V. Capobianco, J. Carretero, M. Castellano, S. Cavauoti, A. Cimatti, R. Cledassou, G. Congedo, L. Conversi, Y. Copin, L. Corcione,

- M. Cropper, A. Da Silva, H. Degaudenzi, M. Douspis, F. Dubath, C. A. J. Duncan, X. Dupac, S. Dusini, A. Ealet, S. Farrens, S. Ferriol, M. Frailis, E. Franceschi, M. Fu-
mana, B. Gillis, C. Giocoli, A. Grazian, F. Grupp, L. Guzzo, S. V. H. Haugan, H. Hoek-
stra, W. Holmes, F. Hormuth, K. Jahnke, M. Kümmel, S. Kermiche, A. Kiessling,
T. Kitching, M. Kunz, H. Kurki-Suonio, S. Ligorì, P. B. Lilje, I. Lloro, E. Maiorano,
O. Mansutti, O. Marggraf, F. Marulli, R. Massey, M. Melchior, M. Meneghetti,
G. Meylan, M. Moresco, B. Morin, L. Moscardini, E. Munari, S. M. Niemi, C. Padilla,
S. Paltani, F. Pasian, K. Pedersen, V. Pettorino, S. Pires, G. Polenta, M. Poncet,
L. Popa, F. Raison, A. Renzi, J. Rhodes, E. Romelli, R. Saglia, B. Sartoris, P. Schnei-
der, T. Schrabback, A. Secroun, G. Seidel, C. Sirignano, G. Sirri, L. Stanco, C. Surace,
P. Tallada-Crespí, D. Tavagnacco, A. N. Taylor, I. Tereno, R. Toledo-Moreo, F. Tor-
radeflot, E. A. Valentijn, L. Valenziano, T. Vassallo, Y. Wang, J. Weller, G. Zamorani,
J. Zoubian, S. Andreon, D. Maino, and S. de la Torre. Euclid: Fast two-point corre-
lation function covariance through linear construction. *Astronomy and Astrophysics*,
666:A129, October 2022. doi: 10.1051/0004-6361/202244065.
- [98] Idit Zehavi, Zheng Zheng, David H. Weinberg, Joshua A. Frieman, Andreas A. Berlind,
Michael R. Blanton, Roman Scoccimarro, Ravi K. Sheth, Michael A. Strauss, Issha
Kayo, Yasushi Suto, Masataka Fukugita, Osamu Nakamura, Neta A. Bahcall, Jon
Brinkmann, James E. Gunn, Greg S. Hennessy, Željko Ivezić, Gillian R. Knapp, Jon
Loveday, Avery Meiksin, David J. Schlegel, Donald P. Schneider, Istvan Szapudi, Max
Tegmark, Michael S. Vogeley, Donald G. York, and SDSS Collaboration. The Luminos-
ity and Color Dependence of the Galaxy Correlation Function. *Astrophysical Journal*,
630(1):1–27, September 2005. doi: 10.1086/431891.
- [99] Esteban Jiménez, Sergio Contreras, Nelson Padilla, Idit Zehavi, Carlton M. Baugh,
and Violeta Gonzalez-Perez. Extensions to the halo occupation distribution model for
more accurate clustering predictions. *Monthly Notices of the RAS*, 490(3):3532–3544,
December 2019. doi: 10.1093/mnras/stz2790.

- [100] Chris Blake, Adrian Collister, and Ofer Lahav. Halo-model signatures from 380000 Sloan Digital Sky Survey luminous red galaxies with photometric redshifts. *Monthly Notices of the RAS*, 385(3):1257–1269, April 2008. doi: 10.1111/j.1365-2966.2007.11925.x.
- [101] Kevork Abazajian, Zheng Zheng, Idit Zehavi, David H. Weinberg, Joshua A. Frieman, Andreas A. Berlind, Michael R. Blanton, Neta A. Bahcall, J. Brinkmann, Donald P. Schneider, and Max Tegmark. Cosmology and the Halo Occupation Distribution from Small-Scale Galaxy Clustering in the Sloan Digital Sky Survey. *Astrophysical Journal*, 625(2):613–620, June 2005. doi: 10.1086/429685.
- [102] Florian Beutler, Chris Blake, Matthew Colless, D. Heath Jones, Lister Staveley-Smith, Lachlan Campbell, Quentin Parker, Will Saunders, and Fred Watson. The 6dF Galaxy Survey: dependence of halo occupation on stellar mass. *Monthly Notices of the RAS*, 429(4):3604–3618, March 2013. doi: 10.1093/mnras/sts637.
- [103] Ramin A. Skibba, M. Stephen M. Smith, Alison L. Coil, John Moustakas, James Aird, Michael R. Blanton, Aaron D. Bray, Richard J. Cool, Daniel J. Eisenstein, Alexander J. Mendez, Kenneth C. Wong, and Guangtun Zhu. PRIMUS: Galaxy Clustering as a Function of Luminosity and Color at $0.2 < z < 1$. *Astrophysical Journal*, 784(2):128, April 2014. doi: 10.1088/0004-637X/784/2/128.
- [104] F. Marulli, M. Bolzonella, E. Branchini, I. Davidzon, S. de la Torre, B. R. Granett, L. Guzzo, A. Iovino, L. Moscardini, A. Pollo, U. Abbas, C. Adami, S. Arnouts, J. Bel, D. Bottini, A. Cappi, J. Coupon, O. Cucciati, G. De Lucia, A. Fritz, P. Franzetti, M. Fumana, B. Garilli, O. Ilbert, J. Krywult, V. Le Brun, O. Le Fèvre, D. Maccagni, K. Malek, H. J. McCracken, L. Paioro, M. Polletta, H. Schlagenhauser, M. Scodreggio, L. A. M. Tasca, R. Tojeiro, D. Vergani, A. Zanicelli, A. Burden, C. Di Porto, A. Marchetti, C. Marinoni, Y. Mellier, R. C. Nichol, J. A. Peacock, W. J. Percival, S. Phleps, M. Wolk, and G. Zamorani. The VIMOS Public Extragalactic Redshift Survey (VIPERS) . Luminosity and stellar mass dependence of galaxy cluster-

- ing at $0.5 < z < 1.1$. *Astronomy and Astrophysics*, 557:A17, September 2013. doi: 10.1051/0004-6361/201321476.
- [105] Yue Shen, Michael A. Strauss, Masamune Oguri, Joseph F. Hennawi, Xiaohui Fan, Gordon T. Richards, Patrick B. Hall, James E. Gunn, Donald P. Schneider, Alexander S. Szalay, Anirudda R. Thakar, Daniel E. Vanden Berk, Scott F. Anderson, Neta A. Bahcall, Andrew J. Connolly, and Gillian R. Knapp. Clustering of High-Redshift ($z \approx 2.9$) Quasars from the Sloan Digital Sky Survey. *Astronomical Journal*, 133(5): 2222–2241, May 2007. doi: 10.1086/513517.
- [106] Nicholas P. Ross, Yue Shen, Michael A. Strauss, Daniel E. Vanden Berk, Andrew J. Connolly, Gordon T. Richards, Donald P. Schneider, David H. Weinberg, Patrick B. Hall, Neta A. Bahcall, and Robert J. Brunner. Clustering of Low-redshift ($z \approx 2.2$) Quasars from the Sloan Digital Sky Survey. *Astrophysical Journal*, 697(2):1634–1655, June 2009. doi: 10.1088/0004-637X/697/2/1634.
- [107] 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, Hélión 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. *Monthly Notices of the RAS*, 500(1):736–762, January 2021. doi: 10.1093/mnras/staa2800.
- [108] 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 Sheldon, 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. *Monthly Notices of the RAS*, 426(4):2719–2737, November 2012. doi: 10.1111/j.1365-2966.2012.21779.x.
- [109] Seshadri Nadathur, Alex Woodfinden, Will J. Percival, Marie Aubert, Julian Bautista, Kyle Dawson, Stéphanie Escoffier, Sebastien Fromenteau, Héctor Gil-Marín, James Rich, Ashley J. Ross, Graziano Rossi, Mariana Vargas Magaña, Joel R. Brownstein, and Donald P. Schneider. The completed SDSS-IV extended baryon oscillation spectroscopic survey: geometry and growth from the anisotropic void-galaxy correlation function in the luminous red galaxy sample. *Monthly Notices of the RAS*, 499(3): 4140–4157, December 2020. doi: 10.1093/mnras/staa3074.
- [110] Joaquin Armijo, Carlton M. Baugh, Peder Norberg, and Nelson D. Padilla. A new test of gravity - II. Application of marked correlation functions to luminous red galaxy samples. *Monthly Notices of the RAS*, 528(4):6631–6636, March 2024. doi: 10.1093/mnras/stae449.
- [111] Joaquín Armijo, Yan-Chuan Cai, Nelson Padilla, Baojiu Li, and John A. Peacock. Testing modified gravity using a marked correlation function. *Monthly Notices of the RAS*, 478(3):3627–3632, August 2018. doi: 10.1093/mnras/sty1335.
- [112] Lucia A. Perez, Sangeeta Malhotra, James E. Rhoads, and Vithal Tilvi. Void Prob-

