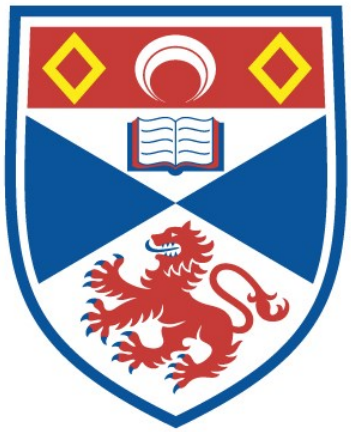
LEARNING AND INTERPRETING THE GALAXY-HALO CONNECTION IN COSMIC SIMULATIONS

Harry George Chittenden

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2023

LEARNING AND INTERPRETING THE GALAXY-HALO CONNECTION IN COSMIC SIMULATIONS

Harry George Chittenden

University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of

Doctor of Philosophy (PhD)

at the University of St Andrews

August 25, 2023

Declaration

Candidate's declaration

I, Harry George Chittenden, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 55,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in September 2019.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date August 25, 2023 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date August 25, 2023 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Harry George Chittenden, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date August 25, 2023 Signature of candidate

Date August 25, 2023 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Harry George Chittenden, understand that by declaring that I have original research data or digital outputs, I should make every effort in meeting the University's and research funders' requirements on the deposit and sharing of research data or research digital outputs.

Date August 25, 2023 Signature of candidate

Permission for publication of underpinning research data or digital outputs

We understand that for any original research data or digital outputs which are deposited, we are giving permission for them to be made available for use in accordance with the requirements of the University and research funders, for the time being in force.

We also understand that the title and the description will be published, and that the underpinning research data or digital outputs will be electronically accessible for use in accordance with the license specified at the point of deposit, unless exempt by award of an embargo as requested below.

The following is an agreed request by candidate and supervisor regarding the publication of underpinning research data or digital outputs: No embargo on underpinning research data or digital outputs.

Date August 25, 2023 Signature of candidate

Date August 25, 2023 Signature of supervisor

Abstract

In modern galactic astronomy, cosmological simulations and observational galaxy surveys work hand in hand, offering valuable insights into the historical evolution of galaxies on both cosmological scales and an individual basis. As dark matter halos constitute a significant portion of the mass in galaxies, clusters, and cosmic structures, they profoundly impact the properties of galaxies. This relationship is known as the galaxy-halo connection.

Galaxies possess a complex nature necessitating computationally intensive modelling. Accurately and consistently modelling galaxy-halo coevolution across all scales thus presents a challenge, and compromises are usually made between simulation size and resolution. However, it is possible to conduct pure dark matter simulations on larger scales, requiring a fraction of the power of complete simulations. As observational surveys expand in size and detail, however, simulations of this magnitude become crucial in supporting their findings, surpassing the limitations of galaxy simulations.

In this thesis, I present a machine learning model which encodes the galaxy-halo connection within a cosmohydrodynamical simulation. This model predicts the star formation and metallicity of galaxies over time, from properties of their halos and cosmic environment. These predictions are used to emulate observational data using spectral synthesis models, and subsequently the model is applied to a large dark matter simulation.

Through these predictions, the model replicates the correlations responsible for galaxy evolution, as well as observable quantities reflecting this galaxy-halo connection, with similar results in dark matter simulations. The model computes accurate galaxy-halo statistics and reveals important physical relationships; specifically, variables associated with halo accretion influence a galaxy's mass and star formation, while environmental variables are linked to its metallicity. While the predictions from dark matter simulations are reasonably accurate, they are affected by the absence of baryonic processes, the resolution of the simulation, and the calculation of halo properties.

Acknowledgements

This work was supported by the Science & Technology Facilities Council, under grant number ST/T506448/1. I am extremely grateful for their funding and support of my postgraduate research.

Special thanks go to the researchers behind the IllustrisTNG (Nelson et al., 2017, 2019b; Pillepich et al., 2017b; Springel et al., 2017; Marinacci et al., 2018; Naiman et al., 2018) and Uchuu (Ishiyama et al., 2021) cosmological simulations for open access to and technical support with their simulation data.

To Rita, my highly valued PhD supervisor, without whom I would've been going round in circles all this time, thank you for everything. I have learned a lot from you professionally, academically and personally, and I am very glad to have worked with you.

To the galaxies group in St Andrews, I thoroughly enjoyed the insightful, friendly meetings and discussions we had on a regular basis. I wish you all the best for the future, and that includes a working, hassle-free seminar room setup. This includes Juan, who insisted upon one evening at The Rule that I had to include the word "burrito" somewhere in my PhD thesis. There, I just did.

To those of you who supported my numerous applications for postdoctoral positions, I truly appreciate you taking the time to do something so great for me. And thank you to Karl Glazebrook for offering me a job in Melbourne - I look forward to working with you.

To Mum, Dad, my little brother Tom and family, thank you for your continued support of this wild decision of mine to pursue astrophysics. Dad, it certainly helps in many practical manners to have another physicist in the family. I'll explain the contents of this thesis again when I next see you.

To Martin, Kate, Liz and Mike - my extended family in Australia, I'll see you soon.

And lastly, to my lovely black cat Kevin, whom I have had from when I was seven years old until you passed away last year. You have always made my day better, no matter how much of it had been spent debugging. Rest easy.

Research Data/Digital Outputs access statement

The code corresponding to the work in this thesis published in Chittenden & Tojeiro (2022) is available at https://github.com/hgc4/TNG-Networks. The data corresponding to this publication, and the pending publications Chittenden, Tojeiro, & Kraljic (in prep.)and Behera, Chittenden, & Tojeiro (in prep.), are available at https://doi.org/10.17630/9732988f-ed9c-4f03-92ff-5545d779e42d.

"*I may not have gone where I intended to go, but I think I have ended up where I needed to be.*"

- Douglas Adams, The Long Dark Tea-Time Of The Soul

"*I produced a detailed tribute to my wrongness.*"
"*That IS science!*"

- Nathan W. Pyle, Strange Planet

Contents

Declaration i

Abstract v

Acknowledgements vii

1 Introduction 1

1.1 Galaxy Evolution . . . 1

1.1.1 The Standard Cosmological Model . . . 1

1.1.2 Formation Of Cosmic Structure . . . 4

1.1.3 The Properties Of Galaxies . . . 5

1.2 The Galaxy-Halo Connection . . . 8

1.2.1 Modelling Approaches . . . 8

1.2.2 Halo And Galaxy Histories . . . 13

1.3 This Thesis . . . 20

1.3.1 Motivation . . . 20

1.3.2 Outline . . . 22

2 Data, Design & Preprocessing 25

2.1 Machine Learning . . . 25

2.2 A Semi-Recurrent Neural Network . . . 29

2.2.1 Architecture . . . 29

2.2.2 Activation Functions . . . 31

2.2.3 Learning Rate . . . 33

2.3 Simulation Data . . . 35

2.3.1 Simulation Suites . . . 35

2.3.2 Data Access . . . 40

2.4 Data In This Work . . . 45
2.4.1 Dark Matter Quantities . . . 46
2.4.2 Baryonic Quantities . . . 56
2.5 Data Preprocessing . . . 58
2.5.1 Quantile Transformation . . . 59
2.5.2 Vector & Scalar Normalisation . . . 60
2.6 Resolution Corrections . . . 62
2.6.1 Scaling Mass And Metallicity At Fixed Redshift . . . 63
2.6.2 Scaling SFH And ZH At Fixed Halo Mass . . . 64
2.7 Summary . . . 68

3 Neural Network Predictions **71**
3.1 Introduction . . . 71
3.2 Predicted Galaxy Properties . . . 72
3.2.1 Evolutionary History . . . 72
3.2.2 Galaxy-Halo Relationships At $z = 0$ . . . 73
3.2.3 Comparing Network Designs . . . 78
3.2.4 Quality Of Predictions . . . 80
3.3 Stochastic Corrections . . . 82
3.3.1 Motivation . . . 82
3.3.2 Methodology . . . 83
3.3.3 Results . . . 86
3.3.4 Interpretation . . . 91
3.4 Feature Importance . . . 92
3.4.1 Connected Input Properties . . . 92
3.4.2 Methodology . . . 93
3.4.3 Results & Discussion . . . 96
3.5 Summary . . . 103

4 Observables **107**
4.1 Introduction . . . 107
4.2 Calculated Observables . . . 109
4.2.1 Spectral Energy Distributions . . . 110

4.2.2 Photometric Luminosity And Colour . . . 112
4.2.3 Emission Line Luminosity . . . 113
4.3 Results . . . 113
4.3.1 Spectral Energy Distributions . . . 113
4.3.2 Hα Line Luminosity . . . 115
4.3.3 Band Magnitudes . . . 117
4.3.4 Photometric Colours . . . 119
4.4 Summary . . . 122

5 Dark Simulations **125**
5.1 Introduction . . . 125
5.2 Dark Simulation Data . . . 127
5.2.1 TNG-Dark . . . 127
5.2.2 Uchuu . . . 128
5.3 Halo Variables . . . 132
5.3.1 Mass Accretion History . . . 133
5.3.2 Half-Mass Radius . . . 136
5.3.3 Maximum Orbital Velocity . . . 139
5.3.4 Local Environment . . . 139
5.4 Galaxy Quantities . . . 141
5.4.1 Star Formation History . . . 141
5.4.2 Metallicity History . . . 145
5.4.3 Spectroscopy . . . 148
5.4.4 Photometry . . . 149
5.5 Future Model Amendments . . . 150
5.5.1 Simulation Resolution . . . 151
5.5.2 Input Parameters . . . 152
5.5.3 Galaxy Clustering . . . 153
5.5.4 Comparison With Other Models . . . 154
5.6 Conclusion . . . 155

6 Conclusions & Outlook **159**
6.1 Chapter Summaries . . . 160

6.2 Research Outlook 163

Bibliography **168**

List of Figures

1.1 Thermal fluctuations of the cosmic microwave background as observed by the COBE (Boggess et al., 1992; Smoot, 1999), WMAP (Bennett et al., 2013) and Planck Collaboration (2016) satellites, showing the level of detail obtained over time. Though the cosmic microwave background is considered near homogeneous and isotropic, indicating a previous state of the universe in which it was condensed into a much smaller space, these perturbations on the scale of one part in 10^5 would evolve into the large scale structures seen in the present-day universe. Image courtesy of NASA. 2

1.2 The ratio of energy densities of baryonic and dark matter and dark energy in the present universe, according to the Planck Collaboration (2016) cosmological model, which measured these densities by combining gravitational lensing, baryonic oscillation and CMB field cross-correlation data. The energy density associated with electromagnetic radiation constitutes a negligible quantity, and thus is not shown. These densities constitute fundamental parameters of the ΛCDM model. These values are assumed for the entirety of this thesis. 3

1.3 A visual summary of empirical and physical methods of modelling the galaxy-halo connection in cosmological simulations, as shown in Wechsler & Tinker (2018). The image on the left shows the dark matter distribution of a $90 \times 90 \times 30\text{Mpc}/h$ region of an N-body simulation, initialised from a random seed. The image on the right shows the distribution of galaxies assigned to halos in this simulation, based on an abundance matching model (see section 1.2.1) adjusted to match observational data. The scale below lists several modelling strategies ranging from physical models on the left to empirical models on the right, additionally listing basic techniques and assumptions. 9

2.1 Eight commonplace activation functions used in neural networks (Aggarwal, 2018). In each scenario, the vertical axis displays the output of a particular node, while the horizontal axis displays the input value to that node. The name of the function is given in the figure titles, and their mathematical expressions are given in the figure. 29

2.2 This diagram depicts the neural network architecture for central galaxies, with each dot representing a fully connected layer and its dimensionality indicated by a number, except for the purple dots that represent a subset of the final 1D output layer. The network has separate input layers for time-dependent and time-independent halo properties, which are combined at a dense layer with 42 nodes. The temporal input layer and recurrent layers are two-dimensional, consisting of eight variables over 33 time steps. The arrows show connections between consecutive layers, with the label indicating the number of times the connection repeats. For example, "3" means there are four consecutive hidden layers for that connection. The dashed line arrow indicates that every fourth hidden layer has three additional nodes until each layer reaches 69 nodes. Finally, the network outputs baryonic data, including star formation and metallicity histories, and three zero-redshift galaxy properties. 30

2.3 Four examples of adaptive learning rates used in neural networks, where the name of the adaptive learning rate is given in the title of each figure, and the mathematical formula for the learning rate Γ as a function of the epoch number N is given in the figure. In these mathematical expressions, κ and η are fixed, tailored hyperparameters, H is the Heaviside step function, and ζ is a random uniform variable. The stochastic formula is designed such that for every epoch in the training phase, there is a 1 in κ probability that the current learning rate will be halved. 34

2.4 Nine halo and galaxy variables mapped over a $110 \times 14 \times 37$Mpc region of space in the TNG100-1 simulation at $z = 0$, illustrating the wealth of information directly available from the IllustrisTNG public data release. The variables and their physical scales are indicated in the legend at the base of the figure, and ordered according to their appearance from the top to the bottom of the figure. Image taken from Nelson et al. (2019b). 37

2.5 (Left) Density map of the Uchuu dark matter simulation at $z = 0$, with white dashed boxes showing the three TNG simulations to scale. (Right) Magnified regions of the Uchuu simulation enclosed by the boxes in the left figure, showing the details resolved by Uchuu in a volume equivalent to the TNG simulations. 39

2.6 This diagram is a simplified representation of a halo merger tree, as seen in the TNG simulations, and is taken from Jiang & van den Bosch (2014). It is organised into rows, with each row representing a snapshot from the earliest time in the top row to the latest time in the bottom row. The size of each sphere corresponds to the mass of the halo it represents. In each row, the purple sphere represents the main (0^{th} order) progenitor of the host halo at the final snapshot, and the purple lines depict the main progenitor branch. Any other halos in the diagram are considered secondary progenitors, with the overlap of a smaller sphere over a larger one indicating that the smaller halo has been accreted by the larger halo as a bound subhalo. The small rectangles represent smooth accretion of matter that is not associated with halos. The boxed region provides an example of a subsection of the main tree which is also considered a subtree in its own right. For any such subtree, the highest order branch is the main progenitor branch of the subtree, while all other members of the ensemble are secondary progenitors; as is the case for the complete merger tree. 43

2.7 This figure depicts the x-y plane projection of the dark matter density distribution of subhalos surrounding three central subhalos of mass $\log_{10} M_h^{z=1}/M_\odot = 11.17$, taken from the $z = 1$ snapshot of the TNG100-1 simulation. The target subhalos are not visualised in these images, nor do they influence the skew calculation. Dark matter cells which are not gravitationally bound to the target subhalo and lie within a sphere of radius 3Mpc, centered on the target subhalo's center of mass, are selected for this image. The terms "low" and "high" skews are used to refer to the lower and upper quantiles of the skew dataset, respectively, while "medium" skews are close to the median skew. 52

2.8 The translation of the halo masses from TNG100-1 dataset (shown on the vertical axis) into both a uniform distribution (left) and Gaussian distribution (right). The histograms along each axis, including the halo mass distribution on the vertical axis, are also presented. The graph illustrates that a considerable range of the data, particularly the halo masses above $10^{12.5} M_\odot$, corresponds to a very narrow range in the uniform distribution, which leads to high sensitivity towards slight variations in the transformed data. Consequently, the uniform distribution is not ideal for making predictions from this data. Therefore, we opted to transform the data to a Gaussian distribution. 60

2.9 This graphic displays the distribution of monotonically increasing data after applying scalar (left) and vector (right) GQT normalisation. The data points at each time step are differentiated by various colours. Scalar normalisation transforms the data independently of other time steps, resulting in each time step sharing the same normal distribution. In contrast, vector normalisation transforms the data according to the complete range of the quantity's value over time, thus making each time step's distribution relative to another. When the full set of time step histograms is combined, it results in the Gaussian distribution of the complete dataset, irrespective of the method of normalisation used. 61

2.10 This figure displays depicts how the ζ corrections vary with halo mass at a redshift of zero, at all halo masses where we sample TNG300-1. For masses exceeding the range displayed, i.e. above $10^{14}M_{\odot}$, the zeta function values are calculated as the average over the $[10^{13}M_{\odot},\ 10^{14}M_{\odot}]$ interval. 63

2.11 The relationship between the SHMR and the HMZR at a redshift of zero is presented in this figure. The TNG100-1 data (green) is compared with the original (red) and adjusted (purple) TNG300-1 distributions, which have been adjusted using the zeta functions. The error bars with matching colours correspond to the median and the range between the 15^{th} and 85^{th} percentiles of either the stellar mass or metallicity within a particular halo mass bin for each dataset. 65

2.12 The figure displays the ψ variables with respect to cosmic time, with four representative halo mass bins shown as examples. The upper panel shows the resolution correction for the star formation history, ψ_S, while the lower panel shows the correction for the metallicity history, ψ_Z. 66

2.13 The figure illustrates the average star formation and metallicity histories of central galaxies with halo masses ranging from 10^{12} to $10^{12.2}$ solar masses, where the size of the shaded regions represents the standard deviation of these data as a function of time. The TNG100-1 data (green) is aligned with the modified TNG300-1 curves (purple), which have been adjusted from the original TNG300 data (red) using the ψ parameters. To eliminate erroneous characteristics caused by a small sample size at early times, time steps that contain fewer than 100 nonzero values across all data are excluded from the figure. This only applies to cosmic times earlier than 1Gyr. 67

3.1 This figure illustrates the evolutionary history of a satellite galaxy in IllustrisTNG with intermediate mass. It displays the original star formation history in blue, and its corresponding prediction by the neural network in cyan; as well as the true stellar metallicity history (purple) and the predicted metallicity history (magenta), indicating the time-dependent metallicity of stars formed according to the corresponding star formation rate. The sample's subhalo mass and predicted stellar mass and metallicity values are presented in the header of the figure. While the sample shows a decent match to the shapes of the star formation and metallicity histories, it fails to replicate the fluctuations on short time scales. 73

3.2 This figure shows the mean and standard deviation in the Fourier amplitudes of the star formation histories (left) and metallicity histories (right) of central galaxies in the halo mass range $10^{12.4} - 10^{12.6}M_{\odot}$, containing 638 samples. The Fourier transforms of the predicted data (green) show a clear decline in amplitude with respect to the original data (purple) for frequencies around 0.15Gyr^{-1} and higher. This indicates the lack of high frequency data in the neural network predictions. These Fourier transforms are plotted up to the Nyquist frequency of approximately 1.1954Gyr^{-1}. . . 74

3.3 The numerical stellar-halo mass relation assessed with the fiducial and predicted star formation rates is shown in this figure for central galaxies (left) and satellite galaxies (right). For the former, this is depicted as a function of halo mass, and for the latter, as a function of subhalo mass. The original TNG dataset's datapoints are presented in red and predictions of the networks are depicted in blue. Red and blue errorbars display the median and 15^{th} and 85^{th} percentiles of stellar mass in halo mass bins, while blue and red datapoints represent individual galaxies. The similarities between the shape and scatter of the two SHMRs suggest that the star formation histories are predicted similarly overall. 74

3.4 Stellar-halo mass relations for central galaxies (top panels) and satellites (bottom panels) are presented in 2D histograms which are coloured in accordance with the mean redshift per bin at which the central or satellite halo's final mass is produced. Plotting according to the individual star masses allows each SHMR to be seen independently for both the original TNG data and the predictions made by the networks. The network's predictions for this dependence of halo formation redshift on the SHMR are consistent with the original data for both central and satellite galaxies. The reader should be aware that the low occupancy of bins on the borders of the SHMRs makes them susceptible to slight variations in scatter, which deceptively gives the SHMRs an apparent distortion. 75

3.5 Similar figure to fig. 3.4, instead showing the dependence of the mass-weighted stellar ages of galaxies of the SHMR. These results show that this relationship, in which the galaxy age correlates positively with both mass and scatter, is captured by the neural networks and shows in the models' predictions. 76

3.6 The graph displays the relationship between the mass-weighted age of galaxies and their mass, for both central galaxies (diamond points, solid lines) and satellite galaxies (hexagonal points, dashed lines). Each plotted point and error bar represents the median and interquartile range of age values for a specific range of masses. The ages estimated from the predicted history of stellar mass assembly (blue) match the overall trend of ages with respect to mass, as computed from the original data (red). However, some error bars have been shifted or reduced, indicating that some of the samples have significant differences in their mass assembly geometry. 77

3.7 In a similar format to fig. 3.3, this figure shows the mass-metallicity relation for central galaxies (left) and satellite galaxies (right), where the original and predicted galaxies are shown in red and purple, respectively, and the errorbars indicate the median and 15^{th} and 85^{th} percentiles of stellar metallicity in a given bin of stellar mass. These results show that the neural networks predict a similar MZR shape in both datasets, but the scatter in metallicity is underpredicted, particularly at high mass. 78

3.8 Comparison of the predictions for the mean star formation history of central galaxies across six halo mass bins. The predictions were generated by two neural network models: a basic dense neural network shown in green, and a semi-recurrent architecture shown in purple. The median and interquartile ranges of the predicted mean SFH are plotted for both networks, based on ten independent runs of each network. The true mean star formation history in each bin is shown in blue, while the averaged mass accretion history is shown in red. The difference in interquartile ranges between the predictions of the two networks demonstrates a significant superiority of the semi-recurrent model to converge accurately. 79

3.9 The amplitudes of the Fourier transforms of the star formation histories (top panels) and metallicity histories (bottom panels) for central galaxies (left column) and satellite galaxies (right column), in three narrow bins of stellar mass. This shows the Fourier amplitudes for the original TNG data (red) against the Fourier amplitudes predicted by the neural network when trained to fit said amplitudes (blue). The quality of the predicted Fourier transforms constitute a superior fit to the original data than the star formation or metallicity histories predicted directly by the network (see fig. 3.2), and therefore the network can be used to produce a stochastic amendment to its own predictions. 84

3.10 Star formation histories (solid lines) and metallicity histories (dashed lines) of an exemplary intermediate mass satellite galaxy, showing the original TNG sample (red), the prediction of the neural network (blue) and the stochastically modified result (green). This shows a typical result where the fluctuative behaviour of the original galaxy is reproduced by the stochastic amendment, yet the overall shape of the star formation and metallicity histories is not significantly changed to match the target SFH and ZH. . . . 87

3.11 Absolute-valued Fourier transforms of the data in fig. 3.10, adopting the same choice of colours and linestyles, with the additional field of the Fourier amplitude predicted by the neural network shown in purple. This shows that the stochastic modification brings the Fourier transforms of the star formation and metallicity histories of this sample close to that of the TNG data, yet the predicted transform itself does not include distinctive features seen in the original transform, which in turn is absent from the modified result. 88

3.12 Stellar-halo mass relations, showing the original TNG data (red), the predictions of the neural network (blue) and the stochastically modified results (green). For each dataset, individual points represent samples, whereas errorbars indicate the median and $15^{\text{th}} - 85^{\text{th}}$ percentile ranges of stellar mass in bins of halo mass. The stochastic correction shows a modest improvement to the amplitude and scatter of the precdicted SHMR. 88

3.13 Stellar mass-metallicity relations, showing the original TNG data (red), the predictions of the neural network (blue) and the stochastically modified results (green). For each dataset, individual points represent samples, whereas errorbars indicate the median and $15^{\text{th}} - 85^{\text{th}}$ percentile ranges of metallicity in bins of stellar mass. The stochastic correction provides a significant improvement to the scatter of the MZR for central galaxies, yet it performs poorly for low mass satellite galaxies. 89

3.14 Median and interquartile ranges of mass-weighted ages as a function of stellar mass for central and satellite galaxies, showing the original TNG data (red), the predictions of the neural network (blue) and the stochastically modified results (green). Median stochastic mass-weighted ages are closer to that of the original data for most samples, showing that the correction improves the inferred geometry of the star formation histories; yet its smaller ranges show that these shapes are too generalised. 90