- ability Function of Simulated Surveys of High-redshift Ly α Emitters. *Astrophysical Journal*, 906(1):58, January 2021. doi: 10.3847/1538-4357/abc88b.
- [113] Jeremy L. Tinker, Charlie Conroy, Peder Norberg, Santiago G. Patiri, David H. Weinberg, and Michael S. Warren. Void Statistics in Large Galaxy Redshift Surveys: Does Halo Occupation of Field Galaxies Depend on Environment? *Astrophysical Journal*, 686(1):53–71, October 2008. doi: 10.1086/589983.
- [114] Darren J. Croton, Peder Norberg, Enrique Gaztañaga, and Carlton M. Baugh. Statistical analysis of galaxy surveys - III. The non-linear clustering of red and blue galaxies in the 2dFGRS. *Monthly Notices of the RAS*, 379(4):1562–1570, August 2007. doi: 10.1111/j.1365-2966.2007.12035.x.
- [115] Charlie Conroy, Alison L. Coil, Martin White, Jeffrey A. Newman, Renbin Yan, Michael C. Cooper, Brian F. Gerke, Marc Davis, and David C. Koo. The DEEP2 Galaxy Redshift Survey: The Evolution of Void Statistics from $z \sim 1$ to $z \sim 0$. *Astrophysical Journal*, 635(2):990–1005, December 2005. doi: 10.1086/497682.
- [116] Beth A. Reid, Will J. Percival, Daniel J. Eisenstein, Licia Verde, David N. Spergel, Ramin A. Skibba, Neta A. Bahcall, Tamas Budavari, Joshua A. Frieman, Masataka Fukugita, J. Richard Gott, James E. Gunn, Željko Ivezić, Gillian R. Knapp, Richard G. Kron, Robert H. Lupton, Timothy A. McKay, Avery Meiksin, Robert C. Nichol, Adrian C. Pope, David J. Schlegel, Donald P. Schneider, Chris Stoughton, Michael A. Strauss, Alexander S. Szalay, Max Tegmark, Michael S. Vogeley, David H. Weinberg, Donald G. York, and Idit Zehavi. Cosmological constraints from the clustering of the Sloan Digital Sky Survey DR7 luminous red galaxies. *Monthly Notices of the RAS*, 404(1):60–85, May 2010. doi: 10.1111/j.1365-2966.2010.16276.x.
- [117] Beth Ann Reid. *Applications of the halo model to large scale structure measurements of the Luminous Red Galaxies*. PhD thesis, Princeton University, New Jersey, January 2008.

- [118] Beth A. Reid and David N. Spergel. Constraining the Luminous Red Galaxy Halo Occupation Distribution Using Counts-In-Cylinders. *Astrophysical Journal*, 698(1):143–154, June 2009. doi: 10.1088/0004-637X/698/1/143.
- [119] Heather D. Berrier, Elizabeth J. Barton, Joel C. Berrier, James S. Bullock, Andrew R. Zentner, and Risa H. Wechsler. Counts-in-cylinders in the Sloan Digital Sky Survey with Comparisons to N-body Simulations. *Astrophysical Journal*, 726(1):1, January 2011. doi: 10.1088/0004-637X/726/1/1.
- [120] Douglas F. Watson, Andreas A. Berlind, Cameron K. McBride, David W. Hogg, and Tao Jiang. The Extreme Small Scales: Do Satellite Galaxies Trace Dark Matter? *Astrophysical Journal*, 749(1):83, April 2012. doi: 10.1088/0004-637X/749/1/83.
- [121] Hong Guo, Zheng Zheng, Peter S. Behroozi, Idit Zehavi, Chia-Hsun Chuang, Johan Comparat, Ginevra Favole, Stefan Gottloeber, Anatoly Klypin, Francisco Prada, Sergio A. Rodríguez-Torres, David H. Weinberg, and Gustavo Yepes. Modelling galaxy clustering: halo occupation distribution versus subhalo matching. *Monthly Notices of the RAS*, 459(3):3040–3058, July 2016. doi: 10.1093/mnras/stw845.
- [122] Zheng Zheng and Hong Guo. Accurate and efficient halo-based galaxy clustering modelling with simulations. *Monthly Notices of the RAS*, 458(4):4015–4024, June 2016. doi: 10.1093/mnras/stw523.
- [123] Ryan Hausen, Brant E. Robertson, Hanjue Zhu, Nickolay Y. Gnedin, Piero Madau, Evan E. Schneider, Bruno Villaseñor, and Nicole E. Drakos. Revealing the Galaxy-Halo Connection through Machine Learning. *Astrophysical Journal*, 945(2):122, March 2023. doi: 10.3847/1538-4357/acb25c.
- [124] Adam Gauci, Kristian Zarb Adami, and John Abela. Machine Learning for Galaxy Morphology Classification. *arXiv e-prints*, art. arXiv:1005.0390, May 2010. doi: 10.48550/arXiv.1005.0390.

- [125] Xiaoying Xu, Shirley Ho, Hy Trac, Jeff Schneider, Barnabas Póczos, and Michelle Ntampaka. A First Look at Creating Mock Catalogs with Machine Learning Techniques. *Astrophysical Journal*, 772(2):147, August 2013. doi: 10.1088/0004-637X/772/2/147.
- [126] Abhishek Jana and Lado Samushia. Constraining Galaxy Halo Connection Using Machine Learning. In *American Astronomical Society Meeting Abstracts*, volume 241 of *American Astronomical Society Meeting Abstracts*, page 306.11, January 2023.
- [127] M. Ntampaka, H. Trac, D. J. Sutherland, N. Battaglia, B. Póczos, and J. Schneider. A Machine Learning Approach for Dynamical Mass Measurements of Galaxy Clusters. *Astrophysical Journal*, 803(2):50, April 2015. doi: 10.1088/0004-637X/803/2/50.
- [128] Richard Stiskalek, Deaglan J. Bartlett, Harry Desmond, and Dhayaa Anbajagane. The scatter in the galaxy-halo connection: a machine learning analysis. *Monthly Notices of the RAS*, 514(3):4026–4045, August 2022. doi: 10.1093/mnras/stac1609.
- [129] A. Ćiprijanović, A. Lewis, K. Pedro, S. Madireddy, B. Nord, G. N. Perdue, and S. M. Wild. DeepAstroUDA: semi-supervised universal domain adaptation for cross-survey galaxy morphology classification and anomaly detection. *Machine Learning: Science and Technology*, 4(2):025013, June 2023. doi: 10.1088/2632-2153/acca5f.
- [130] Xinyue Zhang, Yanfang Wang, Wei Zhang, Yueqiu Sun, Siyu He, Gabriella Contardo, Francisco Villaescusa-Navarro, and Shirley Ho. From Dark Matter to Galaxies with Convolutional Networks. *arXiv e-prints*, art. arXiv:1902.05965, February 2019. doi: 10.48550/arXiv.1902.05965.
- [131] Harry George Chittenden and Rita Tojeiro. Modelling the galaxy-halo connection with semi-recurrent neural networks. *Monthly Notices of the RAS*, 518(4):5670–5692, January 2023. doi: 10.1093/mnras/stac3498.
- [132] Luisa Lucie-Smith, Hiranya V. Peiris, Andrew Pontzen, and Michelle Lochner. Machine

- learning cosmological structure formation. *Monthly Notices of the RAS*, 479(3):3405–3414, September 2018. doi: 10.1093/mnras/sty1719.
- [133] Benjamin P. Moster, Thorsten Naab, Magnus Lindström, and Joseph A. O’Leary. GalaxyNet: connecting galaxies and dark matter haloes with deep neural networks and reinforcement learning in large volumes. *Monthly Notices of the RAS*, 507(2): 2115–2136, October 2021. doi: 10.1093/mnras/stab1449.
- [134] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.
- [135] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- [136] Gilles Louppe. Understanding Random Forests: From Theory to Practice. *arXiv e-prints*, art. arXiv:1407.7502, July 2014. doi: 10.48550/arXiv.1407.7502.
- [137] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.
- [138] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.
- [139] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
- [140] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [141] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Number 4 in Information science and statistics. Springer, 2006. ISBN 9780387310732. doi: 10.1117/1.2819119. URL <http://www.library.wisc.edu/selectedtocs/bg0137.pdf>.
- [142] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009. URL <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>.

- [143] Harshil M. Kamdar, Matthew J. Turk, and Robert J. Brunner. Machine learning and cosmological simulations - I. Semi-analytical models. *Monthly Notices of the RAS*, 455(1):642–658, January 2016. doi: 10.1093/mnras/stv2310.
- [144] Harshil M. Kamdar, Matthew J. Turk, and Robert J. Brunner. Machine learning and cosmological simulations - II. Hydrodynamical simulations. *Monthly Notices of the RAS*, 457(2):1162–1179, April 2016. doi: 10.1093/mnras/stv2981.
- [145] Christopher C. Lovell, Sultan Hassan, Francisco Villaescusa-Navarro, Shy Genel, ChangHoon Hahn, Daniel Angles-Alcazar, James Kwon, Natali de Santi, Kartheik Iyer, Giulio Fabbian, and Greg Bryan. A Hierarchy of Normalizing Flows for Modelling the Galaxy-Halo Relationship. In *Machine Learning for Astrophysics*, page 21, July 2023. doi: 10.48550/arXiv.2307.06967.
- [146] Natalí S. M. de Santi, Natália V. N. Rodrigues, Antonio D. Montero-Dorta, L. Raul Abramo, Beatriz Tucci, and M. Celeste Artale. Mimicking the halo-galaxy connection using machine learning. *Monthly Notices of the RAS*, 514(2):2463–2478, August 2022. doi: 10.1093/mnras/stac1469.
- [147] Ana Maria Delgado, Digvijay Wadekar, Boryana Hadzhiyska, Sownak Bose, Lars Hernquist, and Shirley Ho. Modelling the galaxy-halo connection with machine learning. *Monthly Notices of the RAS*, 515(2):2733–2746, September 2022. doi: 10.1093/mnras/stac1951.
- [148] Ferah Munshi, Alyson M. Brooks, Elaad Applebaum, Charlotte R. Christensen, T. Quinn, and Serena Sligh. Quantifying Scatter in Galaxy Formation at the Lowest Masses. *Astrophysical Journal*, 923(1):35, December 2021. doi: 10.3847/1538-4357/ac0db6.
- [149] Kaustav Mitra, Frank C. van den Bosch, and Johannes U. Lange. BASILISK II. Improved constraints on the galaxy-halo connection from satellite kinematics in SDSS.

- Monthly Notices of the RAS*, 533(3):3647–3675, September 2024. doi: 10.1093/mnras/stae2030.
- [150] Volker Springel, Naoki Yoshida, and Simon D. M. White. GADGET: a code for collisionless and gasdynamical cosmological simulations. *New Astronomy*, 6(2):79–117, April 2001. doi: 10.1016/S1384-1076(01)00042-2.
 - [151] R. Teyssier. Cosmological hydrodynamics with adaptive mesh refinement. A new high resolution code called RAMSES. *Astronomy and Astrophysics*, 385:337–364, April 2002. doi: 10.1051/0004-6361:20011817.
 - [152] Volker Springel. The cosmological simulation code GADGET-2. *Monthly Notices of the RAS*, 364(4):1105–1134, December 2005. doi: 10.1111/j.1365-2966.2005.09655.x.
 - [153] Tomoaki Ishiyama, Toshiyuki Fukushige, and Junichiro Makino. GreeM: Massively Parallel TreePM Code for Large Cosmological N -body Simulations. *Publications of the ASJ*, 61:1319, December 2009. doi: 10.1093/pasj/61.6.1319.
 - [154] Volker Springel. E pur si muove: Galilean-invariant cosmological hydrodynamical simulations on a moving mesh. *Monthly Notices of the RAS*, 401(2):791–851, January 2010. doi: 10.1111/j.1365-2966.2009.15715.x.
 - [155] J. Bédorf, E. Gaburov, and S. Portegies Zwart. Bonsai: A GPU Tree-Code. In R. Capuzzo-Dolcetta, M. Limongi, and A. Tornambè, editors, *Advances in Computational Astrophysics: Methods, Tools, and Outcome*, volume 453 of *Astronomical Society of the Pacific Conference Series*, page 325, July 2012. doi: 10.48550/arXiv.1204.2280.
 - [156] R. E. Angulo, V. Springel, S. D. M. White, A. Jenkins, C. M. Baugh, and C. S. Frenk. Scaling relations for galaxy clusters in the Millennium-XXL simulation. *Monthly Notices of the RAS*, 426(3):2046–2062, November 2012. doi: 10.1111/j.1365-2966.2012.21830.x.
 - [157] P. Fosalba, E. Gaztañaga, F. J. Castander, and M. Crocce. The MICE Grand Challenge

- light-cone simulation - III. Galaxy lensing mocks from all-sky lensing maps. *Monthly Notices of the RAS*, 447(2):1319–1332, February 2015. doi: 10.1093/mnras/stu2464.
- [158] Anatoly Klypin, Gustavo Yepes, Stefan Gottlöber, Francisco Prada, and Steffen Heß. MultiDark simulations: the story of dark matter halo concentrations and density profiles. *Monthly Notices of the RAS*, 457(4):4340–4359, April 2016. doi: 10.1093/mnras/stw248.
- [159] Yohan Dubois, Sébastien Peirani, Christophe Pichon, Julien Devriendt, Raphaël Gavazzi, Charlotte Welker, and Marta Volonteri. The HORIZON-AGN simulation: morphological diversity of galaxies promoted by AGN feedback. *Monthly Notices of the RAS*, 463(4):3948–3964, December 2016. doi: 10.1093/mnras/stw2265.
- [160] Douglas Potter, Joachim Stadel, and Romain Teyssier. PKDGRAV3: beyond trillion particle cosmological simulations for the next era of galaxy surveys. *Computational Astrophysics and Cosmology*, 4(1):2, May 2017. doi: 10.1186/s40668-017-0021-1.
- [161] Joseph DeRose, Risa H. Wechsler, Jeremy L. Tinker, Matthew R. Becker, Yao-Yuan Mao, Thomas McClintock, Sean McLaughlin, Eduardo Rozo, and Zhongxu Zhai. The AEMULUS Project. I. Numerical Simulations for Precision Cosmology. *Astrophysical Journal*, 875(1):69, April 2019. doi: 10.3847/1538-4357/ab1085.
- [162] Katrin Heitmann, Hal Finkel, Adrian Pope, Vitali Morozov, Nicholas Frontiere, Salman Habib, Esteban Rangel, Thomas Uram, Danila Korytov, Hillary Child, Samuel Flenner, Joe Insley, and Silvio Rizzi. The Outer Rim Simulation: A Path to Many-core Supercomputers. *Astrophysical Journal, Supplement*, 245(1):16, November 2019. doi: 10.3847/1538-4365/ab4da1.
- [163] Volker Springel, Rüdiger Pakmor, Oliver Zier, and Martin Reinecke. Simulating cosmic structure formation with the GADGET-4 code. *Monthly Notices of the RAS*, 506(2): 2871–2949, September 2021. doi: 10.1093/mnras/stab1855.