3.15 The Ξ values of the stellar-halo mass relation, mass-metallicity relation, and mass-metallicity history relation for each shuffle group are presented in the top, middle, and bottom rows, respectively; for both the central model (left) and satellite model (right). The median and interquartile range of Ξ values obtained from ten independent runs of each network are displayed in each cell's text. Illustrating the most critical shuffle groups, grid cells with smaller Ξ values are shaded in dark blue, transitioning to bright green as the median Ξ becomes larger. Higher Ξ values indicate that the given shuffle group is more important as the model determines the galaxy-halo relation, while a smaller interquartile range indicates greater significance of this result. 96

3.16 The tables presented in this section are similar to fig. 3.15, but instead highlight the effects of shuffling particular subsets within a shuffle group. Each table has three columns: the left column displays the results for the entire shuffle group, as shown in fig. 3.15. The center column shows the results for the input parameters being analysed, and the right column displays the results for the remaining components of the shuffle group. These tables indicate that overdensity is the primary factor in the second shuffle group for central galaxies, while the infall parameters in group 1a have a discernible impact on the satellite galaxy model. 100

3.17 Smoothed probability density functions of the baryonic data in an intermediate halo/subhalo mass bin; for stellar mass, metallicity, and mass-weighted age; based on the predicted star formation and metallicity histories. The distributions are obtained from the original TNG data (grey), the median of ten standard predictions (black), and the median of each randomisation (coloured lines). The distributions resulting from each randomisation do not exhibit a clear improvement over the fiducial data, nor do they bring the network any closer to the true distributions from the TNG simulations. If the horizontal axis is logarithmic, the PDF is a function of the logarithmic value labelled on the figure. 102

4.1 The response functions of the five optical filters used in the SDSS project. These are the weighting functions used to compute band fluxes from the spectra computed from our star formation histories. These filters are adjusted for atmospheric transmission with a typical airmass of 1.2 (Fukugita et al., 1996). The grey, vertical lines indicate the approximate range of optical wavelengths, showing that some of these these filters are sensitive to ultraviolet and infrared wavelengths. Each line is coloured according to the approximate perceived monochromatic colour of the band's effective wavelength, with the exception of the z band, which has no optical wavelengths. 111

4.2 The mean and standard deviation for stacked central (top row) and satellite (bottom row) spectra in bins of stellar mass, shown for predicted star formation and metallicity histories in green, and TNG data in blue. Emission lines have been omitted from these plots for clarity. In the majority of samples, the continuum is generally well recovered, and is of similar amplitude. However, for high mass objects there is a reduced variance at short wavelengths, and lower mass galaxies have a smaller variance overall. This represents a poorer prediction of central galaxy spectra, with lower mean amplitudes and smaller variance than the spectra evaluated from TNG data. 114

4.3 2D histograms of the fractional difference between true and predicted total galaxy luminosity and the mean high-frequency Fourier amplitudes of their star formation histories. These are shown for galaxies between the 75^{th} and 95^{th} percentiles of $z = 0$ star formation rate, and shows data within a frequency range of 0.3-1.2 Gyr^{-1}, i.e. a timescale range of 0.8-3.3 Gyr. This correlation between the two residuals indicates the dependence of the calculated luminosity on high-frequency star formation events. 115

4.4 Distribution of positive Hα luminosities evaluated from the original and predicted spectra, shown in relation to stellar mass, with contour lines indicating the tenth and ninetieth percentiles of the distribution of data, indicated by the legend at the bottom of the figure. Only galaxies with positive Hα fluxes are shown. Predictions show a modest fit to the mass-Hα luminosity relations for central and satellite galaxies, yet the scatter in both is underpredicted due to missing star formation points. 116

4.5 For the same galaxies as in fig. 4.3, this figure shows the correlations between residuals of their high frequency SFH data and their total Hα line luminosity. This indicates the importance of measuring short-timescale star formation events as in fig. 4.3, in particular at low stellar ages with the largest contribution of ionising photons. 117

4.6 Estimates of the five SDSS band magnitudes from the true and predicted spectra of both central and satellite galaxies, shown as a function of stellar mass, with contour lines indicating the tenth, fiftieth and nineteenth percentiles of their 2D distribution. These show a reasonable similarity in all bands despite a slight reduction in the variance of magnitudes in the predicted data. In both central and satellite data, the bimodal distribution of magnitudes can be seen in relation to mass. 118

4.7 Photometric colour distributions across the five bands, showing the differences between two consecutive bands. The distributions, mostly bimodal, are in rough agreement between datasets, however there are clear offsets in some of the data, such as bluer red galaxies in $g - r$, and significantly smaller predicted ranges. 119

4.8 Colour-mass distributions of the three galaxy datasets, shown for central and satellite galaxies, and for four colours evaluated from neighbouring band magnitudes. This shows the distinction between low mass "blue" galaxies and high mass "red" galaxies in all datasets, showing that the network predicts this relationship. However, there is a smaller range of colours in the predicted data which is not noticeably improved by the stochastic amendment. Contour lines indicate the tenth, fiftieth and nineteenth percentiles of the 2D distribution of each dataset. 120

4.9 For the data shown in figs. 4.3 and 4.5, this figure shows 2D histograms of the absolute difference between true and predicted $g - r$ colours and the mean high-frequency Fourier amplitudes of their star formation histories (top row) and metallicity histories (bottom row). This clear correlation indicates the importance of measuring short-timescale star formation and chemical enrichment events in the aim to calculate accurate colours. 121

5.1 This graph depicts the distributions of central halo and satellite subhalo masses in the entire Uchuu forest, shown in purple, and our cross-matched TNG-Dark sample, shown in green. By drawing samples from the former distribution according to the latter, we derive the distribution of Uchuu halos used in our study, represented by the orange histogram. The distribution of central halos closely resembles the TNG-Dark data, but the lack of well-defined satellite subhalos at low mass results in a skewed distribution of satellite subhalos in our Uchuu dataset. 129

5.2 This schematic illustrates the redshifts and cosmic time of the snapshots for the TNG simulations represented in red, and for the Uchuu simulation in blue. Despite having fewer snapshots than TNG, Uchuu has a higher temporal resolution during early times and is more sparsely sampled for $z < 2$. 131

5.3 This schematic illustrates how the mass accretion histories of central halos are distributed according to the halo mass and specific mass accretion gradient. The horizontal axis represents the halo mass and increases in value from left to right, while the vertical axis shows the specific mass accretion gradient and decreases in steepness from top to bottom. The solid lines display the median mass accretion history for each bin, while the shaded regions represent the $15^{\text{th}} - 85^{\text{th}}$ percentile ranges of the binned data. The mass accretion histories for the TNG-Hydro simulations' training and testing datasets are presented in red and blue, respectively, while the green and purple data correspond respectively to the TNG-Dark and Uchuu simulations. 134

5.4 Mass accretion histories of satellite subhalos, categorised similarly to fig. 5.3, with the satellite subhalo mass m_h replacing the central halo mass M_h, and the scaled accretion time a_{max} replacing the specific mass accretion gradient β. The same colour and percentile schemes used in the previous figure are adopted here. One significant difference between these accretion histories and the previous ones is that the satellite subhalos' accretion approaches zero or becomes negative, which is uncommon to central halos. The various times at which the growth of the median subhalo terminates is shown in various growth regimes. 135

5.5 This figure illustrates the evolution of the half-mass radius of central halos over time. It is presented in the same tabular format as in fig. 5.3, with identical halo mass and accretion gradient bins. 136

5.6 The half-mass radius growth of satellite subhalos, displayed in a tabular format similar to fig. 5.4, including the same bins for halo mass and scaled accretion time. 137

5.7 The histograms in each panel illustrate the probability density functions of logarithmic dark matter overdensities, contained in a spherical region of 5Mpc radius, surrounding halos at $z = 0$ in the four simulation datasets. The distributions of overdensity are comparable for most mass and accretion histories across the simulations, with the exception of the Uchuu dataset, which generally has higher densities due to the use of halo tracers. 140

5.8 This figure depicts the star formation histories of central galaxies grouped by halo mass and specific mass accretion gradient. In low mass bins, TNG-Dark overestimates star formation rates, while Uchuu underestimates them. In higher mass bins, the difference between the two simulations becomes less pronounced. 141

5.9 This figure displays the star formation histories of satellite galaxies grouped in the same manner as in fig. 5.8. The figure reveals seemingly inadequate predictions for the star formation histories of low mass galaxies in Uchuu. However, it should be noted that several of these bins in the Uchuu data have low population and are characterised by low-quality haloes. 142

5.10 This figure depicts the quantitative SHMR of central galaxies (left) and satellite galaxies (right) based on numerical integration their star formation rates. Each individual galaxy is denoted by a data point, and the error bars denote the median and the fifteenth and eighty-fifth percentiles of stellar mass within a specific halo mass range. The similarity in the shapes of these relationships implies accurate prediction of the star formation histories in dark simulations. 144

5.11 This figure shows the mass-weighted ages of central galaxies (left) and satellite galaxies (right) as a function of predicted stellar mass for the four simulation datasets, shown using the median and interquartile range of ages in different mass bins. This shows accurate recovery of the trend of age with mass in the dark simulations, yet there is a bias towards higher ages which increases with mass. 145

5.12 Metallicity histories of central galaxies, tabulated by halo mass and specific mass accretion gradient. This shows similar characteristics to the star formation histories in fig. 5.8, yet the low mass Uchuu samples are of particularly poor quality. However, the metallicity histories derived from dark simulations are generally very similar to the hydrodynamical predictions. . 146

5.13 Metallicity histories of satellite galaxies, tabulated according to subhalo mass and scaled accretion time. These show similar characteristics to star formation histories in fig. 5.9, with overprediction in TNG-Dark and underprediction in Uchuu. 146

5.14 The numerical mass-metallicity relation is presented for central galaxies in the left panel and satellite galaxies in the right panel, with total stellar mass and mass-weighted metallicity values obtained by use of eqs. (2.24) and (2.27). 147

5.15 The mean and standard deviation of the evaluated spectra from four simulation datasets are presented in three bins of low, intermediate, and high stellar mass. The top row displays the results for central galaxies, while the bottom row shows those for satellite galaxies. The comparison demonstrates that the two dark simulations for satellite galaxies are well-matched to the hydrodynamic TNG simulation, as they exhibit similar means and variances in the spectra. In contrast, the agreement for central galaxies is less robust, especially for high mass galaxies, where the Uchuu spectra have a higher variance, indicating less constrained stellar mass. To clearly display the mean continuum from each simulation, emission lines are omitted from these spectra. 148

5.16 Colour-Mass diagrams from the four simulation datasets, using four colours defined as the difference between successive SDSS band magnitudes, are shown for central galaxies (left panels) and satellite galaxies (right panels). The contour lines represent the tenth percentiles (light-coloured lines) and ninetieth percentiles (dark-coloured lines) of the 2D histograms. All four datasets show the anticipated photometric colour bimodality and its relationship with mass. The dark simulations, however, show an excess of samples between the peaks of the colour distributions. 150

List of Tables

2.1 A summary of the parameters of the twenty IllustrisTNG simulations (Nelson et al., 2019b) and four Uchuu simulations (Ishiyama et al., 2021). N_x represents the total number of particles or cells of component x, whereas M_x represents the size of one unit of mass in this simulation, i.e. the smallest resolvable mass. The components "dm" and "b" are dark matter and baryonic components, respectively. For each baryonic simulation, units of baryonic and dark mass are split according to the cosmic baryon fraction ($\Omega_\mathrm{b}/\Omega_\mathrm{dm}$), while in their dark equivalent simulations, these units are added together into one total mass unit. All simulations shown in this table use the parameters of the Planck Collaboration (2016) ΛCDM cosmological model. 36

2.2 A summary of the quantities used in both neural networks, grouped by layer and ordered by their placement in said layer. This entails the units of each quantity, and indicates which networks utilise them and how they are normalised. The section column indicates which section of this paper discusses this quantity. The shuffle group (final column) indicates which variables are simultaneously scrambled when testing for feature importance (see section 3.4). 45

4.1 Characteristics of the five SDSS bands for which we calculate photometric magnitudes from our galaxy spectra (Fukugita et al., 1996; SDSS Collaboration, 2002). 112

5.1 Spearman correlation coefficients of various halo characteristics with metallicity and stellar mass, in narrow bins of relatively low (sub)halo mass. All the parameters are evaluated at the final snapshot of the Uchuu simulation. The centrals are considered in a halo mass range of $11.77 < \log M_h^{z=0} < 11.97$, whereas the satellites are taken from a mass range of $10.9 < \log m_h^{z=0} < 11.1$. 143

1 Introduction

1.1 Galaxy Evolution

1.1.1 The Standard Cosmological Model

The history, matter and energy contents, evolution and structure formation of the observable universe are commonly parameterised by the Lambda Cold Dark Matter (ΛCDM) cosmological model. The principal conjecture is that the energy content of universe is comprised of four key components: baryonic matter, electromagnetic radiation, dark matter and dark energy (Planck Collaboration, 2016). The letter Λ refers to dark energy: a vacuum energy force which accelerates the expansion of space; an effect which is apparent on the largest observable scales of the universe (Perlmutter, 2000). The second most abundant component is cold dark matter: "cold" in that it has negligible, non-relativistic kinetic energy, and "dark" in that it is impervious to the electromagnetic force, and does not regularly interact with baryons via collisions (Peebles, 1982).

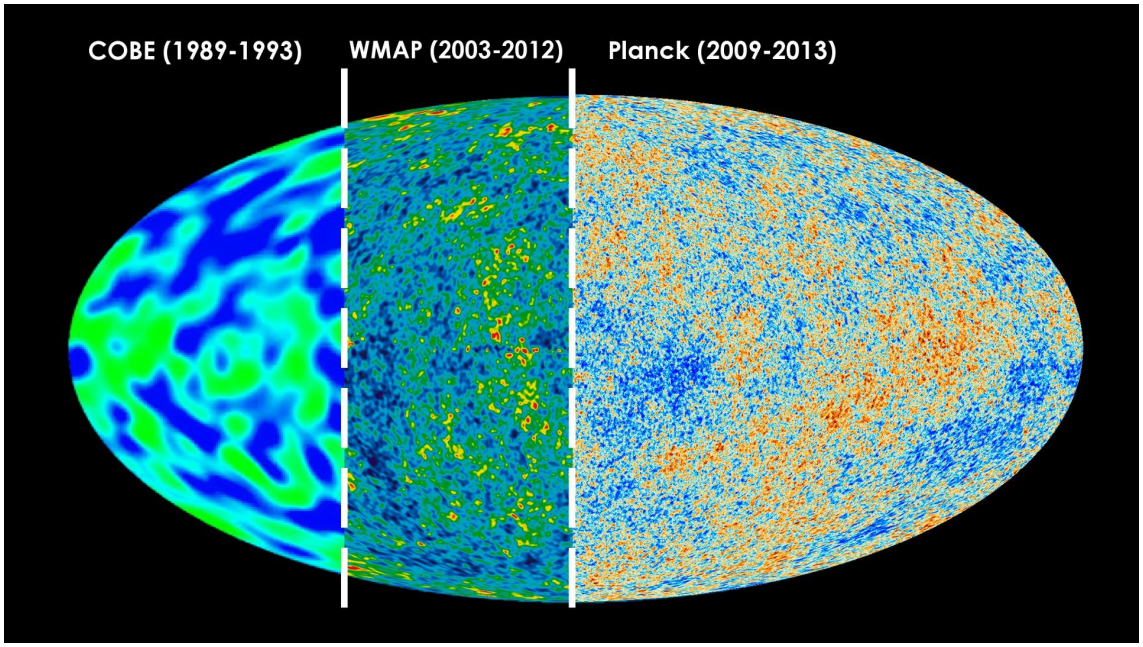


Figure 1.1: Thermal fluctuations of the cosmic microwave background as observed by the COBE (Boggess et al., 1992; Smoot, 1999), WMAP (Bennett et al., 2013) and Planck Collaboration (2016) satellites, showing the level of detail obtained over time. Though the cosmic microwave background is considered near homogeneous and isotropic, indicating a previous state of the universe in which it was condensed into a much smaller space, these perturbations on the scale of one part in 10^5 would evolve into the large scale structures seen in the present-day universe. Image courtesy of NASA.

According to ΛCDM, the universe began as a hot, dense plasma of matter and radiation, which underwent rapid expansion in what is known as the epoch of inflation (Guth, 1981). This expansion would eventually increase the average separation between photons and baryons, transitioning from an opaque plasma to a cool, transparent universe; where the interacting photons would decouple from the plasma, being observed today as the cosmic microwave background (CMB). This allowed protons and electrons to combine to form the first atoms, in an era known as the epoch of recombination; and for groups of matter, originating from small perturbations seen in the CMB, to collapse under their own gravity, forming gravitationally bound structures of dark and baryonic matter (Peebles, 1982; Blumenthal et al., 1984; Weinberg, 2008). The structure of the CMB as observed by telescopes of three consecutive genrations is shown in fig. 1.1.

One of the most intriguing classes of objects which forms from these collapsing overdensities is galaxies: the individual, gravitationally bound structures into which approximately 10% of the baryonic mass of the universe collapses (Fukugita & Peebles, 2004). Within them, hydrogen gas is converted into heavy elements by the formation of stars, which go on to form solar systems (White & Rees, 1978). As astronomical surveys in recent

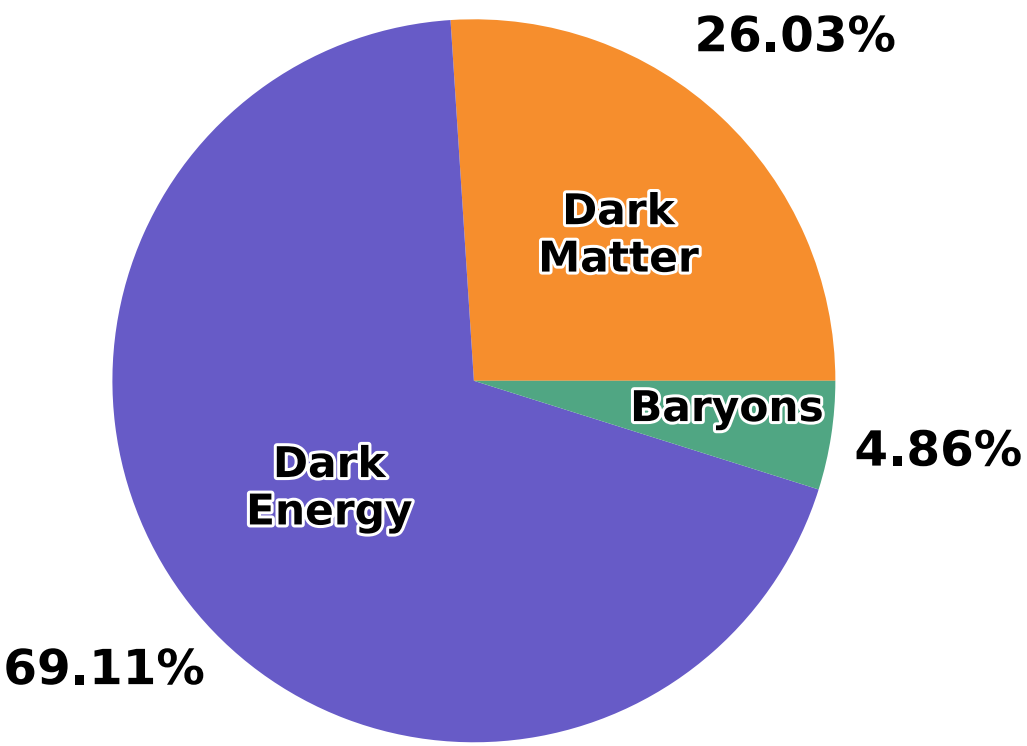


Figure 1.2: The ratio of energy densities of baryonic and dark matter and dark energy in the present universe, according to the Planck Collaboration (2016) cosmological model, which measured these densities by combining gravitational lensing, baryonic oscillation and CMB field cross-correlation data. The energy density associated with electromagnetic radiation constitutes a negligible quantity, and thus is not shown. These densities constitute fundamental parameters of the ΛCDM model. These values are assumed for the entirety of this thesis.

years have increased in sensitivity, more galaxies have been discovered very early in the universe's history, and individual galaxies have been mapped and analysed in increasingly fine detail. Galaxies are therefore useful for studying the mechanics of star formation and the large scale properties of the local and distant universe.

However, the dominant mass of a galaxy is not what we observe. The ΛCDM cosmological model argues that the majority of the mass of the universe is comprised of dark matter (Planck Collaboration, 2016), while the dominant form of energy in the universe is *dark energy*: a form of vacuum energy which drives the expansion of space (Peebles & Ratra, 2003). The relative densities of baryonic and dark matter and dark energy according to Planck Collaboration (2016) data is given in fig. 1.2.

The dynamics of galaxies and clusters show that the gravitational potential of a galaxy is dominated by some unseen potential which does not correspond to the potential of

the visible matter alone, and must be produced by an invisible, inert substance which interacts only via gravitational forces (Zwicky, 1933; Rubin & Ford, 1970; Rubin et al., 1980), otherwise known as dark matter. The invisible object in which galaxies are nested which produces this potential is commonly known as a dark matter halo.

1.1.2 Formation Of Cosmic Structure

After recombination, matter was scattered randomly throughout the universe. In theory, the universe began as a homogeneous, isotopic fluid (Bartelmann, 2010), but the CMB exhibits density perturbations in the cosmic fluid, shown by fig. 1.1. While on the scale of only $\sim 10^{-5}$ times the mean density of the universe, these fluctuations have developed into regions of high concentrations of matter, and by contrast, regions of space with little to no matter at all (Schneider, 2015). These overdense or underdense regions would evolve much like a smaller universe of their own, with different modes of expansion. Areas with little gravitating matter would expand more rapidly and potentially even source dark energy (Yusofi et al., 2022), while dense volumes of space would be dominated by their local gravity, and collapse into a gravitationally bound ensemble.

Dark matter halos, which form from the gravitational collapse of overdensities, are the fundamental building blocks of cosmic structure. As these halos originated from overdense regions of space, their gravity would serve to accrete the surrounding material, causing these halos to continually grow in size (Bartelmann, 2010). The attraction between any two halos would cause them to grow more and more rapidly through merger events: the collision of two or more halos, which can have profound effects on the internal structure and dynamics of the merging objects (Robaina et al., 2010; Welker et al., 2014; Hani et al., 2020; McAlpine et al., 2020). The result of this mechanism of halo evolution is a hierarchical scheme of structure formation: small halos would coalesce to form large halos, large halos would produce massive halos, and so on (Schneider, 2015). This timeline of structure evolution from smaller to larger objects is an artefact of the ΛCDM model, whereas energetic "hot" dark matter favours the growth of larger structures first, fragmenting into smaller systems over time (Dodelson, 2003; Frenk & White, 2012).

The anisotropic initial conditions for this structure formation would advance into highly non-uniform structure growth. The extreme gravity of multiple high mass halos would result in clusters: groups of hundreds to thousands of galaxies amassing around

10^{14-16} solar masses ($M_{\odot}$). By the same mechanism, halos can form along the pathways between these high density regions, which form cosmic filaments bounded by expanding voids (Bond et al., 1996). This network of large scale structures, composed of dark matter halos and illuminated by their galaxies and intergalactic medium (IGM), forms the distribution of the present-day universe which is named the cosmic web.

Forming from different quantities of matter in different gravitational environments, there is great diversity of growth and interaction histories of dark matter ensembles and the galaxies within them. As the dominant mass of a halo-galaxy system, the evolution of halo properties over cosmic time are bound to have a profound effect on the properties of the galaxies which live in and interact with said halo (Wechsler & Tinker, 2018). The surrounding clusters, filaments and other large scale structures further affect how these halos and galaxies evolve, either by growing their mass in different ways or channeling star-forming gas into the galaxy (Poudel et al., 2017; Veena et al., 2018; Castignani et al., 2022; Donnan et al., 2022).