- [164] Y. Dubois, C. Pichon, C. Welker, D. Le Borgne, J. Devriendt, C. Laigle, S. Codis, D. Pogosyan, S. Arnouts, K. Benabed, E. Bertin, J. Blaizot, F. Bouchet, J. F. Cardoso, S. Colombi, V. de Lapparent, V. Desjacques, R. Gavazzi, S. Kassin, T. Kimm, H. McCracken, B. Milliard, S. Peirani, S. Prunet, S. Rouberol, J. Silk, A. Slyz, T. Sousbie, R. Teyssier, L. Tresse, M. Treyer, D. Vibert, and M. Volonteri. Dancing in the dark: galactic properties trace spin swings along the cosmic web. *Monthly Notices of the RAS*, 444(2):1453–1468, October 2014. doi: 10.1093/mnras/stu1227.
- [165] 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*, 13:12–37, November 2015. doi: 10.1016/j.ascom.2015.09.003.
- [166] Ian G. McCarthy, Joop Schaye, Simeon Bird, and Amandine M. C. Le Brun. The BAHAMAS project: calibrated hydrodynamical simulations for large-scale structure cosmology. *Monthly Notices of the RAS*, 465(3):2936–2965, March 2017. doi: 10.1093/mnras/stw2792.
- [167] Romeel Davé, Daniel Anglés-Alcázar, Desika Narayanan, Qi Li, Mika H. Rafieeantsoa, and Sarah Appleby. SIMBA: Cosmological simulations with black hole growth and feedback. *Monthly Notices of the RAS*, 486(2):2827–2849, June 2019. doi: 10.1093/mnras/stz937.
- [168] 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.
- [169] Y. P. Jing, H. J. Mo, and G. Börner. Spatial Correlation Function and Pairwise Velocity Dispersion of Galaxies: Cold Dark Matter Models versus the Las Campanas Survey. *Astrophysical Journal*, 494(1):1–12, February 1998. doi: 10.1086/305209.

- [170] Zheng Zheng, Alison L. Coil, and Idit Zehavi. Galaxy Evolution from Halo Occupation Distribution Modeling of DEEP2 and SDSS Galaxy Clustering. *Astrophysical Journal*, 667(2):760–779, October 2007. doi: 10.1086/521074.
- [171] Zheng Zheng, Idit Zehavi, Daniel J. Eisenstein, David H. Weinberg, and Y. P. Jing. Halo Occupation Distribution Modeling of Clustering of Luminous Red Galaxies. *Astrophysical Journal*, 707(1):554–572, December 2009. doi: 10.1088/0004-637X/707/1/554.
- [172] M. Celeste Artale, Idit Zehavi, Sergio Contreras, and Peder Norberg. The impact of assembly bias on the halo occupation in hydrodynamical simulations. *Monthly Notices of the RAS*, 480(3):3978–3992, November 2018. doi: 10.1093/mnras/sty2110.
- [173] 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. *Monthly Notices of the RAS*, 490(4):5693–5711, December 2019. doi: 10.1093/mnras/stz2546.
- [174] Ignacio G. Alfaro, Facundo Rodriguez, Andrés N. Ruiz, and Diego Garcia Lambas. How galaxies populate haloes in very low-density environments. An analysis of the halo occupation distribution in cosmic voids. *Astronomy and Astrophysics*, 638:A60, June 2020. doi: 10.1051/0004-6361/201937431.
- [175] Ignacio G. Alfaro, Andrés N. Ruiz, Heliana E. Luparello, Facundo Rodriguez, and Diego Garcia Lambas. Galaxy populations in haloes in high-density environments. *Astronomy and Astrophysics*, 654:A62, October 2021. doi: 10.1051/0004-6361/202039838.
- [176] Ignacio G. Alfaro, Facundo Rodriguez, Andrés N. Ruiz, Heliana E. Luparello, and Diego Garcia Lambas. How galaxies populate halos in extreme density environments: An analysis of the halo occupation distribution in SDSS. *Astronomy and Astrophysics*, 665:A44, September 2022. doi: 10.1051/0004-6361/202243542.

- [177] Shankar Agarwal, Romeel Davé, and Bruce A. Bassett. Painting galaxies into dark matter haloes using machine learning. *Monthly Notices of the RAS*, 478(3):3410–3422, August 2018. doi: 10.1093/mnras/sty1169.
- [178] Isidro Gómez-Vargas, Ricardo Medel Esquivel, Ricardo García-Salcedo, and J. Alberto Vázquez. Neural network reconstructions for the Hubble parameter, growth rate and distance modulus. *arXiv e-prints*, art. arXiv:2104.00595, April 2021. doi: 10.48550/arXiv.2104.00595.
- [179] Andreas Nygaard, Emil Brinch Holm, Steen Hannestad, and Thomas Tram. CONNECT: a neural network based framework for emulating cosmological observables and cosmological parameter inference. *Journal of Cosmology and Astroparticle Physics*, 2023(5):025, May 2023. doi: 10.1088/1475-7516/2023/05/025.
- [180] Lucia A. Perez, Shy Genel, Francisco Villaescusa-Navarro, Rachel S. Somerville, Austen Gabrielpillai, Daniel Anglés-Alcázar, Benjamin D. Wandelt, and L. Y. Aaron Yung. Constraining Cosmology with Machine Learning and Galaxy Clustering: The CAMELS-SAM Suite. *Astrophysical Journal*, 954(1):11, September 2023. doi: 10.3847/1538-4357/accd52.
- [181] Xiaoying Xu, Shirley Ho, Hy Trac, Jeff Schneider, Barnabas Poczós, and Michelle Ntampaka. A First Look at creating mock catalogs with machine learning techniques. *Astrophys. J.*, 772:147, 2013. doi: 10.1088/0004-637X/772/2/147.
- [182] Michelle Ntampaka, Daniel J. Eisenstein, Sihan Yuan, and Lehman H. Garrison. A Hybrid Deep Learning Approach to Cosmological Constraints from Galaxy Redshift Surveys. *Astrophysical Journal*, 889(2):151, February 2020. doi: 10.3847/1538-4357/ab5f5e.
- [183] 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.
- [184] Xiaojun Xu, Saurabh Kumar, Idit Zehavi, and Sergio Contreras. Predicting halo occupation and galaxy assembly bias with machine learning. *Monthly Notices of the RAS*, 507(4):4879–4899, November 2021. doi: 10.1093/mnras/stab2464.
 - [185] Ana Maria Delgado, Digvijay Wadekar, Boryana Hadzhiyska, Sownak Bose, Lars Hernquist, and Shirley Ho. Modelling the galaxy–halo connection with machine learning. *Mon. Not. Roy. Astron. Soc.*, 515(2):2733–2746, 2022. doi: 10.1093/mnras/stac1951.
 - [186] Tyann Dumerchat and Julian Bautista. Galaxy clustering multi-scale emulation. *arXiv e-prints*, art. arXiv:2311.10425, November 2023. doi: 10.48550/arXiv.2311.10425.
 - [187] Marc Victor Leonard Metchnik. *A fast N-body scheme for computational cosmology*. PhD thesis, University of Arizona, January 2009.
 - [188] Lehman H. Garrison, Daniel J. Eisenstein, and Philip A. Pinto. A high-fidelity realization of the Euclid code comparison N-body simulation with ABACUS. *Monthly Notices of the RAS*, 485(3):3370–3377, May 2019. doi: 10.1093/mnras/stz634.
 - [189] Lehman H. Garrison, Daniel J. Eisenstein, Douglas Ferrer, Nina A. Maksimova, and Philip A. Pinto. The ABACUS cosmological N-body code. *Monthly Notices of the RAS*, 508(1):575–596, November 2021. doi: 10.1093/mnras/stab2482.
 - [190] Lehman H. Garrison, Daniel J. Eisenstein, Douglas Ferrer, Jeremy L. Tinker, Philip A. Pinto, and David H. Weinberg. The Abacus Cosmos: A Suite of Cosmological N-body Simulations. *Astrophysical Journal, Supplement*, 236(2):43, June 2018. doi: 10.3847/1538-4365/aabfd3.
 - [191] Lehman H. Garrison, Daniel J. Eisenstein, Douglas Ferrer, Marc V. Metchnik, and Philip A. Pinto. Improving Initial Conditions for Cosmological N-Body Simulations. *Mon. Not. Roy. Astron. Soc.*, 461(4):4125–4145, 2016. doi: 10.1093/mnras/stw1594.

- [192] Kyle S. Dawson, David J. Schlegel, Christopher P. Ahn, Scott F. Anderson, Éric Aubourg, Stephen Bailey, Robert H. Barkhouser, Julian E. Bautista, Alessandra Beifiori, Andreas A. Berlind, Vaishali Bhardwaj, Dmitry Bizyaev, Cullen H. Blake, Michael R. Blanton, Michael Blomqvist, Adam S. Bolton, Arnaud Borde, Jo Bovy, W. N. Brandt, Howard Brewington, Jon Brinkmann, Peter J. Brown, Joel R. Brownstein, Kevin Bundy, N. G. Busca, William Carithers, Aurelio R. Carnero, Michael A. Carr, Yanmei Chen, Johan Comparat, Natalia Connolly, Frances Cope, Rupert A. C. Croft, Antonio J. Cuesta, Luiz N. da Costa, James R. A. Davenport, Timothée Delubac, Roland de Putter, Saurav Dhital, Anne Ealet, Garrett L. Ebelke, Daniel J. Eisenstein, S. Escoffier, Xiaohui Fan, N. Filiz Ak, Hayley Finley, Andreu Font-Ribera, R. Génova-Santos, James E. Gunn, Hong Guo, Daryl Haggard, Patrick B. Hall, Jean-Christophe Hamilton, Ben Harris, David W. Harris, Shirley Ho, David W. Hogg, Diana Holder, Klaus Honscheid, Joe Huehnerhoff, Beatrice Jordan, Wendell P. Jordan, Guinevere Kauffmann, Eyal A. Kazin, David Kirkby, Mark A. Klaene, Jean-Paul Kneib, Jean-Marc Le Goff, Khee-Gan Lee, Daniel C. Long, Craig P. Loomis, Britt Lundgren, Robert H. Lupton, Marcio A. G. Maia, Martin Makler, Elena Malanushenko, Viktor Malanushenko, Rachel Mandelbaum, Marc Manera, Claudia Maraston, Daniel Margala, Karen L. Masters, Cameron K. McBride, Patrick McDonald, Ian D. McGreer, Richard G. McMahon, Olga Mena, Jordi Miralda-Escudé, Antonio D. Montero-Dorta, Francesco Montesano, Demitri Muna, Adam D. Myers, Tracy Naugle, Robert C. Nichol, Pasquier Noterdaeme, Sebastián E. Nuza, Matthew D. Olmstead, Audrey Oravetz, Daniel J. Oravetz, Russell Owen, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Kaike Pan, John K. Parejko, Isabelle Pâris, Will J. Percival, Ismael Pérez-Fournon, Ignasi Pérez-Ràfols, Patrick Petitjean, Robert Pfaffenberger, Janine Pforr, Matthew M. Pieri, Francisco Prada, Adrian M. Price-Whelan, M. Jordan Raddick, Rafael Rebolo, James Rich, Gordon T. Richards, Constance M. Rockosi, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Graziano Rossi, J. A. Rubiño-Martin, Lado Samushia, Ariel G. Sánchez, Conor Sayres, Sarah J. Schmidt, Donald P. Schneider, C. G. Scóccola, Hee-Jong Seo, Alaina Sheldon, Erin Sheldon, Yue Shen, Yiping Shu,

Anže Slosar, Stephen A. Smee, Stephanie A. Snedden, Fritz Stauffer, Oliver Steele, Michael A. Strauss, Alina Streblyanska, Nao Suzuki, Molly E. C. Swanson, Tomer Tal, Masayuki Tanaka, Daniel Thomas, Jeremy L. Tinker, Rita Tojeiro, Christy A. Tremonti, M. Vargas Magaña, Licia Verde, Matteo Viel, David A. Wake, Mike Watson, Benjamin A. Weaver, David H. Weinberg, Benjamin J. Weiner, Andrew A. West, Martin White, W. M. Wood-Vasey, Christophe Yèche, Idit Zehavi, Gong-Bo Zhao, and Zheng Zheng. The Baryon Oscillation Spectroscopic Survey of SDSS-III. *Astronomical Journal*, 145(1):10, January 2013. doi: 10.1088/0004-6256/145/1/10.

- [193] DESI Collaboration, Amir Aghamousa, Jessica Aguilar, Steve Ahlen, Shadab Alam, Lori E. Allen, Carlos Allende Prieto, James Annis, Stephen Bailey, Christophe Balleland, Otger Ballester, Charles Baltay, Lucas Beaufore, Chris Bebek, Timothy C. Beers, Eric F. Bell, José Luis Bernal, Robert Besuner, Florian Beutler, Chris Blake, Hannes Bleuler, Michael Blomqvist, Robert Blum, Adam S. Bolton, Cesar Briceno, David Brooks, Joel R. Brownstein, Elizabeth Buckley-Geer, Angela Burden, Etienne Burtin, Nicolas G. Busca, Robert N. Cahn, Yan-Chuan Cai, Laia Cardiel-Sas, Raymond G. Carlberg, Pierre-Henri Carton, Ricard Casas, Francisco J. Castander, Jorge L. Cervantes-Cota, Todd M. Claybaugh, Madeline Close, Carl T. Coker, Shaun Cole, Johan Comparat, Andrew P. Cooper, M. C. Cousinou, Martin Crocce, Jean-Gabriel Cuby, Daniel P. Cunningham, Tamara M. Davis, Kyle S. Dawson, Axel de la Macorra, Juan De Vicente, Timothée Delubac, Mark Derwent, Arjun Dey, Govinda Dhungana, Zhejie Ding, Peter Doel, Yutong T. Duan, Anne Ealet, Jerry Edelstein, Sarah Eftekharzadeh, Daniel J. Eisenstein, Ann Elliott, Stéphanie Escoffier, Matthew Evatt, Parker Fagrelus, Xiaohui Fan, Kevin Fanning, Arya Farahi, Jay Farihi, Ginevra Favole, Yu Feng, Enrique Fernandez, Joseph R. Findlay, Douglas P. Finkbeiner, Michael J. Fitzpatrick, Brenna Flaugher, Samuel Flender, Andreu Font-Ribera, Jaime E. Forero-Romero, Pablo Fosalba, Carlos S. Frenk, Michele Fumagalli, Boris T. Gaensicke, Giuseppe Gallo, Juan Garcia-Bellido, Enrique Gaztanaga, Nicola Pietro Gentile Fusillo, Terry Gerard, Irena Gershkovich, Tommaso Giannan-

tonio, Denis Gillet, Guillermo Gonzalez-de-Rivera, Violeta Gonzalez-Perez, Shelby Gott, Or Graur, Gaston Gutierrez, Julien Guy, Salman Habib, Henry Heetderks, Ian Heetderks, Katrin Heitmann, Wojciech A. Hellwing, David A. Herrera, Shirley Ho, Stephen Holland, Klaus Honscheid, Eric Huff, Timothy A. Hutchinson, Dragan Huterer, Ho Seong Hwang, Joseph Maria Illa Laguna, Yuzo Ishikawa, Dianna Jacobs, Niall Jeffrey, Patrick Jelinsky, Elise Jennings, Linhua Jiang, Jorge Jimenez, Jennifer Johnson, Richard Joyce, Eric Jullo, Stéphanie Juneau, Sami Kama, Armin Karcher, Sonia Karkar, Robert Kehoe, Noble Kennamer, Stephen Kent, Martin Kilbinger, Alex G. Kim, David Kirkby, Theodore Kisner, Ellie Kitanidis, Jean-Paul Kneib, Sergey Koposov, Eve Kovacs, Kazuya Koyama, Anthony Kremin, Richard Kron, Luzius Kronig, Andrea Kueter-Young, Cedric G. Lacey, Robin Lafever, Ofer Lahav, Andrew Lambert, Michael Lampton, Martin Landriau, Dustin Lang, Tod R. Lauer, Jean-Marc Le Goff, Laurent Le Guillou, Auguste Le Van Suu, Jae Hyeon Lee, Su-Jeong Lee, Daniela Leitner, Michael Lesser, Michael E. Levi, Benjamin L’Huillier, Baojiu Li, Ming Liang, Huan Lin, Eric Linder, Sarah R. Loebman, Zarija Lukić, Jun Ma, Niall MacCrann, Christophe Magneville, Laleh Makarem, Marc Manera, Christopher J. Manser, Robert Marshall, Paul Martini, Richard Massey, Thomas Matheson, Jeremy McCauley, Patrick McDonald, Ian D. McGreer, Aaron Meisner, Nigel Metcalfe, Timothy N. Miller, Ramon Miquel, John Moustakas, Adam Myers, Milind Naik, Jeffrey A. Newman, Robert C. Nichol, Andrina Nicola, Luiz Nicolati da Costa, Jundan Nie, Gustavo Niz, Peder Norberg, Brian Nord, Dara Norman, Peter Nugent, Thomas O’Brien, Minji Oh, Knut A. G. Olsen, Cristobal Padilla, Hamsa Padmanabhan, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Antonella Palmese, Daniel Pappalardo, Isabelle Pâris, Changbom Park, Anna Patej, John A. Peacock, Hiranya V. Peiris, Xiyan Peng, Will J. Percival, Sandrine Perruchot, Matthew M. Pieri, Richard Pogge, Jennifer E. Pollack, Claire Poppett, Francisco Prada, Abhishek Prakash, Ronald G. Probst, David Rabinowitz, Anand Raichoor, Chang Hee Ree, Alexandre Refregier, Xavier Regal, Beth Reid, Kevin Reil, Mehdi Rezaie, Constance M. Rockosi, Natalie Roe, Samuel Ronayette, Aaron Roodman, Ashley J. Ross, Nicholas P. Ross, Graziano