Galaxies, however, are far more complex objects than any dark matter structures. While halos and cosmic filaments are governed entirely by gravitational interactions, galaxies are subject to a long list of processes which affect their mass, size, shape, luminosity, colour, chemical composition, star formation activity, gas loss and other physical and observational features (Somerville & Davé, 2015; Vogelsberger et al., 2020). A popular open question in astronomy asks how accurately and precisely the behaviour of galaxies can be explained by their dark matter components alone.

1.1.3 The Properties Of Galaxies

Opposed to the dark matter halo which dominates the gravitational potential of a galaxy, a galaxy is predominantly comprised of three phases of baryonic matter: gas, stars and black holes. Galaxies begin as clouds of hydrogen gas, gravitationally bound to their dark matter halo. Where this gas becomes cool and dense, it collapses under its own gravity and heats up via self-friction. These regions of collapse eventually reach the necessary temperature and density to achieve nuclear fusion; initiating the conversion process of gas into stars.

The internal fusion within stars converts this hydrogen into helium, and hydrogen

and helium into metals. In astronomy, "metals" refer to elements heavier than either hydrogen or helium, and "metallicity" refers to the fraction of a baryonic mass, such as a collection of gas, or stars, which consists of metals. The metallicity of a system can be calculated from the theoretical effective yield of metals of a specific star-forming region (Lia et al., 2002; Chruslinska & Nelemans, 2019); or as an average over units of gas or stars, weighted by their mass, star formation rate, luminosity, or other properties gauging the abundance of matter per unit (Tantalo & Chiosi, 2004). Alternatively, metallicity can be constrained observationally by measuring the relative abundances of elements from relative line emission luminosities or widths (Nagao et al., 2011; Kewley et al., 2019). In this thesis, references to metallicity consider a stellar metallicity weighted by the mass of star particles in a cosmic simulation (see section 2.4.2) unless explicitly stated otherwise.

Elements heavier than iron are created in supernovae: the explosions of high mass stars resulting from the imbalance between gravitational collapse and electron degeneracy pressure. Lower mass stars never reach this stage in their lifetime, and instead lose most of their mass into the interstellar medium (ISM) by stellar winds, by which time they have formed lighter elements overall. The contrast in the evolutionary properties of different stars results in very different properties of old and young, high and low mass stars, and consequently, galaxies exhibit very different stellar populations depending on their star formation history, and emit very different spectral energy distributions (SEDs/spectra) accordingly.

Recently formed galaxies have had little time to form stars, and so the stellar populations they do have include hot, massive blue stars which ionise the surrounding gases; or cool, low mass stars such as red dwarf stars, which exhibit very different evolutionary paths due to their difference in size, mass and luminosity. Over time, the supply of star-forming gas will increase due to the accumulation of mass, yet eventually the supply will be used up, slowing star formation to a halt. Additionally, the collision with another galaxy can heat and redistribute this gas, making it unusable for star formation. Older galaxies are therefore typically more massive and redder in colour due to the abundance of old, cool, red stars; and the abscence of high mass, UV-luminous stars which have a shorter lifespan (Gallazzi et al., 2005). These aged galaxies also have a stellar population of assorted sizes, masses and colours, due to the distinct evolutionary trajectories of high and low mass stars.

The distribution of galaxy colours is in fact bimodal in nature: "blue cloud" galaxies are separated from "red sequence" galaxies by the "green valley": a significantly less populated phase where galaxies are undergoing rapid changes such as increase in mass through collisions with other galaxies, transformation from spiral to elliptical morphologies, and swift decline in their star formation activity. The observed colour of galaxies is therefore reflective of several dichotomies in the physical properties of the galaxy population (Baldry et al., 2004).

The energetic radiation emanating from stars serves to ionise the surrounding gases. Aside from the heating and acceleration of the gas resulting from this, metal-rich stars can deposit their contents into the ISM, enriching the gas with metals. The ionising radiation produces emission lines when interacting with the gas, which are unique to each element, and therefore the strengths and ratios of different emission lines are a practical tracer of the chemical composition, density, temperature and star formation rate of the gas. These lines can also be broadened by the Doppler shift resulting from the motion of the gas, and shifted due to gravitational or cosmological redshifts, and therefore are valuable probes of the kinematics, densities and distances to their galaxies.

For more massive galaxies, a large black hole typically forms in the galactic centre, and by accreting matter from its surroundings can produce a bright, compact region of emission, known as an active galactic nucleus (AGN). The black holes which source the AGN are believed to form through a combination of processes that involve the collapse of massive clouds of gas, the merger of smaller black holes, and the growth of existing black holes through accretion of matter. For galaxies hosting highly luminous AGN, commonly known as quasars, there are substantial, fundamental differences in their spectral features due to the AGN outshining the galaxy almost completely. Most notably, quasar spectra are highly luminous at short wavelengths, exhibit a characteristic power law shape and have strong, broad emission lines owing to intense ionisation of gases in the accretion disk. The class of AGN-hosting (active) galaxies other than quasars is the Seyfert galaxy, where the AGN emission is weaker and the host galaxy remains detectable.

In summary, the numerous physical processes which take place in a galaxy's evolutionary history, from the hierarchical builup of the large dark matter halos which contain them to the onset of collisions and gravitational capture of nearby galaxies, are identifiable in

several of the galaxy's spectroscopic and photometric features. By conducting surveys of the galaxies which we see in the sky, we can observe these features and draw conclusions about how these galaxies have grown. The interplay between these phenomena is nonetheless very complex, and in order to understand these observational statistics, we have to model this enigma with great precision and accuracy. Fortunately, the dark matter component of galaxies and their surroundings have a profound influence on galaxy evolution, and the relative simplicity of dark matter structures can offer a relatively simple approach to modelling galaxy evolution, while providing useful information about the convoluted relationship between galaxies and their halos and environment.

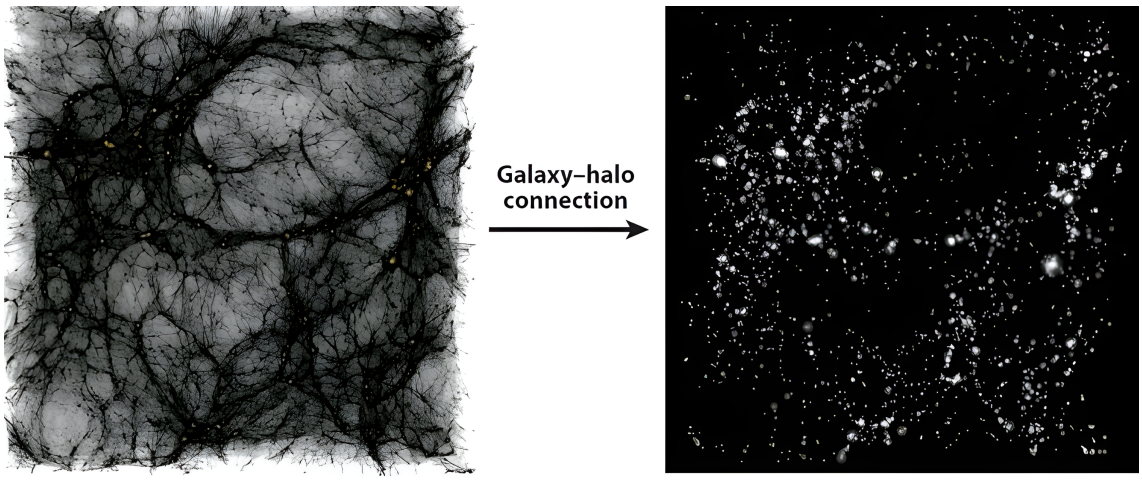
1.2 The Galaxy-Halo Connection

With dark matter constituting approximately 84% of the mass of the universe, as measured by the Planck Collaboration (2016), the halos and cosmic filaments which naturally manifest from dark matter over time form the framework for galaxy evolution. Over time, the gas would cool and condense enough to undergo nuclear fusion, creating stars, and these stars would coalesce to form galaxies.

The behaviour of the halo and its surroundings have profound effects on how its galaxies evolve over time. As halos collided with each other and merged into larger objects, the gas contained in these halos would be rapidly compressed and form stars at an accelerated pace. As halos gain angular momentum, much of the mass of the galaxies is moulded into bar structures, funneling much of the star-forming gas into the galactic centre, as shown by Saha & Naab (2013). Gas is also accreted onto galaxies in halos in close proximity to cosmic filaments, and taken by larger halos during flybys. The dark matter component of galaxies and cosmic structure has a profound influence on the properties of the galaxies themselves, from interstellar to cosmological scales.

1.2.1 Modelling Approaches

Astronomers have long sought to understand the logistics of the galaxy-halo connection (GHC), being of great importance for understanding not only how halos and galaxies evolve, but for understanding the nature of cosmic expansion and the conditions of the early universe. Approaches to modelling the GHC have largely consisted of statistical modelling, either through assigning observational galaxy properties to halo properties in N-



Approaches to modeling the galaxy–halo connection

Physical models ← → Empirical models

Hydrodynamical simulations	Semianalytic models	Empirical forward modeling	Subhalo abundance modeling	Halo occupation models
Simulate halos and gas; star formation and 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 and subhalos) plus assumptions about galaxy–(sub)halo connection	Collapsed objects (halos) plus model for distribution of galaxy number given host halo properties

Figure 1.3: A visual summary of empirical and physical methods of modelling the galaxy-halo connection in cosmological simulations, as shown in Wechsler & Tinker (2018). The image on the left shows the dark matter distribution of a $90 \times 90 \times 30\text{Mpc}/h$ region of an N-body simulation, initialised from a random seed. The image on the right shows the distribution of galaxies assigned to halos in this simulation, based on an abundance matching model (see section 1.2.1) adjusted to match observational data. The scale below lists several modelling strategies ranging from physical models on the left to empirical models on the right, additionally listing basic techniques and assumptions.

body simulations, or full simulations of the galaxy formation physics within a cosmological volume. A schematic illustrating typical methods of computing the GHC is shown in fig. 1.3.

Empirical Modelling

In observational studies of galaxies, empirical models of the GHC have provided insights into our understanding of the abundance of halos, the number of satellites of a given halo, the clustering statistics of halos, and numerous other properties which go beyond the visible statistics of galaxies. These constraints have gone further to describe the evolutionary tracks of galaxies which occupy certain halos and large scale environments, by developing such models which incorporate historical properties of halos and galaxies. The advantage of these empirical models is in their simplicity: the assumptions or probabilistic mappings between galaxies and halos can be applied to volumes of any size. On the contrary, their predictions depend fundamentally on assumptions on galaxy and halo evolution which may cease to be valid in unseen circumstances, and are limited in their flexibility by the known results of the variable parameter range.

The galaxy population of a particular halo or cluster of halos can be parameterised by means of a halo occupation distribution (HOD) (Berlind & Weinberg, 2002): a conditional probability function of the expected number of galaxies within a halo, typically of a certain mass, and as a function of just a handful of halo parameters (Paranjape et al., 2015; Hearin et al., 2016). As a function of galaxy properties (morphology, luminosity, etc.) it is a powerful tool for relating the abundance and distribution of distinct galaxy subtypes to the evolutionary history of their halos (Zheng et al., 2005). Similarly, one can predict the intrinsic properties of galaxies by evaluating a probability distribution modulated by halo properties. An example is the conditional luminosity function (CLF): a measure of the distribution of galaxy luminosities based on halo mass (Yang et al., 2003).

Abundance matching is another popular empirical technique in which a heirarchy between halos and galaxies is assumed; most simplistically, halos and galaxies being ordered identically according to their mass (Kravtsov & Klypin, 1999; Guo et al., 2010). Ordered properties of halos and galaxies are then used to produce statistics based on this assumption. These models are typically updated to include additional halo properties which are shown to strongly influence galaxies, and vice versa. A similar algorithm is subhalo abun-

dance matching (SHAM) (Kravtsov et al., 2004; Vale & Ostriker, 2004; Chaves-Montero et al., 2016): in which subhalos above a certain mass threshold are expected to host galaxies, and the central, massive halos which harbour subhalos will host a certain number of galaxies depending on the central halo. By specifying the relationship between galaxies and subhalos, Abundance matching methods can be used to accurately predict galaxy properties, such as the stellar masses and star formation rates of galaxies within the host halo (Guo et al., 2010; Simha et al., 2012; Chaves-Montero et al., 2016).

An example of a galaxy evolution model with an empirical framework is UniverseMachine, developed by Behroozi et al. (2019a), where star formation rates of nested galaxies are parameterised in terms of halo potential and redshift, which are integrated over the halo's merger tree to recover population statistics such as the stellar mass function, and establish clear galaxy-halo correlations, such as between galaxy and halo growth, and between halo mass and quenched fractions at different redshifts. While these results are indeed observationally accurate and insightful into the specifics of the GHC, they depend significantly on general assumptions about the histories of unique galaxy populations, such as a dichotomy between the mean star formation history of quenched and star-forming galaxies (Behroozi et al., 2019a).

Empirical models in general can fail to reproduce the results of more robust implementations, such as the correlation functions of mass and colour separated galaxy populations (Hadzhiyska et al., 2020). Modelling galaxy evolution using fundamental physics rather than statistical inferences can reproduce similar results while naturally incorporating the complete history of the galaxy and halo's growth and interactions.

Physical Modelling

In cosmic simulations of galaxy evolution (examples: Schaye et al., 2014; Lacey et al., 2016; Davé et al., 2019; Nelson et al., 2019b; Henriques et al., 2020), a set of particles is initialised at an extremely high redshift, and propagated in time according to a series of equations of motion, down to the present universe. These simulations are typically conducted in a cubic, comoving volume with periodic boundary conditions, and similar simulation runs, such as those differing only in resolution or some physical parameter, are initialised from the same random seed, creating near-identical halos and large scale structures. This allows one to compare the properties of the same halos, clusters and

galaxies under the variation of this parameter.

Some of these cosmic galaxy simulations fall into the class of semi-analytic models (SAMs). This approach invokes an analytical formula to prescribe baryonic properties and processes based on the properties of halos in an N-body dark matter simulation (White & Frenk, 1991). The benefit of these models is their computational efficiency; they can be used to perform numerous calculations such as gas cooling, star formation rate or stellar feedback from properties like the mass and merger history of the host halo. However, these formulae for galaxy evolution are based largely on approximations of the baryonic physics, and the results are dependent on a large number of parameters which are difficult to fine-tune with observations (Saghiha, 2017). These prescriptions also do not directly resolve galactic substructures, and so are only a practical tool for the macroscopic galaxy-halo connection.

The most self-consistent approach to modelling galaxy evolution is with hydrodynamical simulations, which, unlike SAMs, directly include discretised units of baryonic mass, such as moving mesh cells of the gas density field, alongside dark matter particles in the simulation. This baryonic component is propagated simultaneously with the dark matter component, according to both gravitational equations of motion and hydrodynamic equations of state (Wechsler & Tinker, 2018; Vogelsberger et al., 2020). As well as providing physical realisations of galaxy diversity at all times following the initial redshift, the morphological components of galaxies are resolved in detail, allowing the inflow and dynamics of stars and gas to be studied using these simulations. These simulations are also unique in that the effects that baryons have on the structure of dark matter halos are an explicit result of the model, and so these simulations are usually accompanied by an equivalent N-body run to illustrate the difference that is made.

Despite the advantages of hydrodynamical simulations, they, like N-body simulations, are subject to resolution effects, and again the behaviour of these simulated galaxies is governed by approximations of the physics below the resolution scale; typically known as a "subgrid model" (Wechsler & Tinker, 2018; Vogelsberger et al., 2020). Hydrodynamic simulations are computationally expensive, and there is usually a compromise to be made between the volume of the simulation, encompassing large scale structure, and the resolution of the simulation, resolving the internal structures of galaxies. This broad

range of physical scales includes a long list of aspects of the GHC, discussed at length in section 1.2.2.

One solution to this problem, in the context of fine resolution of objects dispersed in high volumes, is the use of zoom simulations: a special class of cosmological simulations (for various examples see Lovell et al., 2020; Roca-Fàbrega et al., 2021; Wetzel et al., 2022; Nadler et al., 2023) which compute low resolution simulations of the full cosmological volume, and then rerun with high resolution in selected regions of space. These simulations are practical for modelling individual galaxies or clusters in high resolution, and have provided valuable insights into the properties of unique galaxy populations from dwarf galaxies to high redshift clusters. The compromise is of course in the loss of large scale information through resolution reduction, which implies that the influence of cosmic structure, however relevant to galaxy and cluster evolution, is neglected in these simulations. Using hydrodynamical simulations to self-consistently compute the GHC on all scales of the ΛCDM cosmological model is beyond the power of today's computational resources.

1.2.2 Halo And Galaxy Histories

In spite of the various shortcomings of all of the aforementioned models of the galaxy-halo connection, they remain extremely valuable to our interpretation of the various halo properties which drive different aspects of galaxy evolution. Several of the most important properties, correlations and processes involved in the GHC are described below.

Halo Mass

Abundance matching models of the GHC typically invoke the premise that the most massive galaxies are bound to the most massive halos. While there are inevitably secondary correlations between galaxies and other halo properties, this assumption alone describes several galaxy properties very accurately.

One important result of the GHC is the clear trend between the total mass of stars formed in a galaxy and the mass of its halo - a relationship known as the stellar-halo mass relation (SHMR). The stellar mass of a galaxy additionally relates to key properties such as the morphology, star formation activity and environment of said galaxy, and thus the SHMR is closely related to distinct classes of galaxies; for instance, red, elliptical

galaxies typically poulate the high mass end of the relation, or have above-average stellar mass at intermediate halo masses (Correa & Schaye, 2020). Halos and galaxies of any given redshift or central/satellite phase possess a unique SHMR, highlighting the typical efficiency of star formation at the given epoch and mass regime. Similarly to the SHMR, the total luminosity of galaxies is very tightly correlated with halo mass and is additionally a useful observational constraint of halo properties (More et al., 2009).

The SHMR, while showing a simplistic scaling relationship between galaxy and halo masses, is not a tight relation between galaxies and halos, and does not describe galaxy evolution in its entirety. The variance in galaxy masses of a fixed halo mass, referred to as the "scatter" of the SHMR, is an inherent property of the data and is a result of distinct evolutionary histories of these galaxies. Galaxies with above-average scatter have undergone rapid star formation to reach their stellar mass, and those with below-average scatter have formed stars inefficiently for galaxies of their halo mass. Halo properties other than their present mass with well-defined correlations with this scatter are therefore considered important for developing the mass of galaxies.

Other than stellar mass, the halo mass can influence a number of additional galaxy properties. A simple inference from the SHMR is that the halo mass dependence of star formation efficiency dictates the rate of metal synthesis in these galaxies, and thus there is a similar relation between halo mass and stellar metallicity. A relationship which is not so intuitive is the prevalence of quenched galaxies in higher mass halos, which owes to several factors such as the interaction history (Davies et al., 2022) and internal dynamics of the halo (Bluck et al., 2020).

These secondary halo properties can introduce additional differences, such as the enhancement of the clustering of halos of a given mass, known as halo assembly bias, which in turn introduces bias in the distribution of galaxies of a given mass. The halo mass is thus a powerful but not absolute measure of how galaxies evolve, and the properties discussed below are known to contribute to the additional diversity of galaxy populations.

Mass Accretion History

The time at which halos begin to form and the rate at which they grow in size and mass is additionally important to the growth of galaxies. Halos can grow smoothly through the

accretion of matter from their surroundings, or rapidly through collisions with other halos. These events in the halo's growth history have unique effects on their galaxies, and the future evolution of the halos themselves. The mass accretion history is thus a fundamental component of most galaxy formation models, yet due to its complexity it is difficult to characterise with a small set of parameters.

The time at which a halo forms a significant fraction of its final mass, or maximum mass achieved throughout its lifetime, is widely used to characterise the age of a halo and its characteristic rate of mass accretion. This definition is subjective, and multiple definitions are used by different studies, showing different degrees of correlation with different galaxy properties (Tojeiro et al., 2017). Half-mass formation time shows clear correlation with the scatter of the SHMR (Cui et al., 2021), which indicates the rapidity of halo mass accretion to correlate significantly with the buildup of star-forming gas; while Zhao et al. (2009) show a clear scaling relation between the time of formation of 4% of the final halo mass and the increase in the concentration of halo density over time, indicating the formation of central structure to manifest early in the halo's history through rapid accumulation of mass. Formation times are therefore a useful quantity to relate the mass accretion history to the GHC in parametric form, whether this indicates a direct or indirect effect on the galaxy; however other parameterisations of accretion history show further distinctions in galaxy properties, such as their gas fraction or photometric colour (Shi et al., 2020; Montero-Dorta et al., 2021).

Further to smooth accretion of material from the cosmic web, halos can change their mass in much shorter frames of time by means of interaction with other halos. Merger events refer to the coalescence of halos as they collide, which in turn have dramatic effects on the galaxies residing in merging halos, yet the nature of this influence on galaxy evolution depends on the scale of the merger event (Lambas et al., 2012; Ownsworth et al., 2014).

A minor merger event is one in which the larger of the two halos is much greater in mass than the other, and the acquisition of the smaller halo has little effect on it or its galaxy. The galaxy of the smaller halo, on the other hand, can become a satellite of the central galaxy and be subject to effects such as tidal or ram pressure stripping of its gas and stars, or can lose much of its mass to the larger halo in a flyby event. Either way, the

event has a profound and permanent effect on the future of its galaxy evolution.

A major merger event is one in which the masses of the colliding halos are similar, and the effect on each halo and galaxy is just as great. These objects merge to form an object much greater in mass than either of its progenitors, and of course undergo entirely different modes of mass accretion afterwards. The galaxies can be affected in that their gas can be rapidly compressed by the shockwave which results from the merger, leading to swift acceleration of their star formation. On the contrary, the merger can rapidly heat this gas and thereby prevent it from condensing and forming stars, or fuel the central black hole of the galaxy which expels this gas from the system with its relativistic jet. Major mergers are rarer occurrences than the minor events which collectively equate to smooth accretion and satellite infall, but individually they are some of the most important events in a galaxy's history. Yet the extent to which merger events vary in their frequency, time of occurrence, progenitor mass ratio, and stellar and gas mass fractions, adds to the difficulty of understanding their role in the galaxy-halo connection.

Internal Structure And Dynamics

Dark matter halos exhibit a simple, radially symmetric structure when in thermal equilibrium. The most common function used to describe the radial density profile of a dark matter halo is the Navarro, Frenk, & White (1996) (NFW) profile. This can be defined in terms of the concentration parameter c, which indicates the density of matter towards the centre of the halo. This is defined as follows:

$$c = \frac{r_{\rm vir}}{r_{\rm s}} \tag{1.1}$$

where $r_{\rm s}$ is the scale radius of the NFW function (eq. (1.2)), indicating where the gradient of the density profile represents an isothermal sphere; and $r_{\rm vir}$ is the virial radius of the halo: the radius enclosing a region of sufficient density to collapse under its self-gravity.

The NFW profile can be written:

$$\rho(r) = \frac{\rho_{\rm crit}\ \delta_c}{\frac{r}{r_{\rm s}}\left(1+\frac{r}{r_{\rm s}}\right)^2} \tag{1.2}$$

where $\rho_{\rm crit}$ is the critical density of the universe and δ_c is the characteristic density contrast of the halo, defined as follows:

$$\delta_c = \frac{200}{3} \times \frac{c^3}{\ln(1+c) - \frac{c}{1+c}}. \tag{1.3}$$

A similar and commonly used function is the Einasto (1965) profile:

$$\rho(r) = \rho_{\rm s} \exp\left[\frac{2}{\alpha}\left(1 - \left(\frac{r}{r_{\rm s}}\right)^{\alpha}\right)\right] \tag{1.4}$$

which is a better fit to halos with gradually changing gradients (Navarro et al., 2004). α represents a gradient parameter which is a function of halo mass and redshift, and $\rho_{\rm s}$ is the density at isothermal radius $r_{\rm s}$.