Rossi, Eduardo Roza, Vanina Ruhlmann-Kleider, Eli S. Rykoff, Cristiano Sabiu, Lado Samushia, Eusebio Sanchez, Javier Sanchez, David J. Schlegel, Michael Schneider, Michael Schubnell, Aurélia Secroun, Uros Seljak, Hee-Jong Seo, Santiago Serrano, Arman Shafieloo, Huanyuan Shan, Ray Sharples, Michael J. Sholl, William V. Shourt, Joseph H. Silber, David R. Silva, Martin M. Sirk, Anze Slosar, Alex Smith, George F. Smoot, Debopam Som, Yong-Seon Song, David Sprayberry, Ryan Staten, Andy Stefanik, Gregory Tarle, Suk Sien Tie, Jeremy L. Tinker, Rita Tojeiro, Francisco Valdes, Octavio Valenzuela, Monica Valluri, Mariana Vargas-Magana, Licia Verde, Alistair R. Walker, Jiali Wang, Yuting Wang, Benjamin A. Weaver, Curtis Weaverdyck, Risa H. Wechsler, David H. Weinberg, Martin White, Qian Yang, Christophe Yèche, Tianmeng Zhang, Gong-Bo Zhao, Yi Zheng, Xu Zhou, Zhimin Zhou, Yaling Zhu, Hu Zou, and Ying Zu. The DESI Experiment Part I: Science, Targeting, and Survey Design. *arXiv e-prints*, art. arXiv:1611.00036, October 2016. doi: 10.48550/arXiv.1611.00036.

- [194] Euclid Collaboration, R. Scaramella, J. Amiaux, Y. Mellier, C. Burigana, C. S. Carvalho, J. C. Cuillandre, A. Da Silva, A. Derosa, J. Dinis, E. Maiorano, M. Maris, I. Tereno, R. Laureijs, T. Boenke, G. Buenadicha, X. Dupac, L. M. Gaspar Venancio, P. Gómez-Álvarez, J. Hoar, J. Lorenzo Alvarez, G. D. Racca, G. Saavedra-Criado, J. Schwartz, R. Vavrek, M. Schirmer, H. Aussel, R. Azzollini, V. F. Cardone, M. Cropper, A. Ealet, B. Garilli, W. Gillard, B. R. Granett, L. Guzzo, H. Hoekstra, K. Jahnke, T. Kitching, T. Maciaszek, M. Meneghetti, L. Miller, R. Nakajima, S. M. Niemi, F. Pasian, W. J. Percival, S. Pottinger, M. Sauvage, M. Scodeggio, S. Wachter, A. Zacchei, N. Aghanim, A. Amara, T. Auphan, N. Auricchio, S. Awan, A. Balestra, R. Bender, C. Bodendorf, D. Bonino, E. Branchini, S. Brau-Nogue, M. Brescia, G. P. Candini, V. Capobianco, C. Carbone, R. G. Carlberg, J. Carretero, R. Casas, F. J. Castander, M. Castellano, S. Cavuoti, A. Cimatti, R. Cledassou, G. Congedo, C. J. Conselice, L. Conversi, Y. Copin, L. Corcione, A. Costille, F. Courbin, H. Degaudenzi, M. Douspis, F. Dubath, C. A. J. Duncan, S. Dusini, S. Farrens, S. Ferriol, P. Fos-alba, N. Fourmanoit, M. Frailis, E. Franceschi, P. Franzetti, M. Fumana, B. Gillis,

C. Giocoli, A. Grazian, F. Grupp, S. V. H. Haugan, W. Holmes, F. Hormuth, P. Hudelot, S. Kermiche, A. Kiessling, M. Kilbinger, R. Kohley, B. Kubik, M. Kümmel, M. Kunz, H. Kurki-Suonio, O. Lahav, S. Ligori, P. B. Lilje, I. Lloro, O. Mansutti, O. Marggraf, K. Markovic, F. Marulli, R. Massey, S. Maurogordato, M. Melchior, E. Merlin, G. Meylan, J. J. Mohr, M. Moresco, B. Morin, L. Moscardini, E. Munari, R. C. Nichol, C. Padilla, S. Paltani, J. Peacock, K. Pedersen, V. Pettorino, S. Pires, M. Poncet, L. Popa, L. Pozzetti, F. Raison, R. Rebolo, J. Rhodes, H. W. Rix, M. Roncarelli, E. Rossetti, R. Saglia, P. Schneider, T. Schrabback, A. Secroun, G. Seidel, S. Serrano, C. Sirignano, G. Sirri, J. Skottfelt, L. Stanco, J. L. Starck, P. Tallada-Crespí, D. Tavagnacco, A. N. Taylor, H. I. Teplitz, R. Toledo-Moreo, F. Torradeflot, M. Trifoglio, E. A. Valentijn, L. Valenziano, G. A. Verdoes Kleijn, Y. Wang, N. Welikala, J. Weller, M. Wetzstein, G. Zamorani, J. Zoubian, S. Andreon, M. Baldi, S. Bardelli, A. Boucaud, S. Camera, D. Di Ferdinando, G. Fabbian, R. Farinelli, S. Galeotta, J. Graciá-Carpio, D. Maino, E. Medinaceli, S. Mei, C. Neissner, G. Polenta, A. Renzi, E. Romelli, C. Rosset, F. Sureau, M. Tenti, T. Vassallo, E. Zucca, C. Baccigalupi, A. Balaguera-Antolínez, P. Battaglia, A. Biviano, S. Borgani, E. Bozzo, R. Cabanac, A. Cappi, S. Casas, G. Castignani, C. Colodro-Conde, J. Coupon, H. M. Courtois, J. Cuby, S. de la Torre, S. Desai, H. Dole, M. Fabricius, M. Farina, P. G. Ferreira, F. Finelli, P. Flose-Reimberg, S. Fotopoulou, K. Ganga, G. Gozaliasl, I. M. Hook, E. Keihänen, C. C. Kirkpatrick, P. Liebing, V. Lindholm, G. Mainetti, M. Martinelli, N. Martinet, M. Maturi, H. J. McCracken, R. B. Metcalf, G. Morgante, J. Nightingale, A. Nucita, L. Patrizii, D. Potter, G. Riccio, A. G. Sánchez, D. Sapone, J. A. Schewtschenko, M. Schultheis, V. Scottez, R. Teyssier, I. Tutusaus, J. Valiviita, M. Viel, W. Vriend, and L. Whittaker. Euclid preparation. I. The Euclid Wide Survey. *Astronomy and Astrophysics*, 662:A112, June 2022. doi: 10.1051/0004-6361/202141938.