These profiles provide a good fit to the density profiles of virialised halos. The concentration parameter is a useful quantity; as it can describe the smooth distribution of dark matter, or the radial number density of subhalos within a halo of galaxy or cluster scale mass. The concentration parameter follows a scaling relation with halo mass at any given redshift, which, at low redshifts, shows higher mass halos to have lower concentration, while this relation becomes flatter at earlier times (Child et al., 2018). The concentration parameter is also tightly related to the formation time of the halo, and shown to grow with time as a result of smooth accretion (Zhao et al., 2009). It is a practical indicator of the mass accretion histories of halos of a given mass.

These profile fits, however, represent an idealised halo, for which the mass, shape, size and internal dynamics are self-consistent. In reality, halos vary in their shapes and their velocity structures, which can be reconciled with the tidal distortions induced by large scale structure and merger activity, or AGN and supernova driven winds and outflows (Chua et al., 2022).

The dispersion, or variance in the velocities of halo particles, is tightly correlated with the halo's total mass if the system is virialised, which makes this a practical observational

measure of the gravitational potential. The scatter in this relation signifies the non-virial dynamics of the halo and its satellites, which is usually measured by velocity anisotropy: quantifying the fraction of matter which is moving isotropically, or radially with respect to the halo centre, as opposed to a tangential orbit. This motion is often the result of tidal anisotropy within the local environment or internal to the halo, and has significant effect on the morphology and distribution of local galaxies. Furthermore, this anisotropic density field can enhance or suppress mass accretion and merger rates, again having notable effects on the galaxies' star formation and chemical enrichment.

Cosmic Environment

"Cosmic environment" refers to the density, distribution and dynamics of subhalos and large scale structures which surround a halo of interest. Environments are typically characterised using overdensity: the density of matter within a spherical region of space centered on the target halo or galaxy, as well as the properties of nearby nodes, filaments and voids, and measures of tidal anisotropy owing to the asymmetric spatial and velocity distribution of matter in the local vicinity. Different cosmic structures form from the coalescence of halos according to the history of their environment (see section 1.1.2).

The differences in the conditions within regions of different densities and distributions of halos is very important to the properties of galaxies in these regions, which have distinct star formation, morphological and chemical properties depending on their environments (Scoville et al., 2013; Papovich et al., 2018; Galárraga-Espinosa et al., 2023). In denser regions, e.g. cluster centres, galaxies are typically more massive, elliptical in shape, have low rates of star formation and are predominantly comprised of old, metal-rich stars and gas. Galaxies in less dense environments usually have ongoing star formation, spiral morphology, and a recently formed, metal-poor supply of stars.

The relationship between galaxy properties and their local environment is complex, and difficult to characterise with only parameterisations of the surrounding mass distribution. It is in fact believed to result from various physical processes, such as the presence of gas for star formation and the magnitude and frequency of gravitational interactions between halos. In high-density areas, star-forming gas can easily become heated by shocks or expelled by ram pressure, restricting a galaxy's ability to condense its gas reservoir and form stars. Additionally, the presence of numerous galaxies in dense environments can

result in mergers and other interactions, which as discussed in section 1.2.2, can significantly alter the properties of the involved galaxies. Conversely, in lower density regions, the gas is able to cool with little intrusion, leading to an increased rate of star formation.

It is clear that halos which reside in different cosmic environments will have distinct evolutionary histories, as well as the galaxies which they host, having grown in regions of different density and tidal asymmetry, affecting their mass accretion and merger history. It is well established that this is true for galaxies which occupy a known, low-redshift environment, however the history of the cosmic structures themselves are likely to have played a significant role in the galaxy's development, by governing the rate of interactions, supply of gas and so on. In this thesis, we have included measures of the cosmic environment as a function of time as a predictor of galaxy evolution, discussed in chapter 2.

Central And Satellite Phases

The gravitational capture of a galaxy by a much larger halo or cluster has substantial effects on its future, leading to significant differences between the statistics of central and satellite galaxies. Galaxies which have remained central galaxies throughout their lifetime typically grow through the accretion of matter and merger and flyby interactions with other galaxies, and lose their mass predominantly via supernova-driven winds and AGN jets. Satellite galaxies, by definition, fall into the gravitational potential of a much larger halo, and are subject to dynamical friction as it traverses the halo, and as discussed above, is more likely to be subject to tidal and ram pressure stripping in regions of higher density.

The properties of satellite galaxies are expectedly very different from central galaxies, having lost much of their star forming gas and mass in their outer regions, and having become morphologically distorted by the host halo and other satellites. The local environment of satellite galaxies is in fact critical to their quenching (Bluck et al., 2020). For central galaxies, cosmic filaments play a role in sourcing their star-forming gas, which is something which does not significantly influence satellites (Simpson et al., 2018). The different importances of different quantities illustrates the distinction in evolutionary physics between central and satellite galaxies.

An important quantity deciding the fate of satellites is their mass with respect to their host: lower mass satellites are more vulnerable to environmental quenching effects due

to weaker gravitational binding energy. The time of their infall into the central halo's potential is additionally important; these stripping effects are greater on galaxies which have spent more time within the halo, and on galaxies which grew little in baryonic mass prior to their infall. The gravitational forces and ram pressure are also stronger for satellites which follow a radial trajectory towards the halo centre, or at higher velocities. For the central galaxy, the trajectory of these satellites can determine whether these will remain as satellites, exit the potential as a flyby interaction, or merge with the central galaxy, and the gas content of the infalling satellite can enhance its own star formation.

1.3 This Thesis

1.3.1 Motivation

Decades of research have established that the physical and observational statistics of galaxies are dependent on a wide range of phenomena, yet a substantial part of this is either directly or indirectly modulated by dark matter halos. The mass accretion and merger histories of their host halos, the gas and star content of the many millions of halo progenitors, the tidal forces provided by halos, clusters and filaments in their local region, are just some of the factors which contribute to the wide complexity of galaxy evolution.

It has long been established that the halos and cosmic environments which encase galaxies and clusters have made a dramatic impact on the local galaxy population. The time at which these halos form, the preferential orbits of their interior and exterior companions, and the quantity of material which they accumulate over time are all factors which influence how quickly their galaxies form most of their stellar mass, how their disks and bulges develop, and how these galaxies will affect the galaxies which interact with them. Despite valuable insights into certain correlations, e.g. galaxy colour bimodality being driven by the cold gas content of early-forming halos (Cui et al., 2021), the number of potential mechanisms of galaxy-halo coevolution is multifarious. There may be several phenomena, such as the collapse of cosmic filaments, which influence galaxies in an as-yet unexplained fashion.

Modelling the coherent properties of galaxies and halos on all scales, in an effort to explain the GHC in greater detail, has been a challenging endeavour. Hydrodynamical simulations incorporate numerous computationally intensive codes to compute the cooling

of gas, formation of stars, stellar and AGN feedback and other processes. This results in a tradeoff between the volume of the simulation, used to model massive galaxy clusters and large scale structure; and the mass resolution of the simulation, which is imperative to modelling sub-kiloparsec galactic disks. Semi-analytical models may allow galaxies to be modelled on larger scales due to their superior efficiency, but they are not self-consistent and rely on optimisation of a wide parameter space.

The size and diversity of a large, detailed cosmological simulation would advocate the comprehensive study of the correlations of the galaxy-halo connection, describing in detail the physics of galaxy formation and evolution from the dark matter perspective. Furthermore, galaxy surveys in recent years have surveyed larger regions of the sky and probed to higher redshifts as instrumental sensitivity has advanced. These surveys have benefitted from high fidelity baryonic simulations in the production of mock surveys (mocks), using the baryonic content of simulated galaxies to synthesise SEDs, photometry and images. Studies investigating these phenomena have provided valuable results, yet they are fundamentally limited by the size and resolution of the cosmic simulations that are available.

A data model which encapsulates implicit correlations between galaxy and halo properties over time would be an invaluable software to galactic astrophysicists. Identifying the halo and environmental parameters which constrain the evolution of different galaxy populations most profoundly will provide an interpretation into the physical processes common to these galaxies, and the extent to which their evolution can be attributed to their dark matter component and surrounding structures. By applying a predictive model to a high fidelity N-body simulation, one may produce a vast galaxy dataset with which to show these correlations explicitly, and by synthesis of observable quantities from these predictions, one can emulate equally large mocks which reflect this relationship, and by comparing this with some of the deepest, widest surveys to date, can be used to assess the validity of our models of galaxy evolution.

It is the premise of such a practical and versatile design which motivates this thesis. Machine learning offers a means to design a model which establishes connections between a set of halo properties to a set of galaxy properties, and to make predictions on larger scales than the original dataset. Cosmological simulations provide a useful galaxy-halo dataset with which to train a model, and N-body simulations can be used to test predictions and

create mocks. In this thesis, we describe the development and assess the predictions of an artificial neural network, designed to reproduce the evolutionary histories of galaxies in the Illustris: The Next Generation (TNG) hydrodynamical simulation suite.

1.3.2 Outline

Chapter 2 will discuss the neural network design and the necessary data acquisition and preparation steps before training and testing the model. This chapter will explore the advantages of utilising machine learning techniques to forecast galaxy formation from dark matter data exclusively, with a comprehensive analysis of the neural network algorithms utilised for these predictions, along with preprocessing and network design considerations.

Chapter 3 will focus on the results of the neural network when applied to the dark matter halos in the baryonic TNG simulations, while dark simulations will be addressed in later sections of the thesis. Similar to chapter 2, this will outline the halo and environmental properties used, and we will emphasise the recognised correlations between halo and galaxy data, establishing the differences between original and predicted statistics, to demonstrate the model's direct accomplishments and its limitations.

The primary focus of chapter 4 will be on the general approach for creating observable quantities, such as SEDs, using the predicted formation histories of galaxies. We will compare the spectral statistics of our sample with those generated from the true hydrodynamical simulation models to evaluate the effectiveness of the algorithm in producing realistic observables. The chapter will also explore modifications that could improve the algorithm's performance in this regard.

Chapter 5 will discuss the technical steps towards application of the completed algorithm to a pure dark matter simulation. The chapter will also present the test results of the algorithm's application to the dark TNG and Uchuu simulations. Additionally, we will discuss and compare the set of baryonic properties and observables with expected observational and theoretical statistics. Finally, the chapter will explore the anticipated outcomes of applying the algorithm to a high fidelity N-body simulation.

In conclusion, chapter 6 will discuss the successes and shortcomings of the neural network model, in the context of its suitability for predicting evolutionary galaxy properties, encompassing the physics of the galaxy-halo connection, populating N-body simulations

and complementing ongoing and future galaxy surveys. This will involve the potential uses of this methodology in the wide field of extragalactic astronomy, such as precise modelling of galactic star-forming regions and predicting the properties and abundance of massive galaxy clusters at high redshifts.

The data and methodology presented in this chapter is based on the methods presented in Chittenden & Tojeiro (2022).

2

Data, Design & Preprocessing

In this chapter, we introduce the methods of machine learning outline the design of the neural network model used to compute the galaxy-halo connection, justifying the choice of layout and hyperparameters. We introduce the cosmological simulations in the TNG and Uchuu suites, and describe the layout of historical halo and galaxy data in these simulations, as well as how this data was acquired, and secondary data was calculated, for use in the machine learning model. This also includes preprocessing methods, from the necessary numerical procedure of normalisation to physical corrections of data from different simulations.

2.1 Machine Learning

Machine learning is a subset of artificial intelligence that involves training algorithms to automatically learn from data and make predictions or decisions. It is a data-driven approach which enables computers to recognise patterns, learn from past experiences, and make predictions or decisions based on new data. Machine learning algorithms are designed

to improve their performance over time through continuous learning and adaptation.

Machine learning can be applied to statistical problems concerning the analysis of large and complex data sets. Traditional statistical methods (e.g. Bayesian modelling) often struggle to cope with large data sets which machine learning algorithms can easily handle (Aggarwal, 2018; Alpaydin, 2020). By identifying patterns and relationships between quantities in an input and output dataset, machine learning can uncover insights and make predictions which would be difficult or impossible to achieve using traditional statistical methods.

Machine learning takes on two specific forms: supervised and unsupervised machine learning. In supervised learning, samples in the dataset are assigned labels or values, and the model is trained to learn the mapping from one set of such data to another. Following each evaluation step, the model parameters are adjusted to improve the accuracy of predictions, usually based on some numerical metric comparing predictions and targets. The trained model can then be used to predict results using previously unseen input data, making the predictive model useful for applications such as language processing and data classification (Aggarwal, 2018). Examples of its practical utility in galaxy astronomy include the classification of star forming and quiescent galaxies according to black hole properties (Bluck et al., 2023) and the prediction of photometric quasar redshifts using SDSS quasar spectra (Hong et al., 2022).

Unsupervised learning involves training a model on an unlabeled dataset, in which there are no predefined target variables, and thus unlike supervised learning, training an unsupervised model relies on data patterns rather than predictive feedback. The model is designed to find patterns or structure in the data by grouping similar datapoints or identifying clusters of data, resulting in a descriptive model which identifies said patterns and structures. Descriptive models are therefore useful for applications involving data compression and anomaly detection (Alpaydin, 2020; Géron, 2020). An example of its use in galactic astronomy is the use of Gaussian mixture models (Viroli & McLachlan, 2017) to distinguish galaxy populations of different halo properties according to their star formation and metallicity histories (Fraser, Tojeiro, & Chittenden, 2022).

In this thesis, a predictive machine learning model has been developed to predict the evolution of galaxies from the historical properties of their halos and environment.

The development of a predictive model commonly relies on training and testing phases: routines where the model is optimised to fit known input and output data, and evaluated on unseen input data, respectively and consecutively. It is crucial to use training and testing sets to avoid overfitting: a phenomenon where the model fits the training data too closely and fails to generalise to unseen data (Aggarwal, 2018).

In a procedure known as parameter tuning, the components of a model undergoing training are modified in order to improve the fitting of the translation from input to output data, which is done recursively and automatically after evaluating the quality of the fit (Aggarwal, 2018). When the model is optimised as much as possible, the model is applied to testing data to assess the quality and accuracy of its predictions. One can evaluate the adequacy of the trained model by comparing the predicted results with a set of true data, and measuring the quality of fit using techniques like simple linear regression or χ^2 error calculation. Alternatively, one can compile the model several times and assess the convergence of the predictions, which may be achieved by calculating the mean and variance of multiple, independent outputs.

Artificial neural networks are a specific class of a predictive machine learning model, inspired by the composition and behaviour of the human brain. These models consist of multiple layers of nodes, or neurons, which are computational units connected to multiple nodes in the preceding and following layer by some scalar normalising function of the data, known as an activation function (Aggarwal, 2018; Géron, 2020). The translation from the input to output layers of the neural network are therefore equivalent to a series of linear operations, in which the coefficients of each operation, the weights and biases, are optimised in the training phase to provide the best translation from input to output data (Deisenroth et al., 2020). An activation function in the $(n+1)^{\text{th}}$ layer ϕ_{n+1} produces an output:

$$u^i_{n+1} = \phi_{n+1}(w^{ij}_n u^j_n + b^i_n) \tag{2.1}$$

where u^i_n is a vector of neuron activations, w^{ij}_n is a matrix of weight values for every connection between layers, and b^i_n is a vector of biases.

The training process of a neural network involves feeding the model with a set of

data including input quantities and the corresponding output values. In a process known as backpropagation, the weights and biases are subsequently adjusted to minimise the error between the network's predicted output values and the target output data. This error is usually characterised by a multidimensional "loss function" between the predicted and target data, and the adjustments are usually made according to gradient descent minimisation of this function (Alpaydin, 2020). A common choice of loss function is the mean squared error (MSE):

$$MSE\left(\{y_i\},\{f_i\}\right) = \frac{1}{N}\sum_{i+1}^{N}(y_i - f_i)^2 \tag{2.2}$$

where $\{y_i\}$ and $\{f_i\}$ are true and predicted datasets of size N. This loss function is used in all models discussed in this thesis.

The accuracy of the model's predictions depends on the quality of the training data and the complexity of the model. It is therefore necessary to adjust the training parameters, or hyperparameters, to advocate the most efficient and accurate training possible (Aggarwal, 2018; Géron, 2020). One such hyperparameter is the learning rate Γ: the size of incremental steps in the adjustment of weights and biases. This can make the network slow to converge if too small, and can fail to converge or even diverge if too large. Another is the number of hidden layers in the network, or the number of nodes per layer, which must be suitably large if the mapping from input to output data is highly nonlinear, but can fail to converge due to the many degrees of freedom in the model if this is too large.

The choice of activation and loss function is also important to specific datasets and problems; a poor choice can slow or even derail the convergence of the model and produce incorrect or misleading predictions (Géron, 2020). Eight examples of activation functions are given in fig. 2.1, showing the variety in the neuron outputs of different models. The sigmoid function, for example, is an appropriate choice for problems with binary classification, as the output of multiple sigmoid operations will converge towards 0 or 1. It is also a useful function to prevent the divergence of outputs, which will of course cause the network to fail. The sigmoid function is a poor choice for precise regression calculations due to gradient saturation: where the gradient becomes small enough that the weights and biases are barely updated in the training phase. The MSE is also a common choice

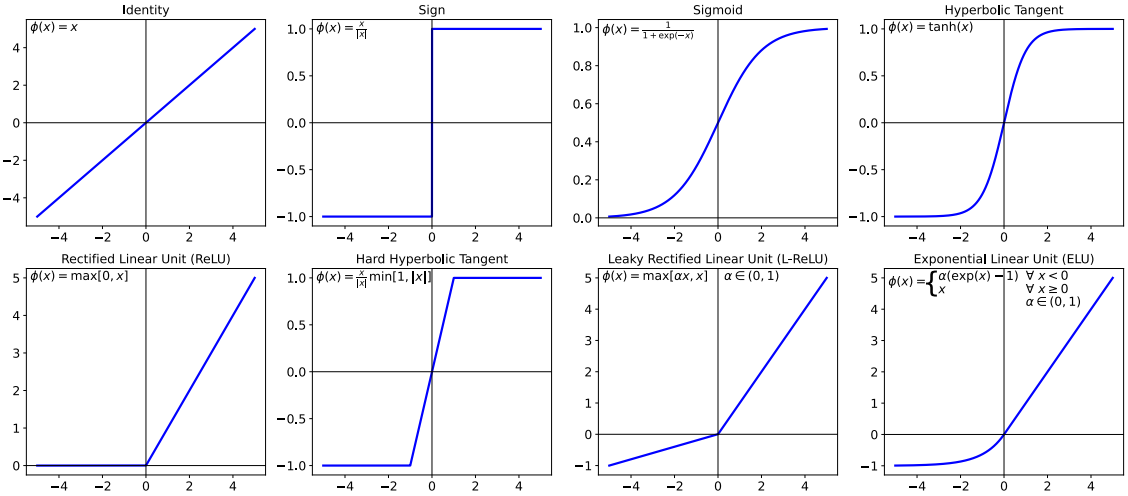


Figure 2.1: Eight commonplace activation functions used in neural networks (Aggarwal, 2018). In each scenario, the vertical axis displays the output of a particular node, while the horizontal axis displays the input value to that node. The name of the function is given in the figure titles, and their mathematical expressions are given in the figure.

of loss function for regressional problems as the error is a continuous variable, whereas binary cross-entropy is a more suitable choice for classification models, where the output is a discrete variable.

2.2 A Semi-Recurrent Neural Network

2.2.1 Architecture

In the neural network model which has been designed for this research with the TensorFlow Python library (Abadi et al., 2016), a fairly non-conventional design is employed. We have the specific problem of reproducing galaxy evolution from the detailed histories and multiple physical properties of their halos and environments. This requires an input of several hundred nodes which are causally connected with one another, in conjunction with time-independent variables. Training a standard neural network for this task would require a parameter space of extreme multiplicity to converge towards a highly specific, causally connected framework. The addition of hundreds of nodes per layer make training difficult due to the introduction of many more degrees of freedom.

Recurrent neural networks are a particular class of neural networks in which a one-directional activation sequence exists between successive nodes in a layer (Lipton et al., 2015). This allows information from previous nodes to be passed to those later in the sequence and for predictions to be made from the internal memory of inputs in any time

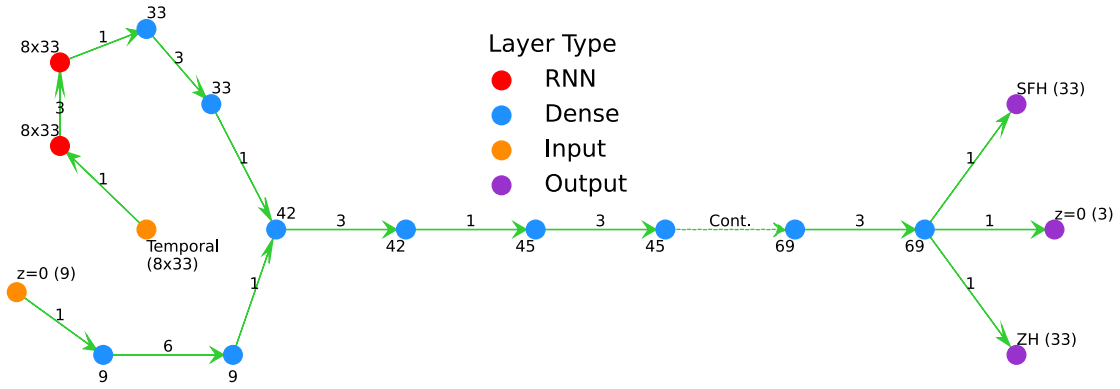


Figure 2.2: This diagram depicts the neural network architecture for central galaxies, with each dot representing a fully connected layer and its dimensionality indicated by a number, except for the purple dots that represent a subset of the final 1D output layer. The network has separate input layers for time-dependent and time-independent halo properties, which are combined at a dense layer with 42 nodes. The temporal input layer and recurrent layers are two-dimensional, consisting of eight variables over 33 time steps. The arrows show connections between consecutive layers, with the label indicating the number of times the connection repeats. For example, "3" means there are four consecutive hidden layers for that connection. The dashed line arrow indicates that every fourth hidden layer has three additional nodes until each layer reaches 69 nodes. Finally, the network outputs baryonic data, including star formation and metallicity histories, and three zero-redshift galaxy properties.

frame. These recurrent layers take in three-dimensional input datasets: adding a temporal axis to the sample and variable axes of a standard neural network input. The advantages of this design include the ability to recognise temporal dependencies between variables, and to predict sequential outputs from multiple sequential inputs.

This work invokes the design of a semi-recurrent neural network: a network with two input layers. The historical properties of the halos and environment are included in a recurrent input layer, where they share the same time steps. Variables with no time dependence are included in a second, dense layer. These inputs and the layers which follow it are concatenated into a single dense layer which eventually outputs the baryonic properties we aim to predict. The full architecture of this network is shown in fig. 2.2.

We implement temporal features of the neural network as 33-element vectors. As discussed in section 2.3, the simulation data used in this work is contained in multiple "snapshots" in time from a high redshift to a redshift of zero. The TNG data on which the network is trained consists of 100 snapshots. To reduce the complexity of the model and improve the speed of convergence, we evaluate temporal quantities in TNG for every third snapshot in the simulation, not including the first snapshot due to finite differencing between snapshots. The Uchuu simulation consists of 50 snapshots, and when calculating

Uchuu data we interpolate the relevant quantities over the reduced TNG time domain.

Two different networks are designed for predicting central and satellite galaxies, due to fundamental differences between their evolutionary histories and summary statistics (Pasquali et al., 2010; Bluck et al., 2020; Engler et al., 2020), and the inclusion of quantities which are only valid or relevant for one of the two datasets, such as the time of infall of satellite galaxies. The network designs are very similar, with the design of the network for central galaxies shown in fig. 2.2, while the network for satellite galaxies has seven temporal variables where the central network has eight, and the dense input layer has eleven variables instead of nine. In the satellite network, these sequences combine to make a 44-node dense layer rather than 42 nodes. From this point onwards, the two networks are identical, progressing from a 45-node dense layer to a 69-node output.

In the process of developing this model, the minimum number of hidden dense or recurrent layers needed to achieve convergence was used to determine the total number of layers in the network. Each input layer was succeeded by the optimal number of dense or recurrent layers to ensure the network recognised their equal importance before they were combined. The number of remaining hidden layers was also optimised to reach the minimum number required for consistent and accurate predictions, while gradually increasing in dimensionality to match the number of output nodes.

2.2.2 Activation Functions

As discussed in section 2.1, the choice of activation function is imperative to the performance of the neural network, and usually depends on the nature of the problem. Examples of commonly used activation functions and their mathematical definitions are given in fig. 2.1. In our case, the task of making a regressional fit to galaxy formation histories using static and temporal halo and environment properties has required a continuous, non-saturating activation in all network layers.

The highly skewed distribution of a number of the network's input features, such as mass accretion rates and overdensity histories, has made saturation effects a prevailing problem. When trialling activation functions such as the sigmoid or hyperbolic tangent, the network has performed extremely poorly due to a large range of data having extreme outputs and vanishing gradients with these activations.

The Rectified Linear Unit (ReLU) activation function, defined as follows:

$$\text{ReLU}(x) \equiv \max[0, x] \tag{2.3}$$

is a common choice of activation function to avoid gradient saturation. As the function has a constant derivative for positive inputs, backpropagation through ReLU layers does not allow the gradients to vanish, and can advocate faster gradient descent through the backpropagation of large gradients (Glorot et al., 2011). Where inputs are negative, ReLU output is zero-valued, which eliminates irrelevant connections in the network, effectively simplifying the model and improving its training efficiency and computational cost (Goodfellow et al., 2016).

An issue with ReLU activation is the Dying ReLU Problem (Lu et al., 2020). When the input to a ReLU neuron becomes negative, the derivative of the ReLU function falls to zero. The neuron effectively "dies" when the weights of the node are adjusted such that its input is strictly negative, and this zero-valued derivative causes any subsequent iterations from this node to be zero-valued, and thus the node no longer contributes to training the network. When a large number of ReLU neurons die, it can severely impact the capacity of the network to learn and lead to poor performance. This is the case with our model.

The similar Leaky Rectified Linear Unit (L-ReLU) activation function, defined as follows:

$$\text{L-ReLU}(x) \equiv \max[\alpha x, x]\,;\ \alpha \in (0, 1) \tag{2.4}$$

has a nonzero gradient for negative values and thus is a practical alternative to the ReLU activation function. While this does reduce the abundance of dead neurons in our model, we have found that using an L-ReLU activated network leads to arbitrary discontinuities in our predictions. Like ReLU, L-ReLU has an undefined gradient when the input equates to zero, which can result in different treatment of inputs approaching this value depending on the adjustment of weights. To mitigate each of these problems, we train the network with Exponential Linear Unit (ELU) activation:

$$\mathrm{ELU}(x) \equiv \begin{cases} x & \text{if } x \geq 0 \\ \alpha\left(\exp(x) - 1\right) & \text{if } x < 0 \end{cases} \tag{2.5}$$

with an α value of 1. This function and its gradient are continuous for all input values, is non-saturating and similar in form to ReLU, offering similar advantages such as linear behaviour and efficient gradient descent.

While the network performs well with ELU activation and converges more accurately and efficiently than with ReLU or L-ReLU, the model is not fully deterministic. ELU activated networks can be advantageous for normalizing a layer's output to have a mean close to zero and a standard deviation close to one. However, this normalisation behavior is only reliable for sequential network architectures with standardised inputs and initial kernel weights, otherwise the model is potentially unstable (Clevert et al., 2015; Géron, 2020). Particularly high or low gradients at very small or very large inputs may result in diverging or slow gradient optimisation, and thus have negative effects on the model's performance.

In many machine learning endeavours, a series of "dropout" layers may be implemented in the model, which randomly eliminate a specified fraction of neurons, in an effort to prevent overfitting. This reduction of connections is precisely the effect which was impeding the network's performance prior to implementing ELU activation. Even with ELU activation, training the network with dropout layers has worsened the quality of predictions, producing similar discontinuities and failed fits to the ReLU activated network. This illustrates the prevalence of vanishing gradients in this model and the importance of our choice of activation function.

2.2.3 Learning Rate

In the process of gradient descent, the learning rate is a model hyperparameter which decides the step size of each iteration. Specifically, for learning rate Γ and loss function f, the model parameters $\{\theta_j\}$ are updated according to the gradient of the loss function, like so:

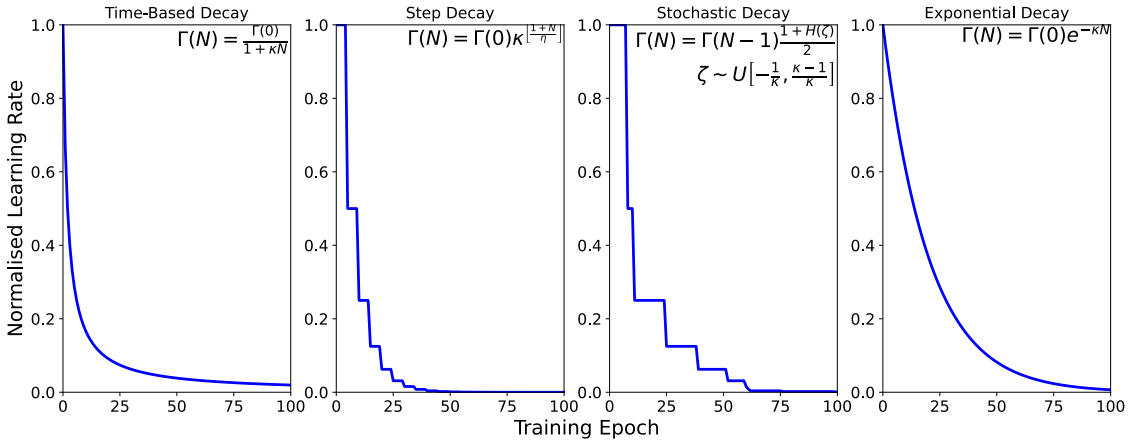


Figure 2.3: Four examples of adaptive learning rates used in neural networks, where the name of the adaptive learning rate is given in the title of each figure, and the mathematical formula for the learning rate Γ as a function of the epoch number N is given in the figure. In these mathematical expressions, κ and η are fixed, tailored hyperparameters, H is the Heaviside step function, and ζ is a random uniform variable. The stochastic formula is designed such that for every epoch in the training phase, there is a 1 in κ probability that the current learning rate will be halved.

$$\theta_j := \theta_j - \Gamma \frac{\partial}{\partial \theta_j} f(\{\theta_j\}) \tag{2.6}$$

The choice of the value of Γ is critical to the performance of the neural network. An exceptionally small value will cause the rate at which the model is updated to be slow, and so it may never achieve convergence. A value which is too large will cause the model to shoot past the optimal solution and potentially diverge from the desired result.

In some cases, a constant learning rate can be problematic; it may be too small at early times in the training phase to approach the optimal solution, but too large at later times to converge to this solution adequately. In circumstances where there is no single number which suits both of these issues, an adaptive learning rate is used, in which the learning rate is reduced with every epoch (Alpaydin, 2020). Four examples of adaptive learning rates are given in fig. 2.3, demonstrating the variety of different adaptive learning rates used for different models. This includes the step decaying learning rate, designed to descend by a certain factor after a certain number of epochs; the stochastic learning rate, which has a probability of randomly decaying by a given factor; and the time-based and exponential decay functions, giving a smooth, analytical expression for the learning rate.

In our model, we have found that an adaptive learning rate was necessary as there was no constant learning rate which lead to adequate convergence. We found that an exponentially decaying learning rate was the optimal choice for this model. Specifically, a

decaying learning rate of the form:

$$\Gamma = \Gamma_0 \exp\left[-\frac{N}{N_0}\right] \tag{2.7}$$

with values $\Gamma_0 = 8 \times 10^{-4}$ and $N_0 = 10$, was found to be the optimal solution for both central and satellite neural networks. This learning rate would eventually decay to a value so small that it would not make noticeable updates to the model, and therefore our training phase is given a total of 70 training epochs; terminating as soon as $\Gamma \leq 10^{-3}\ \Gamma_0$.

2.3 Simulation Data

2.3.1 Simulation Suites

IllustrisTNG

The IllustrisTNG simulations (Nelson et al., 2017, 2019a,b; Pillepich et al., 2017a, 2019; Springel et al., 2017; Marinacci et al., 2018; Naiman et al., 2018) are a suite of twenty cosmological simulations, assorted in simulation volume, mass resolution and number of mass particles in the simulation, with a pure dark matter counterpart for each hydrodynamical simulation. A summary of the properties of each of the TNG simulations is given in table 2.1, along with those of the Uchuu simulations (see below).

The TNG simulations begin at a redshift of 127, with an initial distribution of dark matter and baryonic gas that mirrors the conditions of the early universe. By propagation according to the moving-mesh magnetohydrodynamical code Arepo (Springel, 2010; Weinberger et al., 2020), the matter in the simulation clusters together and forms halos and galaxies, creating a model universe which aligns with the Planck Collaboration (2016) ΛCDM cosmological model, i.e. assuming the following cosmological parameters: $\Omega_\mathrm{m} = 0.3089$, $\Omega_\Lambda = 0.6911$, $\Omega_\mathrm{b} = 0.0486$, $n_\mathrm{s} = 0.9667$, $\sigma_8 = 0.8159$, $H_0 = 67.74$ km/s/Mpc. The simulation data includes catalogues of numerous dark matter, stellar, gas and black hole properties for all halos and galaxies in the simulation, which is visualised in fig. 2.4.

As discussed in section 1.2.1, the computational expense of hydrodynamical simulations usually results in a compromise between the size and the resolution of the simulation. In

IllustrisTNG Simulations							
Name of Simulation			Volume $(\text{Mpc})^3$	N_{dm}	M_{dm} $(10^6 M_\odot)$	N_{b}	M_{b} $(10^6 M_\odot)$
TNG50	-1		51.7^3	2160^3	0.45	2160^3	0.085
		-Dark			0.55	0	N/A
	-2			1080^3	3.6	1080^3	0.68
		-Dark			4.31	0	N/A
	-3			540^3	29	540^3	5.4
		-Dark			34.5	0	N/A
	-4			270^3	232	270^3	43.4
		-Dark			275	0	N/A
TNG100	-1		110.7^3	1820^3	7.5	1820^3	1.4
		-Dark			8.9	0	N/A
	-2			910^3	59.7	910^3	11.2
		-Dark			70.1	0	N/A
	-3			455^3	478	455^3	89.2
		-Dark			567	0	N/A
TNG300	-1		302.6^3	2500^3	47	2500^3	11
		-Dark			59	0	N/A
	-2			1250^3	470	1250^3	88
		-Dark			588	0	N/A
	-3			625^3	3760	625^3	703
		-Dark			4470	0	N/A

Uchuu Simulations					
Name of Simulation	Volume $(\text{Mpc})^3$	N_{dm}	M_{dm} $(10^6 M_\odot)$	N_{b}	M_{b} $(10^6 M_\odot)$
Uchuu	2952.5^3	12800^3	482.73	0	N/A
Mini - Uchuu	590.5^3	2560^3			
Micro - Uchuu	147.6^3	640^3			
Shin - Uchuu	206.7^3	6400^3	1.3242		

Table 2.1: A summary of the parameters of the twenty IllustrisTNG simulations (Nelson et al., 2019b) and four Uchuu simulations (Ishiyama et al., 2021). N_x represents the total number of particles or cells of component x, whereas M_x represents the size of one unit of mass in this simulation, i.e. the smallest resolvable mass. The components "dm" and "b" are dark matter and baryonic components, respectively. For each baryonic simulation, units of baryonic and dark mass are split according to the cosmic baryon fraction ($\Omega_{\text{b}}/\Omega_{\text{dm}}$), while in their dark equivalent simulations, these units are added together into one total mass unit. All simulations shown in this table use the parameters of the Planck Collaboration (2016) ΛCDM cosmological model.

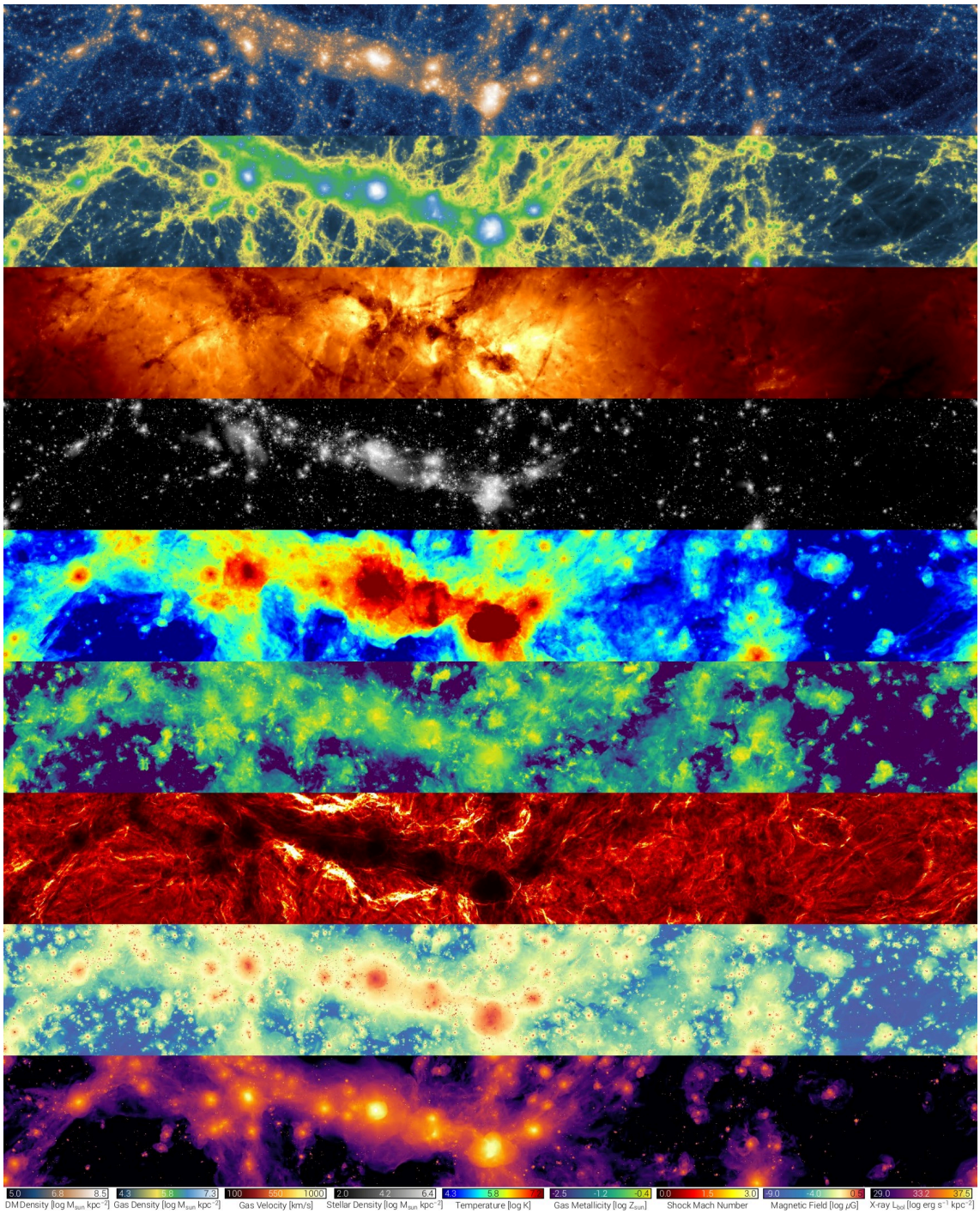


Figure 2.4: Nine halo and galaxy variables mapped over a $110 \times 14 \times 37$Mpc region of space in the TNG100-1 simulation at $z = 0$, illustrating the wealth of information directly available from the IllustrisTNG public data release. The variables and their physical scales are indicated in the legend at the base of the figure, and ordered according to their appearance from the top to the bottom of the figure. Image taken from Nelson et al. (2019b).

TNG, there exist three suites of the same cubic volume, whose approximate side length is indicated in the name of the suite. These are TNG50 (51.7^3Mpc^3), TNG100 (110.7^3Mpc^3) and TNG300 (302.6^3Mpc^3). As shown in table 2.1, the smaller volume TNG50 simulations have finer resolution than the larger TNG100 and TNG300 simulations. TNG50 data is therefore more suited to studies involving the resolved interiors of galaxies (Boecker et al., 2022), while the larger simulations are appropriate to study rare, high mass halos and structures (Hadzhiyska et al., 2021; Montenegro-Taborda et al., 2023).

TNG simulations of a specific volume are also run with different mass and spatial resolutions, i.e. smallest resolvable masses and volumes, initialised with the same random seed. This effectively generates the same galaxies and halos of the simulation with varying levels of resolution, which can be used to investigate the effect of this resolution on galaxy properties, such as the enhancement of star formation and feedback with improved resolution (Pillepich et al., 2017b). TNG simulations of different resolutions are denoted with a trailing index in the simulation name, e.g. "TNG100-2". Where this index is 1, the simulation is the highest resolution of all simulations of this volume, 2 indicates second highest, and so on. For TNG100 and TNG300, the n^{th} TNG300 simulation is intended to approximately recover the $n+1^{\text{th}}$ TNG100 simulation (Pillepich et al., 2017a,b); a relationship which we exploit in section 2.6.

Finally, for each of these hydrodynamical simulations, there exists a counterpart containing only dark matter. These dark simulations are denoted with "-Dark" on the end of the simulation name, e.g. "TNG50-2-Dark". Like the simulations of varying resolution, the halos in these simulations are generated from identical initial conditions, and halos in one simulation can be cross-matched with halos in another, allowing studies to compare the properties of halos with and without baryonic matter and baryon-driven phenomena. We make extensive use of this concordance between baryonic and dark simulations in chapter 5, comparing the predictions of our neural network model in baryonic and dark simulations. The lack of baryonic matter and the phenomena which influence the properties of halos make notable differences between baryonic and dark simulations (Castro et al., 2020; Anbajagane et al., 2021; Haggar et al., 2021; Riggs et al., 2022); yet despite this, the properties and statistics of halos are similar between baryonic and dark simulations.

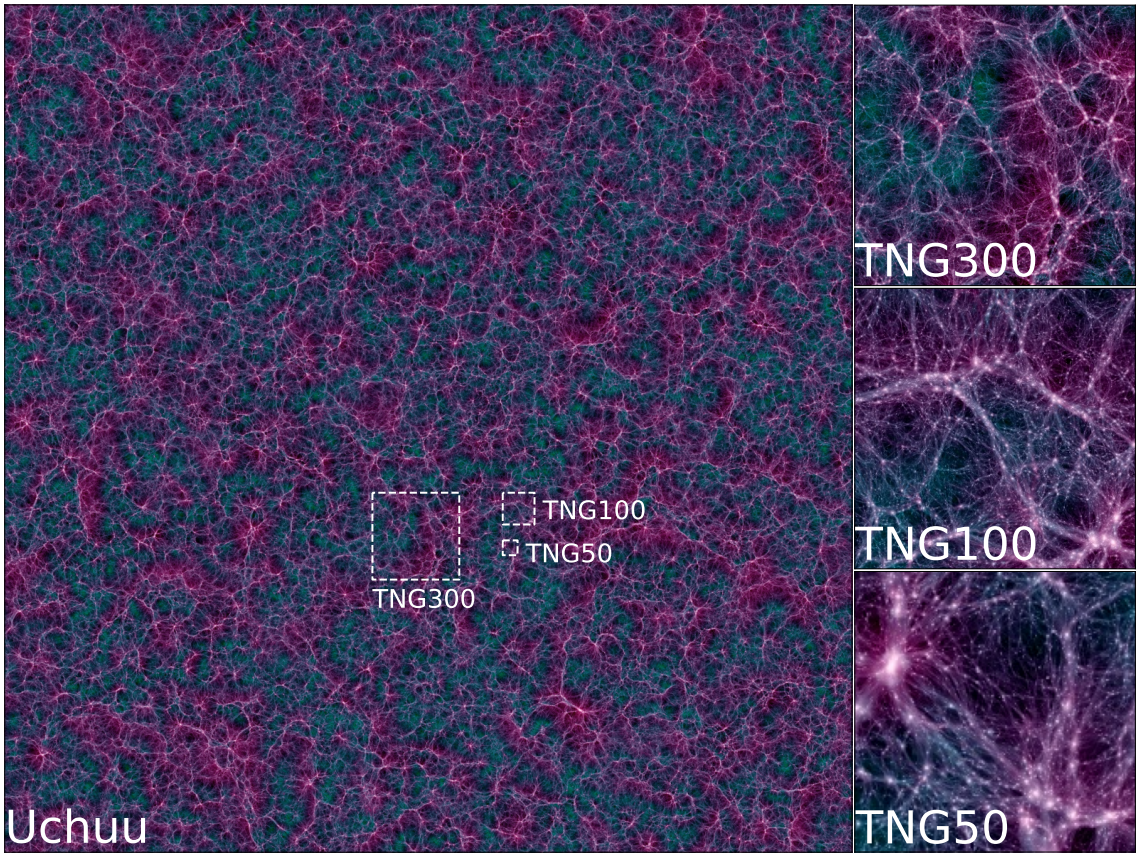


Figure 2.5: (Left) Density map of the Uchuu dark matter simulation at $z = 0$, with white dashed boxes showing the three TNG simulations to scale. (Right) Magnified regions of the Uchuu simulation enclosed by the boxes in the left figure, showing the details resolved by Uchuu in a volume equivalent to the TNG simulations.

Uchuu

Uchuu (Ishiyama et al., 2021) is a set of cosmological N-body simulations, designed to model the growth of dark matter halos from the scales of dwarf galaxies to massive clusters. This includes the main, gigaparsec scale simulation Uchuu, the smaller simulations of identical mass resolution Mini-Uchuu and Micro-Uchuu, and the small, high-resolution Shin-Uchuu simulation, which are used by the Uchuu project to model the growth of low mass halos and the effects of long-wavelength modes in the matter power spectrum. The simulation parameters of the four Uchuu simulations are summarised alongside the TNG properties in table 2.1.

Like TNG, Uchuu assumes the Planck Collaboration (2016) ΛCDM cosmological model, and is initialised from a redshift of 127. This means that the Uchuu and TNG share the same time domain, matter density and initial power spectrum, and are subject to the same dynamics of cosmic expansion, which makes comparing the growth of structures in

the two suites easier.

Unlike TNG, the Uchuu simulations are pure dark matter simulations, and are propagated by the GreeM N-body code (Ishiyama et al., 2009, 2012). Uchuu also defines halos according to the Rockstar halo finder algorithm (Behroozi et al., 2012a), while in TNG, two distinct structures: halos and subhalos (see section 2.3.2), are defined according to the Friends-of-Friends (FoF) and SubFind algorithms, respectively. The simulations are nonetheless alike in the structures that they form, and despite the superior resolution of most TNG simulations, the $\sim 25.7\text{Gpc}^3$ Uchuu simulation resolves structures on the scale of TNG very clearly. A comparison of the size of these simulations and the level of detail shown by the main Uchuu simulation on TNG scales is shown in fig. 2.5.

2.3.2 Data Access

For each of these simulation suites, the data is catalogued in multiple formats to suit different uses of the simulation results. Some studies will require detailed information of the properties of all particles or gas cells which constitute a galaxy or halo, while others will entail the contents of the halos which existed at earlier times, now forming part of the target halo or its interaction history. In both TNG and Uchuu, public data releases of the simulation results meet one of the following data formats.