- [195] D. Spergel, N. Gehrels, C. Baltay, D. Bennett, J. Brekinridge, M. Donahue, A. Dressler, B. S. Gaudi, T. Greene, O. Guyon, C. Hirata, J. Kalirai, N. J. Kasdin,

- B. Macintosh, W. Moos, S. Perlmutter, M. Postman, B. Rauscher, J. Rhodes, Y. Wang, D. Weinberg, D. Benford, M. Hudson, W. S. Jeong, Y. Mellier, W. Traub, T. Yamada, P. Capak, J. Colbert, D. Masters, M. Penny, D. Savransky, D. Stern, N. Zimmerman, R. Barry, L. Bartusek, K. Carpenter, E. Cheng, D. Content, F. Dekens, R. Demers, K. Grady, C. Jackson, G. Kuan, J. Kruk, M. Melton, B. Nemati, B. Parvin, I. Poberezhskiy, C. Peddie, J. Ruffa, J. K. Wallace, A. Whipple, E. Wollack, and F. Zhao. Wide-Field Infrared Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report. *arXiv e-prints*, art. arXiv:1503.03757, March 2015. doi: 10.48550/arXiv.1503.03757.
- [196] Andrew P. Hearin et al. Forward Modeling of Large-Scale Structure: An open-source approach with Halotools. *Astron. J.*, 154(5):190, 2017. doi: 10.5281/zenodo.835895.
- [197] Manodeep Sinha and Lehman H. Garrison. CORRFUNC - a suite of blazing fast correlation functions on the CPU. *Monthly Notices of the RAS*, 491(2):3022–3041, Jan 2020. doi: 10.1093/mnras/stz3157.
- [198] Manodeep Sinha and Lehman H. Garrison. corrfunc – a suite of blazing fast correlation functions on the CPU. *Mon. Not. Roy. Astron. Soc.*, 491(2):3022–3041, 2020. doi: 10.1093/mnras/stz3157.
- [199] Ji Meng Loh. A valid and fast spatial bootstrap for correlation functions. *arXiv e-prints*, art. arXiv:0805.2325, May 2008. doi: 10.48550/arXiv.0805.2325.
- [200] P. Norberg, C. M. Baugh, E. Gaztañaga, and D. J. Croton. Statistical analysis of galaxy surveys - I. Robust error estimation for two-point clustering statistics. *Monthly Notices of the RAS*, 396(1):19–38, June 2009. doi: 10.1111/j.1365-2966.2009.14389.x.
- [201] O. Friedrich, S. Seitz, T. F. Eifler, and D. Gruen. Performance of internal covariance estimators for cosmic shear correlation functions. *Monthly Notices of the RAS*, 456(3): 2662–2680, March 2016. doi: 10.1093/mnras/stv2833.

- [202] Ginevra Favole, Benjamin R. Granett, Javier Silva Lafaure, and Domenico Sapon. Does jackknife scale really matter for accurate large-scale structure covariances? *Monthly Notices of the RAS*, 505(4):5833–5845, August 2021. doi: 10.1093/mnras/stab1720.
- [203] Faizan G. Mohammad and Will J. Percival. Creating jackknife and bootstrap estimates of the covariance matrix for the two-point correlation function. *Monthly Notices of the RAS*, 514(1):1289–1301, July 2022. doi: 10.1093/mnras/stac1458.
- [204] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, October 2011. doi: 10.48550/arXiv.1201.0490.
- [205] Ethan O. Nadler, Yao-Yuan Mao, Risa H. Wechsler, Shea Garrison-Kimmel, and Andrew Wetzel. Modeling the Impact of Baryons on Subhalo Populations with Machine Learning. *Astrophysical Journal*, 859(2):129, June 2018. doi: 10.3847/1538-4357/aac266.
- [206] J. D. Cohn and Nicholas Battaglia. Multiwavelength cluster mass estimates and machine learning. *Monthly Notices of the RAS*, 491(2):1575–1584, January 2020. doi: 10.1093/mnras/stz3087.
- [207] S. Mucesh, W. G. Hartley, A. Palmese, O. Lahav, L. Whiteway, A. F. L. Bluck, A. Alarcon, A. Amon, K. Bechtol, G. M. Bernstein, A. Carnero Rosell, M. Carrasco Kind, A. Choi, K. Eckert, S. Everett, D. Gruen, R. A. Gruendl, I. Harrison, E. M. Huff, N. Kuropatkin, I. Sevilla-Noarbe, E. Sheldon, B. Yanny, M. Aguena, S. Allam, D. Bacon, E. Bertin, S. Bhargava, D. Brooks, J. Carretero, F. J. Castander, C. Conselice, M. Costanzi, M. Crocce, L. N. da Costa, M. E. S. Pereira, J. De Vicente, S. Desai, H. T. Diehl, A. Drlica-Wagner, A. E. Evrard, I. Ferrero, B. Flaugher, P. Fosalba,

- J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, J. Gschwend, G. Gutierrez, S. R. Hinton, D. L. Hollowood, K. Honscheid, D. J. James, K. Kuehn, M. Lima, H. Lin, M. A. G. Maia, P. Melchior, F. Menanteau, R. Miquel, R. Morgan, F. Paz-Chinchón, A. A. Plazas, E. Sanchez, V. Scarpine, M. Schubnell, S. Serrano, M. Smith, E. Suchyta, G. Tarle, D. Thomas, C. To, T. N. Varga, R. D. Wilkinson, and DES Collaboration. A machine learning approach to galaxy properties: joint redshift-stellar mass probability distributions with Random Forest. *Monthly Notices of the RAS*, 502 (2):2770–2786, April 2021. doi: 10.1093/mnras/stab164.
- [208] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. emcee: The MCMC Hammer. *Publications of the ASP*, 125(925):306, March 2013. doi: 10.1086/670067.
- [209] Juergen Schmidhuber. Deep Learning in Neural Networks: An Overview. *arXiv e-prints*, art. arXiv:1404.7828, April 2014. doi: 10.48550/arXiv.1404.7828.
- [210] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. doi: 10.1038/323533a0.
- [211] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4:251–257, 1991. URL <https://api.semanticscholar.org/CorpusID:7343126>.
- [212] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Neural Information Processing Systems*, 2011. URL <https://api.semanticscholar.org/CorpusID:11688126>.
- [213] Lutz Prechelt. Automatic early stopping using cross validation: quantifying the criteria. *Neural networks : the official journal of the International Neural Network Society*, 11 4:761–767, 1998. URL <https://api.semanticscholar.org/CorpusID:2887064>.

- [214] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15:1929–1958, 2014. URL <https://api.semanticscholar.org/CorpusID:6844431>.
- [215] Leo Breiman. Random Forests. *Machine Learning*, 45:5–32, January 2001. doi: 10.1023/A:1010933404324.
- [216] Gérard Biau and Erwan Scornet. A Random Forest Guided Tour. *arXiv e-prints*, art. arXiv:1511.05741, November 2015. doi: 10.48550/arXiv.1511.05741.
- [217] Philipp Probst, Marvin Wright, and Anne-Laure Boulesteix. Hyperparameters and Tuning Strategies for Random Forest. *arXiv e-prints*, art. arXiv:1804.03515, April 2018. doi: 10.48550/arXiv.1804.03515.
- [218] David John Cameron MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992. URL <https://api.semanticscholar.org/CorpusID:1762283>.
- [219] CJ Willmott and K. Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30:79–82, January 2005. doi: 10.3354/cr030079.
- [220] T. Chai and R. R. Draxler. Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3):1247–1250, 2014. doi: 10.5194/gmd-7-1247-2014. URL <https://gmd.copernicus.org/articles/7/1247/2014/>.
- [221] III Gott, J. Richard, Michael S. Vogeley, Silviu Podariu, and Bharat Ratra. Median Statistics, H_0 , and the Accelerating Universe. *Astrophysical Journal*, 549(1):1–17, March 2001. doi: 10.1086/319055.
- [222] Sergios Theodoridis. *Machine learning: a Bayesian and optimization perspective*. Academic press, 2015.