Halo Catalogues

In the TNG and Uchuu data catalogues, "snapshots" refer to a set of data that represents the state of the simulated universe at a specific moment in time. These snapshots include a comprehensive list of all halos, and where applicable, galaxies which exist at this time. Properties of each halo, subhalo or galaxy can be requested from the snapshot archive.

In TNG, one can request the properties of halos or subhalos in a given snapshot. Halo, or group properties, are associated with halos formed by the FoF percolation algorithm (Huchra & Geller, 1982; Press & Davis, 1982; Davis et al., 1985), in which any given particle in the simulation is linked to another if they are separated by less than a specified linking length, which forms a network of directly or indirectly connected particles. This is a practical tool for defining halos as the optimal linking length results in a group which encloses the required density for virial collapse. A linking length equal to a fifth of the mean inter-particle separation of the simulation is commonly chosen, and is in fact chosen

in TNG, as it results in a mass function which is invariant under changes in cosmological parameters (Jenkins et al., 2001; White, 2001).

Subhalos in TNG are identified using the SubFind algorithm (Springel et al., 2001), which identifies locally overdense regions within a nonhomogeneous structure such as a halo, and defines the boundaries of the subhalo according to a contour of fixed density which traces any saddle point in the local density field. This is more practical for identifying structures which are not necessarily virialised; in fact many subhalos will be tidally distorted by their neighbours and by the main halo in which they reside. The Rockstar halo finder used in Uchuu adds to the FoF method by optimising for a phase space linking length within each group, which is used to define halo substructures (Behroozi et al., 2012a).

It is the main halos, the largest single objects in the simulation, which contain the intracluster medium (ICM) and gravitationally bound subhalos, while it is the subhalos which contain galaxies (Zavala & Frenk, 2019). By construction, the central subhalo, hosting the central galaxy of the group, is the most massive SubFind object within the FoF group. All other SubFind objects in the group are considered satellite subhalos, hosting satellite galaxies.

In TNG, quantities in the FoF halo catalogue which can be obtained include, but are not limited to, the sum of masses, star formation rates and gas and star metal fractions of all members of the group, the centre of mass and velocity of the group, the number of SubFind groups, and the total mass and comoving radius of regions enclosing spherical regions of different density. In Uchuu, of course the baryonic properties listed here are non-existent, but the halo catalogues include all of the above quantities where applicable, and in addition include angular momentum and NFW scale radius as fields.

Subhalo fields in TNG include the above quantities for individual galaxies, as well as masses, metallicities and photometric magnitudes of regions enclosed within different radii, such as that of the maximum of the rotation curve. Subhalo catalogues also include indices which point to their host halos. As halos and subhalos are defined using the same algorithm in Uchuu, members of the catalogue include the same fields, in addition to flags which indicate the ID of the host halo, where applicable.

Particle Data

The snapshots of the TNG and Uchuu simulations also include data relating to the individual particles of the simulation, concerning the status of each particle at the time of the snapshot. This particle data can be used to visualise processes internal to the halo or galaxy, and thereby understand the mechanisms of processes such as supermassive black hole growth.

The baryonic TNG simulations are initialised as a set of dark matter particles, and gas "cells" constructed from the Voronoi tesselation of the Euclidean field (Pillepich et al., 2017a). Gas cells will go on to form star and black hole components, thus there exist four types of mass unit in these simulations: dark matter, gas, stars and black holes. Each of these particle types has a unique set of quantities which can be acquired in the "particle" catalogue data.

Dark matter particles, being simplest in nature, have only coordinates, local mass density, velocity and gravitational potential as relevant fields. Gas cells contain an instantaneous cooling rate and star formation rate, as well as thermal energy, abundances of individual metals, and magnetic field strength and divergence. Star particles include similar properties as well as photometry and the time at which they were formed from gas cells, whereas black hole units include various estimates of mass accretion rate and AGN feedback energy.

Merger Trees

The evolution of halos and galaxies in these cosmic simulations is documented in the form of merger trees: a list of all progenitors and descendants of a given halo or subhalo. The basic structure of a merger tree is outlined in fig. 2.6.

For any given halo at a particular snapshot, the merger tree entails the properties of all objects from previous snapshots which would become part of the target halo, and the properties of the halo's descendants at later snapshots. This includes not only halos but the smooth accretion of unbound material. Each unique object forms a "node" of the merger tree. Connecting these objects results in a network of links, or "branches", which represent the growth of the halo and its progenitors in the time between consecutive snapshots.

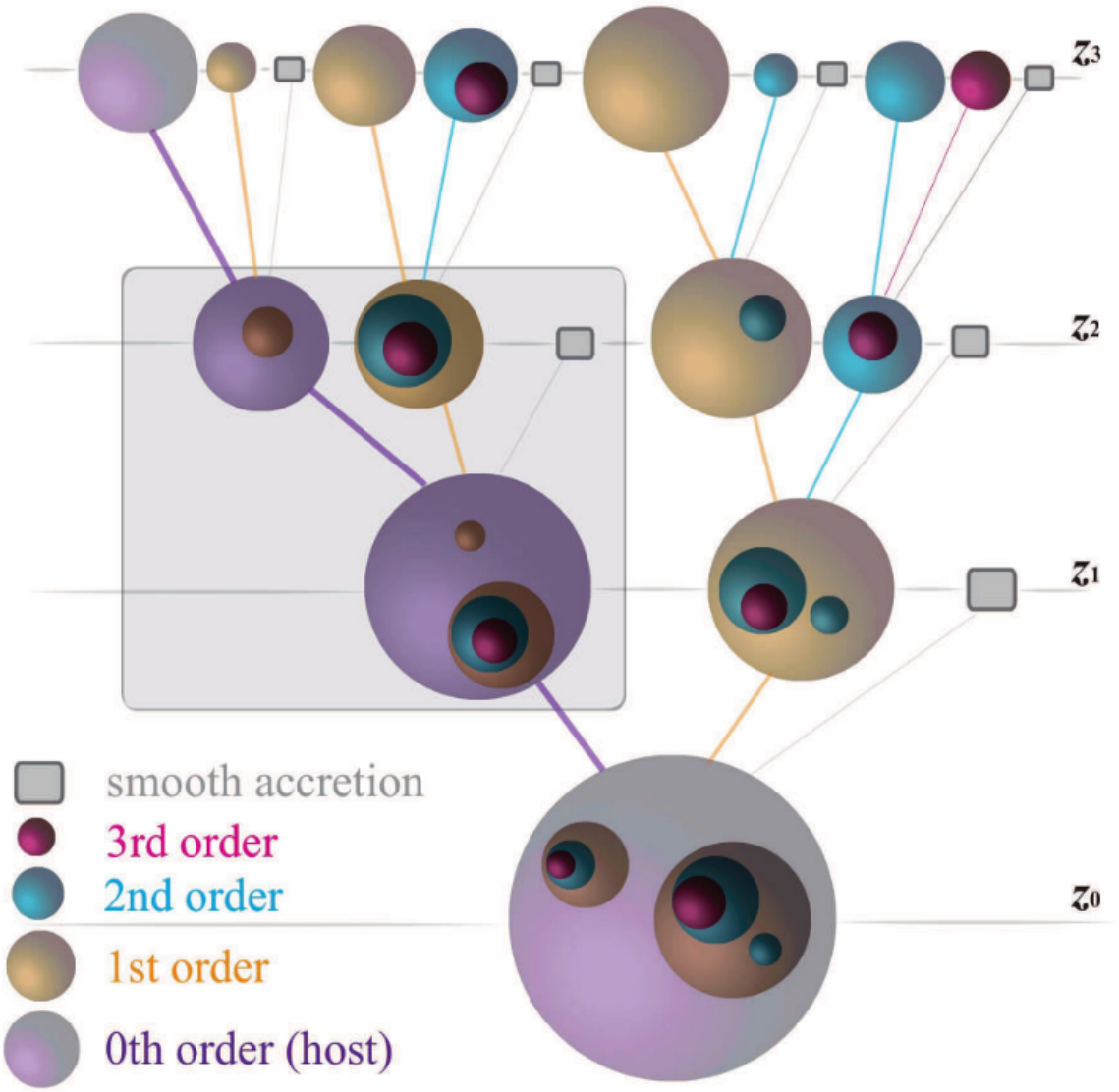


Figure 2.6: This diagram is a simplified representation of a halo merger tree, as seen in the TNG simulations, and is taken from Jiang & van den Bosch (2014). It is organised into rows, with each row representing a snapshot from the earliest time in the top row to the latest time in the bottom row. The size of each sphere corresponds to the mass of the halo it represents. In each row, the purple sphere represents the main (0^{th} order) progenitor of the host halo at the final snapshot, and the purple lines depict the main progenitor branch. Any other halos in the diagram are considered secondary progenitors, with the overlap of a smaller sphere over a larger one indicating that the smaller halo has been accreted by the larger halo as a bound subhalo. The small rectangles represent smooth accretion of matter that is not associated with halos. The boxed region provides an example of a subsection of the main tree which is also considered a subtree in its own right. For any such subtree, the highest order branch is the main progenitor branch of the subtree, while all other members of the ensemble are secondary progenitors; as is the case for the complete merger tree.

The main progenitor branch (MPB) is the single line of connections which trace the growth of the primary halo itself. By construction, this is the most massive progenitor of the target halo for each snapshot. The MPB is used to measure the evolution of the main halo, while secondary branches represent the accretion of smaller halos and other objects. A halo with more progenitors therefore has more secondary branches. These smaller halos may exist as satellites of the main halo for a period of time, in which case their own MPB traces their evolution through both their central and satellite phases. A halo is a central halo if its MPB is not a secondary branch of any other halo.

Significant events in the halo's history can sometimes be seen by eye in the features of the MPB. For example, a sharp rise in the mass of the main halo indicates that a major merger event took place within the time interval between the two snapshots. Alternatively, a satellite halo's mass may decline smoothly if it is stripped away by the ram pressure of the surrounding ICM. These effects can be verified by the presence of suitably large halos in their merger trees.

By requesting the MPB of any halo in TNG or Uchuu, one can acquire the properties of the chosen halo as a function of time, such as its mass components and gas or stellar metallicity. Requesting the full merger tree returns this information for all progenitors or descendants of the chosen halo, such as the gas masses of all accreted subhalos.

In Uchuu, due to the great size of the simulation dataset, the simulation is divided into two thousand merger "forests". Each forest is a self-consistent collection of merger trees: one forest contains all merger trees which have interacted with each other at some point in time. Consequently, each forest is an independent subset of the Uchuu simulation, which occupies its own region of space. One can therefore treat a single forest as an independent dataset, spanning all simulation snapshots.

Supplementary Catalogues

Aside from the data directly available from the public data releases of the simulations, studies have computed additional quantities using the simulation data, following the completion of the simulations themselves. This includes data which were computed by the original collaboration, and by third-party researchers.

Examples from TNG include subhalo matching between baryonic and dark simulations

	Network Data						
	Quantity	Notation	Units	Network	GQT	Logarithmic	Shuffle
Temporal Features	Halo Mass Accretion Rate	$\dot{M}_h$	$M_\odot$/Gyr	Both	Vector	False	1
	Subhalo Mass Accretion Rate	$\dot{m}_h$	$M_\odot$/Gyr	Satellite	Vector	False	1a
	1Mpc Overdensity	δ_1		Both	Scalar	False	2
	3Mpc Overdensity	δ_3		Central	Scalar	False	2
	5Mpc Overdensity	δ_5		Central	Scalar	False	2
	Circular Velocity (proxy)	$\tilde{v}_{\text{vir}}$	$\sqrt{(M_\odot/\text{Mpc})}$	Both	Vector	False	3
	Dark Matter Half-Mass Radius	$R_{\frac{1}{2}}$	Mpc	Both	Vector	False	3
	1Mpc Radial Skew	μ_3		Satellite	Vector	False	4
	3Mpc Radial Skew	μ_3		Central	Vector	False	4
	Distance To Closest Subhalo	d_{μ_3}	Mpc	Both	Vector	False	4
Non-Temporal Features	Specific Halo Mass Accretion Gradient	β (c)β_{halo} (s)	$\log \text{Gyr}^{-2}$	Both	None	False	1
	Specific Subhalo Mass Accretion Gradient	β_{sub}	$\log \text{Gyr}^{-2}$	Satellite	None	False	1a
	Scaled Infall Time	a_{infall}		Satellite	None	False	1a, 2, 4
	Scaled Formation Time	a_{max}		Satellite	None	False	1a
	Infall Mass Ratio	μ		Satellite	None	True	1, 1a
	Infall Velocity	v_{rel}	km/s	Satellite	None	True	2
	$z = 0$ Cosmic Web Distances	d_{CW}	kpc	Central	Scalar	True	2
	Starting Time	t_{start}	Gyr	Both	Scalar	False	All
	$z = 0$ Halo Mass	M_h	$M_\odot$	Both	Scalar	True	1
	Maximum Absolute Halo Accretion Rate	$\lvert \dot{M}_h \rvert$	$M_\odot$/Gyr	Both	Scalar	True	1
	$z = 0$ Subhalo Mass	m_h	$M_\odot$	Satellite	Scalar	True	1a
	Maximum Absolute Subhalo Accretion Rate	$\lvert \dot{m}_h \rvert$	$M_\odot$/Gyr	Satellite	Scalar	True	1a
Targets	Star Formation History	$\mathcal{S}$	$M_\odot$/Gyr	Both	Vector	False	N/A
	Metallicity History	$\mathcal{Z}$	$Z_\odot$	Both	Vector	False	N/A
	$z = 0$ Stellar Metallicity	Z	$Z_\odot$	Both	Scalar	True	N/A
	$z = 0$ Stellar Mass	M_s	$M_\odot$	Both	Scalar	True	N/A
	Mass Weighted Age	MWA	Gyr	Both	Scalar	False	N/A

Table 2.2: A summary of the quantities used in both neural networks, grouped by layer and ordered by their placement in said layer. This entails the units of each quantity, and indicates which networks utilise them and how they are normalised. The section column indicates which section of this paper discusses this quantity. The shuffle group (final column) indicates which variables are simultaneously scrambled when testing for feature importance (see section 3.4).

(Nelson et al., 2015; Rodriguez-Gomez et al., 2015), details of black hole mergers (Blecha et al., 2015; Kelley et al., 2016), distances to key locations in the cosmic web (Duckworth et al., 2019, 2020) and the properties of galactic bars (Rosas-Guevara et al., 2019; Zhao et al., 2020).

2.4 Data In This Work

In this work, we utilise data from each of the categories described in section 2.3.2. We make extensive use of merger trees in TNG and Uchuu, and use our own calculations for any data not included in the above data releases, e.g. using nearby subhalos to define quantities relating to cosmic environment. In this section we discuss the relevant variables used in the machine learning model and discuss how the data was acquired. A full description of the variables used in this network is given in table 2.2.

2.4.1 Dark Matter Quantities

Halo Mass Accretion History

The state of any galaxy will be dramatically shaped by the present mass of its halo, and will have been influenced in the past by the time and rate at which it acquired this mass, through smooth accretion or major mergers.

One of the key temporal variables included in this model is the halo mass accretion rate ($\dot{M}_h$). For this, we request the sum of all dark matter particle masses along the MPB in TNG, and between every third snapshot we apply finite differencing:

$$\dot{M}_h(t_j) = \frac{M_h(t_j) - M_h(t_{j-1})}{t_j - t_{j-1}} \tag{2.8}$$

where t_j is the cosmic time of the simulation at sample snapshot j.

By applying this finite differencing and passing $\dot{M}_h$ to the neural network, we train the network to recognise the effect of changes in the halo's mass, in the form of smooth accretion or mass loss, or shorter timescale events such as mergers, which are reflected in the environmental interaction history (see below). The recurrent framework allows the model to construct an integral of this variable over any time interval, corresponding to the net mass acquired in this time. The full integral of course corresponds to the zero-redshift halo mass, which is included as a static input parameter, being of great importance for scaling the predicted stellar mass. The maximum absolute value of $\dot{M}_h$ is also included as a static input parameter, which in relation to halo mass encompasses the magnitude of the largest interaction event.

A quantity which is derived from the mass accretion history is the specific mass accretion gradient of the halo, defined by Montero-Dorta et al. (2021) as the best-fit value of β in the following approximation:

$$\log_{10}\left(\frac{\dot{M}_h(t)}{M_h(t)}\right) \approx \gamma + \beta \log_{10}(t) \tag{2.9}$$

where γ is a constant, and β is a parameter that distinguishes the halos with the most rapid formation in the early universe, whose galaxies achieve peak star formation rate at

high redshifts, from galaxies forming at later times and having continued star formation at $z = 0$. Similar to halo mass, this parameter also plays a crucial role in categorising galaxies based on their evolutionary status. Montero-Dorta et al. (2021) find that β correlates strongly with stellar-halo mass ratio, quenching time, assembly bias and other galaxy properties. By incorporating β into the static input layer of the neural network, it promotes a metric of specific accretion that considers the halo's growth rate relative to its present mass and the duration it takes for the halo's mass fraction to evolve.

This accretion gradient also introduces a necessary quality cut to our training data. Even after we apply a lower stellar mass cut to all TNG data, there remain a handful of halo mass histories which are sensitive to the resolution limit of the simulation. Their noise-dominated behaviour has been detrimental to the network's performance. We find that the distribution of β values is very well fit by a Gaussian distribution. Poorly resolved mass accretion histories typically appear much flatter than most samples due to Poisson noise, and so have extreme values of β. We discard any samples whose β value exceeds a 5σ difference from the mean of the Gaussian fit to the β distribution.

The time of the formation of the halo ($t_{\rm start}$) is a parameter that is determined by the earliest snapshot in which the MPB of the merger tree is defined. The recurrent layer of the network requires identical time intervals for all data samples, whereas in the dataset, halos germinate at different times. To address this, we have interpolated the time-dependent properties over every third snapshot in TNG. Despite some samples having no data at the earliest times in the simulation, the recurrent layer maintains the causality between time steps. The starting time is used as an additional static parameter to identify the likely characteristics of galaxies whose host halo germinated at a specific time, while ensuring that the relevant time frame of evolution is recognised for each galaxy.

For central halos, we take the sum of masses of all dark matter particles bound to the FoF group as the halo mass in TNG. This field does not exist in the Uchuu merger trees, yet the field $M_{\rm 200c}$: the mass enclosed within a spherical region of density 200 times the critical density of the universe, is very closely matched to this field in TNG. $M_{\rm 200c}$ is slightly affected by baryonic data, but this bias is very small, and of course non-existent in Uchuu. Training and testing the model in the baryonic TNG simulations using either of these definitions of mass makes no noticeable difference to the behaviour of the neural

network.

For satellite subhalos, we are interested in the properties of the host halo as well as the satellite itself, particularly as the host ultimately controls the environment of the satellite phase. The satellite network includes the aforementioned quantities pertaining to the mass accretion history of both objects, whereas the central network includes only $\dot{M}_h$, as the formation history of the Subfind object is usually geometrically congruent to its FoF host, and comprises most of the mass of the system unless undergoing a major merger event.

For the host halo of a satellite subhalo, in TNG, we take the sum of dark matter particle masses of the FoF group as before, while we take the sum of dark matter particle masses of the SubFind group for the satellite subhalo. We denote the satellite subhalo mass as m_h to discern from the halo mass M_h. The presence of two β values in the satellite model, β_{halo} and β_{sub}, means that the Gaussian quality cut applies to both the satellite and the host. These variables are all included in the satellite neural network.

In Uchuu, all halos are defined by the Rockstar algorithm and there is no group/subgroup distinction; we simply use the upid flag to discern central from satellite objects. Maintaining the same central/satellite relationship in TNG, we select satellite halos in Uchuu as first order Rockstar halos and central halos as zeroth order in the merger tree structure (see fig. 2.6).

Halo Substructure

It has been shown that properties pertaining to the distribution of mass and the rotation curve of the halo have significant influence on galaxy evolution, likely indicating the timescale of internal collapse which leads to substructure growth, and accelerates star and black hole formation (Davies et al., 2019; Bluck et al., 2020; Lovell et al., 2021; McGibbon & Khochfar, 2022). Lovell et al. (2021) highlight the maximum circular velocity of the rotation curve (v_{max}) and the radius containing half of the total dark matter mass ($R_{\frac{1}{2}}$) as two properties which constrain galaxy properties at a given redshift, while McGibbon & Khochfar (2022) show that the inclusion of historical circular velocity and velocity dispersion make significant improvements to zero-redshift predictions.

There are two key issues with using these quantities: first, we find that the influence of baryons introduces substantial differences between the values of v_{max} in the baryonic and

dark TNG simulations, particularly for objects in the satellite phase and for high-redshift central halos. Second, the data used by McGibbon & Khochfar (2022) has been specifically restricted to halos with little difference in historical circular velocity and velocity dispersion. Both of these variables are sensitive to the effects of baryons, and desiring a model which can make equivalent predictions in a complete dark matter simulation invalidates the use of these variables in this work.

We compute a proxy for the virial circular velocity as a function of time in terms of dark matter half-mass radius and halo or subhalo mass, for central and satellite galaxies respectively:

$$\tilde{v}_{\rm vir}(t) = \sqrt{\frac{m_h(t)}{R_{\frac{1}{2}}(t)}} \tag{2.10}$$

which is similar to the proxy for NFW concentration used in TNG by Bose et al. (2019):

$$\tilde{v}_{\rm max} = \frac{v_{\rm max}}{H_0 r_{\rm max}} \tag{2.11}$$

In eq. (2.10), we have ignored constant terms such as the Newtonian Gravitational Constant. $\tilde{v}_{\rm vir}$ is not noticeably affected by the presence or abscence of baryons, and thus it is used in the temporal input of the network alongside $R_{\frac{1}{2}}$.

The same proxy is of course calculated for use in the Uchuu data. The radius enclosing half of the halo mass is included as a variable in Uchuu merger trees.

Local Overdensity

As discussed in section 1.2.2, the local environment is an important measure of the star formation, morphological and chemical properties of galaxies. Overdensity is a simple, common metric of the cosmic environment which modulates processes such as the rate of interactions of galaxies, the tidal distortion and ram pressure they experience, the rate at which they accrete gas from the circumgalactic medium (CGM), and other factors.

We calculate overdensities using the Grid Search In Python (GriSPy) package (Chalela et al., 2021): a tool for finding nearest neighbors in a regular grid of any number of dimen-

sions, which has been designed to handle simulations with periodic boundary conditions such as TNG and Uchuu and handle several spacetime metrics. The "bubble neighbors" search function retrieves a group of objects located within a defined distance from a given reference coordinate. In our case, we use this function to identify halos located within a set of comoving Euclidean distances of the center of mass of each target halo.

The local density of an object is calculated as the total mass of all subhalos located within a spherical volume centered on the object's center of mass (Agarwal et al., 2018; Bose et al., 2019). Only subhalos with centers of mass located within this volume are included in the calculation. As a result, the local dark matter density and overdensity of the object are dependent on the size of this volume. Specifically, the local halo density (ρ_r) is established by dividing the sum of masses of local subhalos (m_i^{local}) by the spherical volume which engulfs them (V_r):

$$\rho_r = \frac{\sum_i^{N_{\text{local}}} m_i^{\text{local}}}{V_r} = \frac{3\sum_i^{N_{\text{local}}} m_i^{\text{local}}}{4\pi r^3} \tag{2.12}$$

The mean dark matter density of the simulation volume ($\bar{\rho}$) is taken by dividing the sum of all halos in the simulation (m_i) by the simulation volume (V_{box}):

$$\bar{\rho} = \frac{\sum_i^{N_{\text{total}}} m_i}{V_{\text{box}}} = \frac{\sum_i^{N_{\text{total}}} m_i}{L_{\text{box}}^3} \tag{2.13}$$

The ratio between these respective densities defines the local halo overdensity (δ_r):

$$\delta_r = \frac{\rho_r}{\bar{\rho}} = \frac{V_{\text{box}}\sum_i^{N_{\text{local}}} m_i^{\text{local}}}{V_r\sum_i^{N_{\text{total}}} m_i} = \frac{3L_{\text{box}}^3\sum_i^{N_{\text{local}}} m_i^{\text{local}}}{4\pi r^3\sum_i^{N_{\text{total}}} m_i} \tag{2.14}$$

We use the notation δ_x to represent the overdensities calculated using a radius of x megaparsecs. For central subhalos, we calculate overdensities at radii of 1 Mpc, 3 Mpc, and 5 Mpc, as each of these will capture environmental structures on different scales. For satellite subhalos, we focus on smaller scale overdensities to measure the state of the halo environment. Agarwal et al. (2018) suggest that an overdensity radius of 200kpc is useful for constraining zero-redshift baryonic properties, such as stellar mass, metallicity and neutral and molecular hydrogen masses. However, our investigation of the history of multiple kiloparsec-scale overdensities showed that while their physical values

are inevitably shifted, they are geometrically congruent over time. Smaller overdensities are more likely to have high noise due to fewer subhalos being counted, thus we use the 1 Mpc overdensity (δ_1) as the solitary measure of overdensity for satellites.

Interaction History

We measure the history of interactions of galaxy-hosting subhalos by calculating a mass-weighted skewness of the radial distribution of surrounding subhalos, referred to as "skew" throughout this thesis. As with overdensities, these skews are calculated using periodic nearest-neighbour searches using the GriSPy package (Chalela et al., 2021).

The weighted statistical moments of a dataset x_j with weights w_j are given as follows:

$$\mu_1 = \frac{\sum_{j=1}^{N} w_j x_j}{\sum_{j=1}^{N} w_j} \tag{2.15}$$

$$\mu_2 = \frac{\sum_{j=1}^{N} w_j (x_j - \mu_1)^2}{\sum_{j=1}^{N} w_j} \tag{2.16}$$

$$\mu_n = \frac{\sum_{j=1}^{N} w_j \left(\frac{x_j - \mu_1}{\sqrt{\mu_2}}\right)^n}{\sum_{j=1}^{N} w_j}, \forall n \geq 3 \tag{2.17}$$

By definition, the skew is the third statistical moment, and will be denoted μ_3 hereafter. For the skew of the environmental mass distribution, the radial distances of subhalos from the centre of mass of the target subhalo assume the role of x_j in eq. (2.17) with $n = 3$, while the subhalo masses are the weights w_j.

Several attempts have been made to establish a temporal reference for important merger events using simple parameters like the redshift of the merger or the mass ratio of the merging objects at the time. However, since mergers occur on different timescales, assigning a definitive redshift is not feasible (Rodriguez-Gomez et al., 2015), and these parameters are susceptible to errors due to the complexities of the halo referencing algorithm in TNG (Poole et al., 2017). In practice, using a series of average merger ratios at successive timesteps does not improve the accuracy of constraining galaxy evolution.

The time-dependent skewness parameterisation provides a way to quantify the merger history of each halo, where the largest subhalos, being most massive, exert the strongest

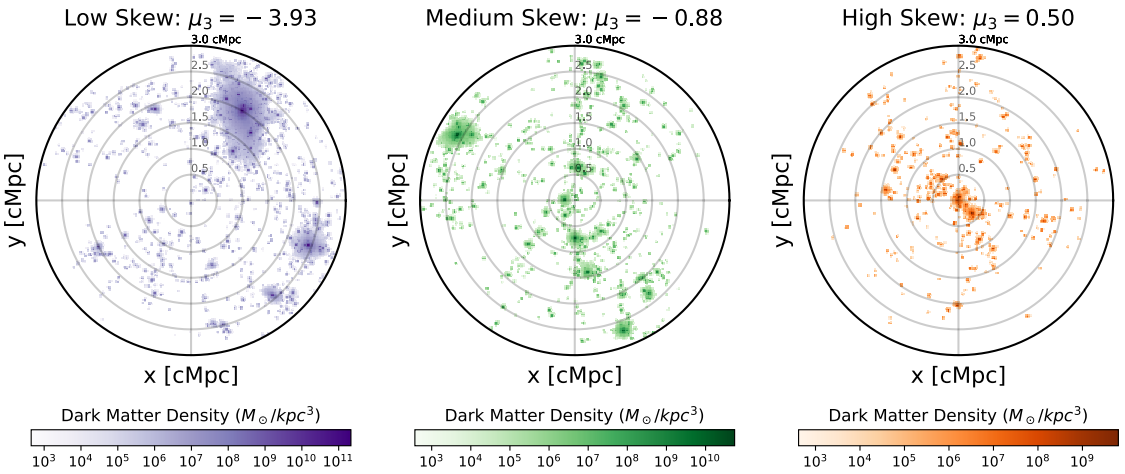


Figure 2.7: This figure depicts the x-y plane projection of the dark matter density distribution of subhalos surrounding three central subhalos of mass $\log_{10} M_h^{z=1}/M_\odot = 11.17$, taken from the $z = 1$ snapshot of the TNG100-1 simulation. The target subhalos are not visualised in these images, nor do they influence the skew calculation. Dark matter cells which are not gravitationally bound to the target subhalo and lie within a sphere of radius 3Mpc, centered on the target subhalo's center of mass, are selected for this image. The terms "low" and "high" skews are used to refer to the lower and upper quantiles of the skew dataset, respectively, while "medium" skews are close to the median skew.

influence on the matter distribution. During infall, the distribution becomes increasingly positively skewed, particularly if the infalling subhalo is massive in contrast with its neighbouring subhalos.

For satellites, in order to assess both the central phase mergers and collisions inside the FoF halo during the satellite phase, we calculate the skewness of the radial distribution up to a 1Mpc radius. For centrals, we evaluate the skewness up to a 3Mpc radius, encompassing the merger activity on both the halo and subhalo levels. This larger radius is necessary to incorporate the external data that affects the central galaxy's accretion activity, particularly for the largest FoF halos. However, satellite galaxies are more influenced by the mass distribution within the FoF halo, hence a smaller scale skew measurement is sufficient.

The typical surroundings of $z = 1$ subhalos with low, medium, and high skews are given in fig. 2.7, where the target subhalo is not included in the diagram to provide a clear view of the external mass distributions. A distribution with a medium skew has an impartial local matter distribution, as demonstrated in the central panel. A low skew distribution is characterised by one or several massive subhalos which move the center of mass far from that of the target subhalo. Highly skewed distributions, on the other hand, have the most massive subhalos concentrated near the target subhalo, increasing

the probability of a merger or significant tidal disruption. Hence the skew is a measure of the local concentration of dark matter at a specific time, while its temporal variation reflects the frequency and magnitude of flybys and collisions with the subhalo.

The neural network takes into account not only μ_3 but also the distance to the nearest subhalo as inputs, which stems from the skew calculation. This distance, denoted as d_{μ_3} and measured in megaparsecs, scales the distribution in a way that links the skew to the actual location of the merging halo. It also serves as a straightforward metric for measuring the proximity of the merging halo itself.

The Cosmic Web

The Discrete Persistent Structure Extractor (DisPerSE) algorithm (Sousbie, 2011) is a geometric method that identifies the stationary points of a density field and measures the structure of connections between critical points. We utilise the supplementary cosmic web catalog data created from the TNG simulations using the DisPerSE algorithm (Duckworth et al., 2019, 2020). More precisely, we extract the distances between the target halo and the nearest critical points and dark filaments at $z = 0$, which are collectively referred to as d_{CW}:

- $d_{\mathbf{node}}$ Distance to the nearest node (maximum) of the density field
- $d_{\mathbf{minima}}$ Distance to the nearest void (minimum) of the density field
- $d_{\mathbf{saddle1}}$ Distance to the nearest saddle point with one minimised dimension
- $d_{\mathbf{saddle2}}$ Distance to the nearest saddle point with two minimised dimensions
- $d_{\mathbf{skel}}$ Distance to the midpoint of the nearest filament

When acquiring cosmic web distances from dark simulations, we simply cross-match the samples in the baryonic TNG simulation to their dark equivalents. This is an adequate procedure as the geometric structure of the two simulations are equivalent, and so the cosmic web distances are preserved. For Uchuu, the DisPerSE algorithm is run on the simulation with a lower mass cut of structure tracers (halos) of $1.5 \times 10^{10} M_{\odot}$, such that the number density of tracers is similar to that of the TNG DisPerSE catalogue; a necessary correction due to the difference in resolution of these simulations.

While the cosmic web distances can provide valuable information for modeling the impact of large-scale environment and its anisotropic nature on central halos, they are not as crucial for objects in the satellite phase, and we have not included cosmic web quantities in the satellite neural network. According to Simpson et al. (2018), most of the satellite quenching results from ram pressure during infall or the host halo's tidal effects on its satellites. Although they suggest that the cosmic web could play a role in quenching some satellites, they indicate that this primarily affects low-mass satellites which intersect the gas inflow from the filament to the host.

Satellite Infall

For satellite galaxies, we wish to implement measures of the time of their acquisition by a larger halo, and the properties of the satellite subhalo in relation to its host.

It has been shown by Shi et al. (2020) that the scaled formation time of a satellite subhalo is a critical measure of their galaxies' star formation rate, gas fraction and other properties. The authors classify satellites as fast-accreting if this value is small, and slow-accreting otherwise; finding that fast-accreting satellite galaxies consequently have a different SHMR. This scaled formation time is defined as follows:

$$a_{\rm max} \equiv \frac{1 + z_{\rm half}}{1 + z_{\rm max}} \tag{2.18}$$

where the maximum redshift at which a subhalo's mass reaches its peak is denoted as $z_{\rm max}$, and $z_{\rm half}$ is the redshift at which half of this maximum mass is achieved for the first time, as per the simulation snapshot.

With a specific focus on the infall of the satellite, we compute a second quantity which we term the scaled infall time. This is defined similarly:

$$a_{\rm infall} \equiv \frac{1 + z_{\rm half}}{1 + z_{\rm infall}} \tag{2.19}$$

where we equate $z_{\rm infall}$, the redshift at which the subhalo becomes bound to a larger halo, to the highest redshift at which the subhalo loses central status, which is indicated by the GrNr flag in TNG and the upid flag in Uchuu.

$a_{\rm infall}$ represents the subhalo's continued growth, or demise, at the time of infall, or in the duration between capture and growth; whereas $a_{\rm max}$ is related to the growth profile during the central phase. For approximately one in 72 TNG samples, however, the subhalo continues to grow to half of its peak mass following capture, which is indicated by a value of $a_{\rm infall}$ less than unity. This is not a possibility for $a_{\rm max}$ as it consists of a ratio of two consecutive times, enforcing a lower bound of one on its value.

Taking into account the velocity of a satellite with respect to its host halo is crucial in determining the rate at which it loses mass due to ram pressure, as well as its orbital velocity, which is indicative of its position in the halo's gravitational potential and its potential for sustained star formation during the satellite phase (Behroozi et al., 2019a; Slone et al., 2021). By incorporating the satellite's velocity at the time of infall, we may obtain valuable insights into its trajectory and the interplay between its mass and environment in the future. We thereby compute the scalar velocity of the satellite subhalo relative to its host at the point of infall, by evaluating the difference between their peculiar velocity vectors at this snapshot:

$$v_{\rm rel}(z_{\rm infall}) = \|\mathbf{v}_{\rm rel}(z_{\rm infall})\| = \|\mathbf{v}^{\rm sub}_{\rm pec}(z_{\rm infall}) - \mathbf{v}^{\rm halo}_{\rm pec}(z_{\rm infall})\| \tag{2.20}$$

Additionally, we include the ratio of the subhalo mass to its host halo's mass at the time of infall, denoted μ:

$$\mu(z_{\rm infall}) = \frac{m_{\rm sat}(z_{\rm infall})}{M_{\rm parent}(z_{\rm infall})} \tag{2.21}$$

This establishes an important criterion for selection which pertains exclusively to the satellite model. The assumption is that the dark matter subhalo constitutes the majority of the mass in the galaxy-halo system. Nonetheless, in the baryonic TNG simulations, low mass subhalos exist where gas or stars constitute the primary mass, which is a feature of tidal dwarf galaxies (Haslbauer et al., 2019). If a satellite galaxy becomes much larger than its future central galaxy, the assumption that the dark matter subhalo is the dominant mass of the galaxy-halo system is no longer valid. To prevent this scenario, we use a criterion which reduces the maximum ratio of satellite to central mass by two orders of magnitude. The criterion is that only satellite galaxies with host halos for which

$m_s(z_{\text{infall}})/M_h(z_{\text{infall}}) < 0.1$ are included, where m_s is the stellar mass of the infalling galaxy and M_h is the halo mass of its host. By building our TNG-Dark sample only by cross-matching with the baryonic TNG sample, this automatically removes TNG-Dark objects which may have different histories from their hydrodynamical counterparts.

2.4.2 Baryonic Quantities

All baryonic quantities considered in this thesis are calculated on the subhalo scale, i.e. for SubFind galaxies. The methods of data acquisition, calculation and preprocessing are also identical for central and satellite galaxies, with the exception of resolution corrections discussed in section 2.6, where both the binning of masses in the calculation of the corrections, and the values of the corrections themselves, are different for central and satellite galaxies.

Star Formation History

The star formation rate and stellar metallicity are fields which are included in the merger trees in the TNG simulations. However, this thesis is in part motivated by the prospect of producing spectroscopic mocks which reflect the galaxy-halo connection, and we aim to evaluate the quality of predictions of the complete stellar population. Requiring historical properties which are compatible with spectral template modelling, we compute a pair of histograms of formation times for all stellar particles bound to the selected subhalo at redshift zero. Each of these is weighted according to stellar mass and mass-weighted metallicity, thereby creating a time distribution of these properties.

The star formation history (SFH) of each galaxy, denoted $\mathcal{S}(t)$, is computed as a histogram of the formation times of all stellar particles, weighted according to their mass. Each SFH per galaxy is initially computed with ten thousand time bins of small and identical width Δt. This simplifies the conversion to a star formation rate; dividing this by the time interval between age bins produces a set of rates which isn't distorted as it would be by time differencing over arbitrary bin sizes, as is the time interval between snapshots in the simulation. We then average these star formation rates over the time intervals between snapshots, matching the data to the time steps of TNG. Thus, an element $\mathcal{S}_n$ of the SFH $\mathcal{S}(t)$, containing N_{part} stellar particles in the n^{th} snapshot time interval, is defined as follows:

$$\mathcal{S}_n = \sum_{j=1}^{N_{\rm part}} \frac{M_{s\ j}^{\rm part}(t)}{\Delta t} \quad \forall t \in (t_{n-1},\ t_n) \tag{2.22}$$

Stellar Metallicity History

We apply the same weighted histogram method to compute the stellar metallicity history (ZH), with two differences. First, there is no conversion to a rate by dividing by the time interval of a single bin. Second, all metallicities are weighted according to the quantity of stellar mass formed in the specified time interval. An element $\mathcal{Z}_n$ of the ZH $\mathcal{Z}(t)$ is defined as follows:

$$\mathcal{Z}_n = \frac{\sum_{j=1}^{N} \mathcal{M}_j\ Z_j}{\sum_{j=1}^{N} \mathcal{M}_j} \quad \forall t \in (t_{n-1},\ t_n) \tag{2.23}$$

where $\mathcal{M}_j$ is the total stellar mass formed in each narrow time interval, and Z_j is the mass-weighted metallicity of all particles formed in said time interval.

Stellar Mass

The total mass of stars formed in the galaxy is defined as the integral of the star formation history over the full time domain of the simulation:

$$M_s^{z=0} \equiv \int_{\infty}^{0} \mathcal{S}(z)dz = \sum_{j=1}^{N_{\rm snap}} \mathcal{S}(t_j) \times (t_j - t_{j-1}) \tag{2.24}$$

where t_j are the cosmic times of sample snapshots. Unless made explicit, when referring to stellar mass this thesis will refer to the integral of the star formation history.

We assess the quality of predictions of star formation histories by computing a numerical SHMR. Despite the merger tree stellar mass being included as a $z = 0$ target in the model, we use the SHMR calculated explicitly from the true and predicted SFHs to assess the quality of the network's predictions[1]. Recovering not only the shape but the scatter of the SHMR, and the correlation of the SHMR with historical halo variables will show the neural network's capability of reproducing the dependence of star formation history on the evolution of its halo and environment. In this vein, we also compute the Mass-Weighted

[1]The SFH-derived mass is larger than the merger tree stellar mass due to the lack of recycling.

Age (MWA) of each galaxy:

$$\text{MWA} = \frac{\sum_{j=n}^{N} \mathcal{M}_n \; t_n^{\text{lookback}}}{\sum_{n=1}^{N} \mathcal{M}_n} \tag{2.25}$$

where t_n^{lookback} is the lookback time to snapshot n, and $\mathcal{M}_n$ is the stellar mass formed in snapshot n:

$$\mathcal{M}_n \equiv \int_{t_{n-1}}^{t_n} \mathcal{S}(t) \; dt = \mathcal{S}_n \times (t_n - t_{n-1}) \tag{2.26}$$

The MWA characterises the time at which most of the galaxy's mass is formed, as well as the basic geometry of the SFH. It is well established that galaxies residing in high mass halos have assembled most of their stellar mass at early times, therefore we measure this trend of halo mass with stellar MWA as another quality metric.

Stellar Metallicity

Similarly to stellar mass and the numerical SHMR, we evaluate the mass-weighted metallicity Z_s of each galaxy as follows:

$$Z_s = \frac{\sum_{n=1}^{N} \mathcal{M}_n \; \mathcal{Z}_n}{\sum_{n=1}^{N} \mathcal{M}_n} \tag{2.27}$$

and use these alongside the integrated stellar mass to obtain a mass-metallicity relation. Again, the shape and scatter of the Mass-Metallicity Relation (MZR) will inform the diversity and accuracy of the neural network's predictions of stellar metallicity histories.

2.5 Data Preprocessing

In this section we discuss the methods of preprocessing the data discussed in section 2.4 for use in the neural network models. It is common practice in machine learning to apply some form of normalisation: bringing all variables to a single scale such that they have equivalent weighting, and that the model remains stable to perturbations in the training phase. It is in fact necessary for the data in this work to be normalised due to vastly different quantities, however a simple scaling relation is not an adequate solution. Many

key variables such as halo mass and overdensity have very few samples with large values and are strongly over-represented at small values, which has introduced extreme biases when the data is simply rescaled. Temporal variables introduce an additional conundrum: the network may recognise the development of a variable over time as important, or may instead rely on values at specific points in time for predictions. We describe two important preprocessing methods for our data below.

2.5.1 Quantile Transformation

The quantile transformation method in the SciKit-Learn Python library (Pedregosa et al., 2011) is a transformation method which is used in most of our model data. Quantile transformation is a method in which the distribution of any given variable is transformed to a simpler distribution, such as a uniform distribution, which is more suitable for unbiased training data.

Quantile Transformations work as follows. Any variable x_i has a normalised probability distribution, which can be interpolated and integrated to establish a Cumulative Distribution Function (CDF) Ψ, which increases monotonically from 0 to 1. Ψ therefore maps the datapoint x_i onto its corresponding quantile value. Thus, if the CDF and the inverse function of a second probability distribution are analytical, it becomes easy to apply the same to map the data x_i onto the second distribution. The mapping of data x_i onto this distribution can be accomplished like so:

$$y_i = \Phi^{-1}\left(\Psi(x_i)\right) \tag{2.28}$$

where Φ is the CDF of the second distribution, i.e. the distribution of y_i.

If our goal is to have a normal distribution for y_i, then eq. (2.28) takes the following form:

$$y_i = \sqrt{2}\,\mathrm{erf}^{-1}\left(2\Psi(x_i) - 1\right) \tag{2.29}$$

and can be reversed when converting the data predicted by the neural network back to original, physical values:

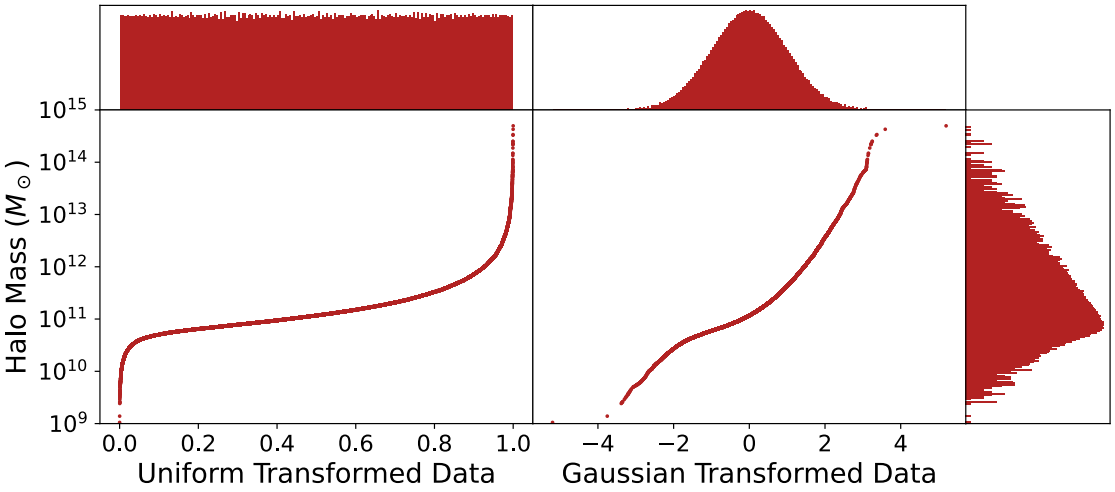


Figure 2.8: The translation of the halo masses from TNG100-1 dataset (shown on the vertical axis) into both a uniform distribution (left) and Gaussian distribution (right). The histograms along each axis, including the halo mass distribution on the vertical axis, are also presented. The graph illustrates that a considerable range of the data, particularly the halo masses above $10^{12.5} M_\odot$, corresponds to a very narrow range in the uniform distribution, which leads to high sensitivity towards slight variations in the transformed data. Consequently, the uniform distribution is not ideal for making predictions from this data. Therefore, we opted to transform the data to a Gaussian distribution.

$$x_i = \Psi^{-1}\left(\frac{1 + \mathrm{erf}\left(y_i/\sqrt{2}\right)}{2}\right) \tag{2.30}$$

We selected a Gaussian Quantile Transformation (GQT) for our data transformation because its domain is limited between 5 and -5, which makes it suitably normalised for the neural network. Although SciKit-Learn provides a transformation to a uniform distribution, the Gaussian transformation is more appropriate for our data. This is illustrated in fig. 2.8, which shows the quantile transformation of halo masses in TNG100-1 to uniform and Gaussian distributions. The distribution of data points in the uniform case is heavily skewed towards the distribution's edges. This narrow margin encompasses a wide range of values, making predictions based on a uniform distribution prone to a significant margin of error when estimating the true data. This is not the case with the Gaussian transformation, and this is therefore our choice of distribution of the normalised data.