- [223] Christian Partmann, Christian Fidler, Cornelius Rampf, and Oliver Hahn. Fast simulations of cosmic large-scale structure with massive neutrinos. *Journal of Cosmology and Astroparticle Physics*, 2020(9):018, September 2020. doi: 10.1088/1475-7516/2020/09/018.
- [224] Davide Gualdi, Héctor Gil-Marín, and Licia Verde. Joint analysis of anisotropic power spectrum, bispectrum and trispectrum: application to N-body simulations. *Journal of Cosmology and Astroparticle Physics*, 2021(7):008, July 2021. doi: 10.1088/1475-7516/2021/07/008.
- [225] William McDonald, Danail Obreschkow, and Lilian Garratt-Smithson. Dissociation of dark matter and gas in cosmic large-scale structure. *Monthly Notices of the RAS*, 516(4):5289–5308, November 2022. doi: 10.1093/mnras/stac2276.
- [226] Jeppe Dakin, Steen Hannestad, and Thomas Tram. Fully relativistic treatment of decaying cold dark matter in N-body simulations. *Journal of Cosmology and Astroparticle Physics*, 2019(6):032, June 2019. doi: 10.1088/1475-7516/2019/06/032.
- [227] Jazhiel Chacón, Isidro Gómez-Vargas, Ricardo Menchaca Méndez, and J. Alberto Vázquez. Analysis of dark matter halo structure formation in N -body simulations with machine learning. *Physical Review D*, 107(12):123515, June 2023. doi: 10.1103/PhysRevD.107.123515.
- [228] Keir K. Rogers and Hiranya V. Peiris. General framework for cosmological dark matter bounds using N -body simulations. *Physical Review D*, 103(4):043526, February 2021. doi: 10.1103/PhysRevD.103.043526.
- [229] 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. *Nature*, 435(7042):629–636, June 2005. doi: 10.1038/nature03597.

- [230] V. Springel, J. Wang, M. Vogelsberger, A. Ludlow, A. Jenkins, A. Helmi, J. F. Navarro, C. S. Frenk, and S. D. M. White. The Aquarius Project: the subhaloes of galactic haloes. *Monthly Notices of the RAS*, 391(4):1685–1711, December 2008. doi: 10.1111/j.1365-2966.2008.14066.x.
- [231] Sownak Bose, Wojciech A. Hellwing, Carlos S. Frenk, Adrian Jenkins, Mark R. Lovell, John C. Helly, and Baojiu Li. The Copernicus Complexio: statistical properties of warm dark matter haloes. *Monthly Notices of the RAS*, 455(1):318–333, January 2016. doi: 10.1093/mnras/stv2294.
- [232] 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. *Monthly Notices of the RAS*, 444(2):1518–1547, October 2014. doi: 10.1093/mnras/stu1536.
- [233] 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. *Monthly Notices of the RAS*, 490(3):3234–3261, December 2019. doi: 10.1093/mnras/stz2306.
- [234] Till Sawala, Carlos S. Frenk, Azadeh Fattahi, Julio F. Navarro, Richard G. Bower, Robert A. Crain, Claudio Dalla Vecchia, Michelle Furlong, John. C. Helly, Adrian Jenkins, Kyle A. Oman, Matthieu Schaller, Joop Schaye, Tom Theuns, James Trayford, and Simon D. M. White. The APOSTLE simulations: solutions to the Local Group’s cosmic puzzles. *Monthly Notices of the RAS*, 457(2):1931–1943, April 2016. doi: 10.1093/mnras/stw145.
- [235] K. Dolag, E. Komatsu, and R. Sunyaev. SZ effects in the Magneticum Pathfinder simulation: comparison with the Planck, SPT, and ACT results. *Monthly Notices of the RAS*, 463(2):1797–1811, December 2016. doi: 10.1093/mnras/stw2035.

- [236] Sihan Yuan, Lehman H. Garrison, Daniel J. Eisenstein, and Risa H. Wechsler. Stringent σ_8 constraints from small-scale galaxy clustering using a hybrid MCMC + emulator framework. *Monthly Notices of the RAS*, 515(1):871–896, September 2022. doi: 10.1093/mnras/stac1830.
- [237] Sihan Yuan, Hanyu Zhang, Ashley J. Ross, Jamie Donald-McCann, Boryana Hadzhiyska, Risa H. Wechsler, Zheng Zheng, Shadab Alam, Violeta Gonzalez-Perez, Jessica Nicole Aguilar, Steven Ahlen, Davide Bianchi, David Brooks, Axel de la Macorra, Kevin Fanning, Jaime E. Forero-Romero, Klaus Honscheid, Mustapha Ishak, Robert Kehoe, James Lasker, Martin Landriau, Marc Manera, Paul Martini, Aaron Meisner, Ramon Miquel, John Moustakas, Seshadri Nadathur, Jeffrey A. Newman, Jundan Nie, Will Percival, Claire Poppett, Antoine Rocher, Graziano Rossi, Eusebio Sanchez, Lado Samushia, Michael Schubnell, Hee-Jong Seo, Gregory Tarlé, Benjamin Alan Weaver, Jiaxi Yu, Zhimin Zhou, and Hu Zou. The DESI one-per cent survey: exploring the halo occupation distribution of luminous red galaxies and quasi-stellar objects with ABACUSSUMMIT. *Monthly Notices of the RAS*, 530(1):947–965, May 2024. doi: 10.1093/mnras/stae359.
- [238] Yevgeniy Dodis, Leonid Reyzin, and Adam Smith. Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. In *International conference on the theory and applications of cryptographic techniques*, pages 523–540. Springer, 2004.
- [239] Arathi Arakala, Jason Jeffers, and Kathy J Horadam. Fuzzy extractors for minutiae-based fingerprint authentication. In *International conference on biometrics*, pages 760–769. Springer, 2007.
- [240] FVC2000: the First International Competition for Fingerprint Verification Algorithms. . <http://bias.csr.unibo.it/fvc2000/>, 2000. [Online; accessed 10-February-2022].
- [241] Dario Maio, Davide Maltoni, Raffaele Cappelli, James L. Wayman, and Anil K. Jain. Fvc2000: Fingerprint verification competition. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):402–412, 2002.

- [242] Stephanie Gootman. Opm hack: The most dangerous threat to the federal government today. *Journal of Applied Security Research*, 11(4):517–525, 2016.
- [243] Josh Taylor. Major breach found in biometrics system used by banks, uk police and defence firms. *The Guardian*, August 2019.
- [244] Zhenyu Liu and Dimitrios A Pados. A decoding algorithm for finite-geometry ldpc codes. *IEEE Transactions on Communications*, 53(3):415–421, 2005.
- [245] Naftali Sommer, Meir Feder, and Ofir Shalvi. Low-density lattice codes. *IEEE Transactions on Information Theory*, 54(4):1561–1585, 2008.
- [246] Metodi P Yankov, Martin A Olsen, Mikkel B Stegmann, Søren Sk Christensen, and Søren Forchhammer. Fingerprint entropy and identification capacity estimation based on pixel-level generative modelling. *IEEE Transactions on Information Forensics and Security*, 15:56–65, 2019.
- [247] Meng-Hui Lim and Pong C Yuen. Entropy measurement for biometric verification systems. *IEEE transactions on cybernetics*, 46(5):1065–1077, 2015.
- [248] Yagiz Sutcu, Elham Tabassi, Husrev T Sencar, and Nasir Memon. What is biometric information and how to measure it? In *2013 IEEE international conference on technologies for homeland security (HST)*, pages 67–72. IEEE, 2013.
- [249] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.
- [250] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [251] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

- [252] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [253] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a” siamese” time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.
- [254] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 459–474, 2018.
- [255] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015. URL <http://arxiv.org/abs/1503.03832>.
- [256] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.
- [257] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pages 241–257. Springer, 2016.
- [258] Xuebo Wang and Wai Ho Mow. Efficient ldpc decoder design from the lattice viewpoint. In *2019 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2019.
- [259] FVC2006: the Fourth International Fingerprint Verification Competition. . <http://bias.csr.unibo.it/fvc2006/>, 2006. [Online; accessed 10-February-2022].
- [260] The Hong Kong Polytechnic University. Contactless 2D to Contact-based 2D Fingerprint Images Database Version 1.0. <http://www4.comp.polyu.edu.hk/~csajaykr/fingerprint.htm>, 2017. [Online; accessed 10-February-2022].

- [261] Google Research. Google bird vocalization classifier: A global bird embedding and classification model, 2023. URL <https://tfhub.dev/google/bird-vocalization-classifier/4>.
- [262] Stefan Kahl, Connor M Wood, Maximilian Eibl, and Holger Klinck. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021.