2.5.2 Vector & Scalar Normalisation

The quantile transformation can take on two distinct forms in the case of time-dependent variables. The first form involves preserving the shape of the variable's history, as the behaviour of a dark matter quantity over time may impact the galaxy's present state. The second form acknowledges that there may be physical or systematic differences between

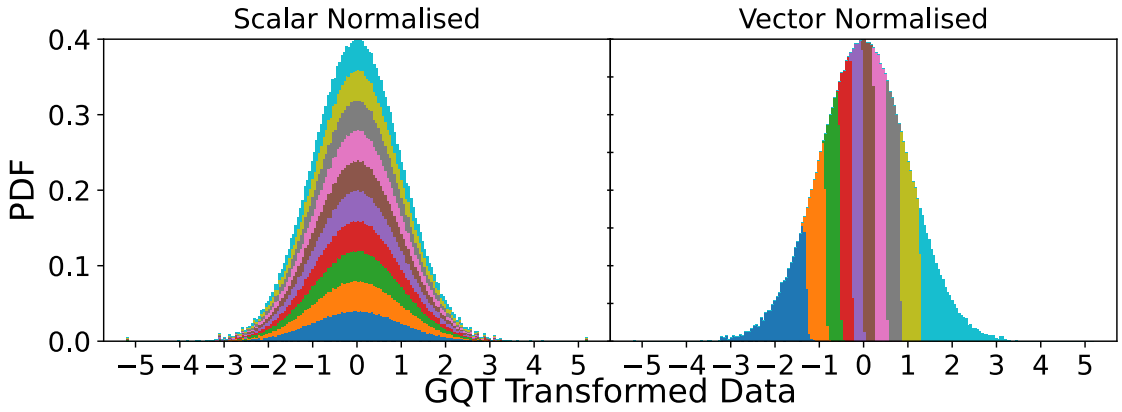


Figure 2.9: This graphic displays the distribution of monotonically increasing data after applying scalar (left) and vector (right) GQT normalisation. The data points at each time step are differentiated by various colours. Scalar normalisation transforms the data independently of other time steps, resulting in each time step sharing the same normal distribution. In contrast, vector normalisation transforms the data according to the complete range of the quantity's value over time, thus making each time step's distribution relative to another. When the full set of time step histograms is combined, it results in the Gaussian distribution of the complete dataset, irrespective of the method of normalisation used.

time steps, making the absolute value at a given point in time more significant than the gradient of the variable with time. To incorporate time-dependent properties differently, we introduce two normalisation techniques: vector and scalar normalisation, which are visualised in fig. 2.9.

The vector normalisation approach involves fitting the GQT to the time-dependent variable at all time steps simultaneously. This results in a transformed variable which is independent of the time at which it is defined, while still preserving the original variable's value. The transformed variable is obtained by applying the GQT to the 2D dataset[2] as follows:

$$y_i(t_j) = \sqrt{2}\,\mathrm{erf}^{-1}\left(2\Psi(x_i(t_j)) - 1\right) \tag{2.31}$$

where Ψ represents the CDF of the 1D set of values of the temporal variable x_i.

Variables which are transformed without considering time are referred to as scalar normalised. In this case, each time step has a distinct CDF for the variable, resulting in a set of unique transformed variables that are separated in time. The GQT is applied as follows:

[2]The dimensions of the variable dataset are data samples and timesteps.

$$y_i^k = \sqrt{2}\,\mathrm{erf}^{-1}\left(2\Psi_k(x_i^k) - 1\right) \tag{2.32}$$

The forms of GQT normalisation for every variable in the model are listed in the GQT column in table 2.2. Most time-dependent variables used in the neural network are vector normalised. In the case of halo quantities such as mass accretion and half-mass radius, the development of these quantities over specific intervals of time, as well as the shape of the quantity's history at early and late times, are all contributing factors to the evolution of their galaxies, which only vector normalisation preserves. The SFH and ZH of a galaxy are themselves important to consider as temporally evolving quantities, and are also vector normalised. The skew parameter is of particular importance as a temporal variable as it traces interactions with neighbouring subhalos, and therefore vector normalisation is strictly necessary to capture these interaction trajectories.

Overdensity histories are the sole temporal variable which are scalar normalised. Due to the expected substantial variation in the structure of the local environment, differences in overdensities between successive times may not have any significance and can be significantly large, rendering a common quantile transformation inadequate in distinguishing subsets of large and small values. Consequently, vector normalisation of the overdensity values can have a negative impact on prediction quality. Instead, the focus is placed on the instantaneous environment over its vectorised history, which is not affected by temporal changes in cosmic structure, and represents the local density field unique to each halo.

2.6 Resolution Corrections

As established in section 2.3.1, the resolution differences between the TNG100-1 and TNG300-1 simulations result in different calculations of star formation rates. Pillepich et al. (2017a,b) illustrate that the stochastic subgrid model of star formation in TNG ultimately depends upon the local density of star-forming gas, which is compromised at lower resolution. Thus, the amplitudes of the SHMR and MZR are weaker in the lower resolution simulation. In this section, we show how adjustments are made to the TNG300-1 data such that it can be used alongside the TNG100-1 data in the neural network, without any bias between the two simulation datasets.

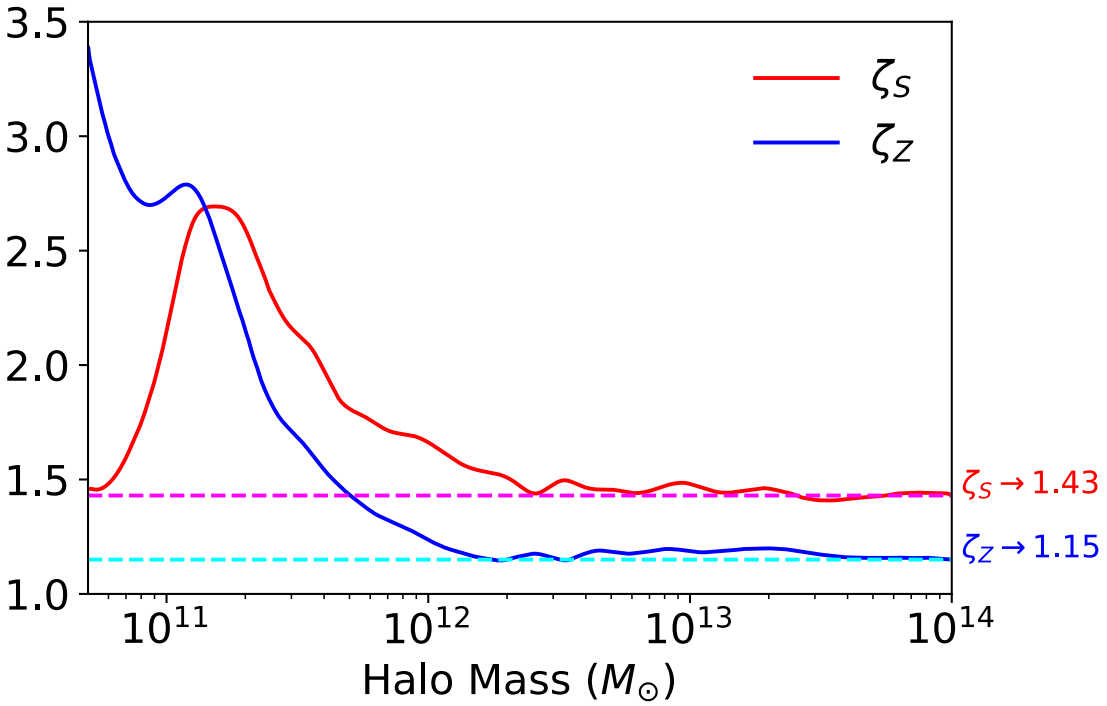


Figure 2.10: This figure displays depicts how the ζ corrections vary with halo mass at a redshift of zero, at all halo masses where we sample TNG300-1. For masses exceeding the range displayed, i.e. above $10^{14} M_\odot$, the zeta function values are calculated as the average over the $[10^{13} M_\odot,\ 10^{14} M_\odot]$ interval.

2.6.1 Scaling Mass And Metallicity At Fixed Redshift

An appendix in Pillepich et al. (2017a) shows that a correction can be made to the SHMR of TNG300-1 by means of a multiplicative function of halo mass. As the resolution of TNG300-1 is equivalent to the resolution of TNG100-2, the SHMR of TNG300-1 can be adjusted to match that of TNG100-1 by taking the ratio of the mean SHMRs of the two TNG100 simulations. We label this correction ζ, and use it to adjust the halo mass-metallicity relation (HMZR) of TNG300-1 as well as the SHMR.

At a given redshift, the ζ fractions are defined:

$$\zeta_S(M_h \mid z) = \begin{cases} \bar{M}^*_{\text{100-1}}(M_h)/\bar{M}^*_{\text{100-2}}(M_h) & \text{if } M_h < 10^{14} M_\odot \\ \mathrm{E}\left[\bar{M}^*_{\text{100-1}}(M_h)/\bar{M}^*_{\text{100-2}}(M_h)\right] \quad \forall \frac{M_h}{M_\odot} \in [10^{13},\ 10^{14}] & \text{if } M_h \geq 10^{14} M_\odot \end{cases} \tag{2.33}$$

$$\zeta_Z(M_h \mid z) = \begin{cases} \bar{Z}^*_{\text{100-1}}(M_h)/\bar{Z}^*_{\text{100-2}}(M_h) & \text{if } M_h < 10^{14} M_\odot \\ \mathrm{E}\left[\bar{Z}^*_{\text{100-1}}(M_h)/\bar{Z}^*_{\text{100-2}}(M_h)\right] \quad \forall \frac{M_h}{M_\odot} \in [10^{13},\ 10^{14}] & \text{if } M_h \geq 10^{14} M_\odot \end{cases} \tag{2.34}$$

where $\bar{M}^*_{\text{100-1}}$ is the mean stellar mass as a function of halo mass for TNG100-1, which we evaluate by binning the SHMR and interpolating through the mean values of each bin. The same method applies to stellar metallicity and to TNG100-2.

As in Pillepich et al. (2017a), these functions are assigned their average value over the halo mass range $10^{13} M_\odot \leq M_h \leq 10^{14} M_\odot$ where $M_h \geq 10^{14} M_\odot$, due to the low sample size at such a high mass. The shapes of ζ_S and ζ_Z at other halo masses at $z = 0$ are shown in fig. 2.10. We show in fig. 2.11 that applying this correction to the merger tree masses and metallicities results in a suitable match between the SHMR and HMZR of the original TNG100 and corrected TNG300 data.

2.6.2 Scaling SFH And ZH At Fixed Halo Mass

The ζ corrections are effective in modifying the SHMR and HMZR for a single snapshot. However, since the computation of the galaxy formation histories of the network relies on stellar age spectra, which are not determined using single-snapshot data, ζ ceases to be a suitable correction for this purpose. Hence we implement a similar correction for the average histories of objects in bins of zero-redshift halo mass. We compute a time-varying variable which we label ψ, and define as follows:

$$\psi_S(z \mid M_h^{z=0}) = \tilde{\mathcal{S}}_{\text{100-1}}(z) \ / \ \tilde{\mathcal{S}}_{\text{100-2}}(z) \tag{2.35}$$

$$\psi_Z(z \mid M_h^{z=0}) = \tilde{\mathcal{Z}}_{\text{100-1}}(z) \ / \ \tilde{\mathcal{Z}}_{\text{100-2}}(z) \tag{2.36}$$

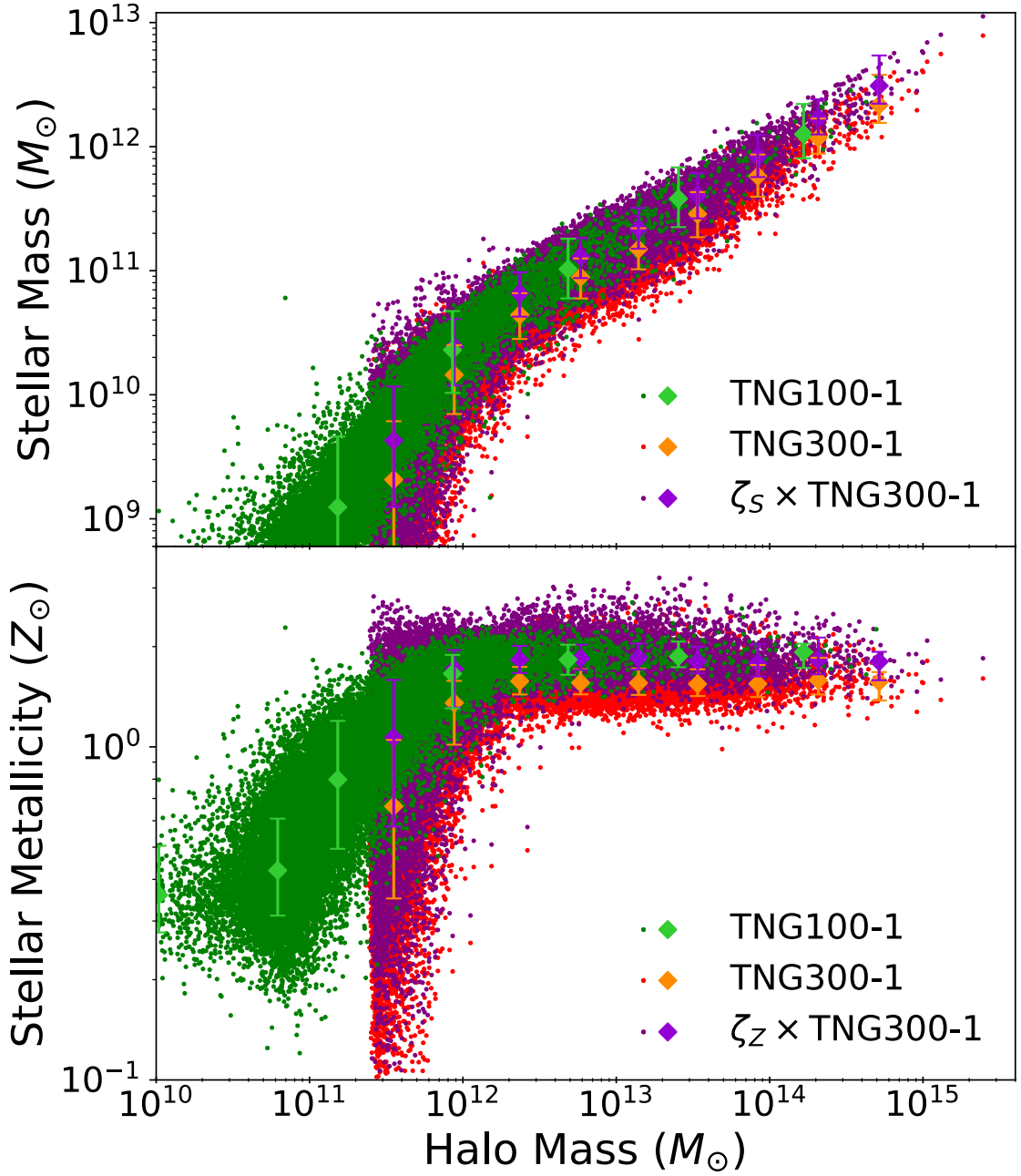


Figure 2.11: The relationship between the SHMR and the HMZR at a redshift of zero is presented in this figure. The TNG100-1 data (green) is compared with the original (red) and adjusted (purple) TNG300-1 distributions, which have been adjusted using the zeta functions. The error bars with matching colours correspond to the median and the range between the 15$^{\text{th}}$ and 85$^{\text{th}}$ percentiles of either the stellar mass or metallicity within a particular halo mass bin for each dataset.

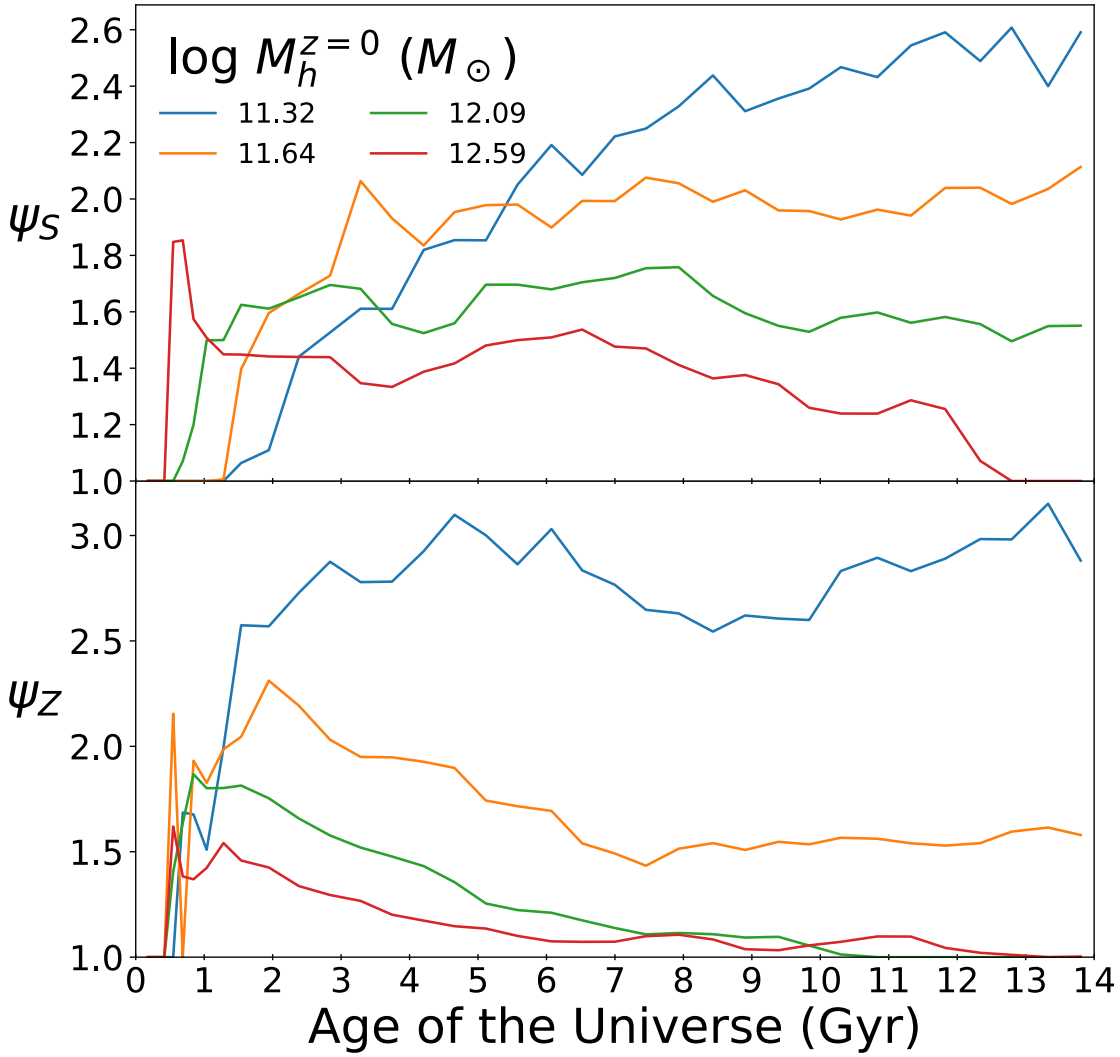


Figure 2.12: The figure displays the ψ variables with respect to cosmic time, with four representative halo mass bins shown as examples. The upper panel shows the resolution correction for the star formation history, ψ_S, while the lower panel shows the correction for the metallicity history, ψ_Z.

where by applying a cubic spline interpolation between ψ values per halo mass bin, we can extrapolate values of ψ outside of the mass range of TNG100-1, where necessary.

The star formation and metallicity properties of ψ as functions of both time and halo mass are illustrated in fig. 2.12. Like the ζ corrections, this results in a reasonable agreement between the two simulations after the TNG300 data is modified, which we show in fig. 2.13. Thus, we use the ψ-corrected star formation and metallicity histories when training our neural networks.

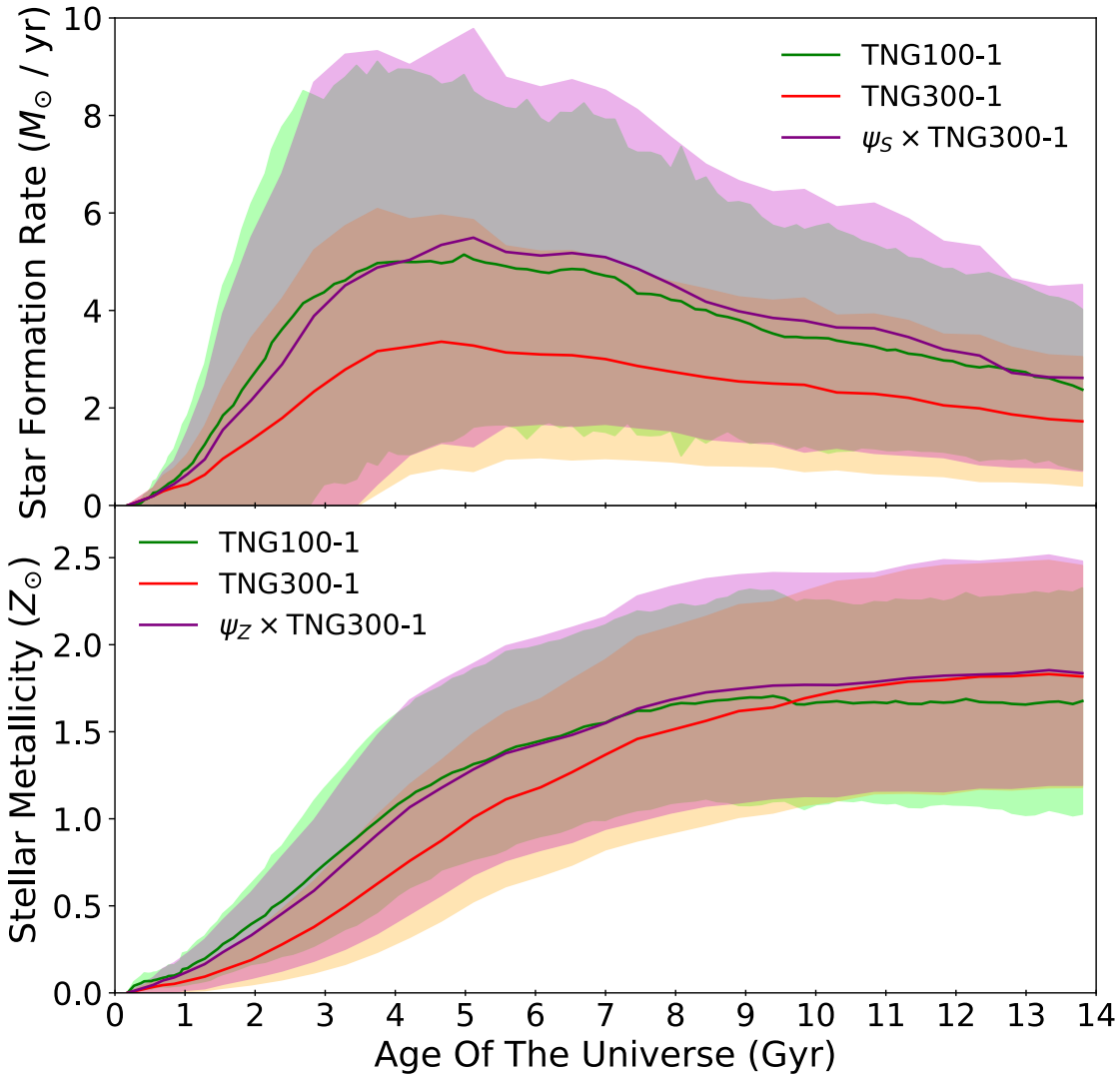


Figure 2.13: The figure illustrates the average star formation and metallicity histories of central galaxies with halo masses ranging from 10^{12} to $10^{12.2}$ solar masses, where the size of the shaded regions represents the standard deviation of these data as a function of time. The TNG100-1 data (green) is aligned with the modified TNG300-1 curves (purple), which have been adjusted from the original TNG300 data (red) using the ψ parameters. To eliminate erroneous characteristics caused by a small sample size at early times, time steps that contain fewer than 100 nonzero values across all data are excluded from the figure. This only applies to cosmic times earlier than 1Gyr.

2.7 Summary

In this chapter, an artificial neural network incorporating both time-dependent and time-independent variables associated with the histories of dark matter halos and their environment was developed using data from the TNG simulations. Its purpose is to establish connections between these variables and the star formation and metallicity histories of the galaxies nested in halos, enabling predictions of said galaxy properties. In addition to the calculation of novel variables from the simulation data, this has required a normalisation method which accounts for the vast differences in the distributions of variables, and a resolution correction to ensure the congruence of galaxy-halo statistics between the TNG100 and TNG300 simulations. These procedures can be summarised as follows:

1. We extract the mass and half-mass radius from the main progenitor branch of simulation merger trees, calculate environmental histories using a periodic-boundary neighbour search algorithm at each snapshot, and calculate star formation and metallicity histories from the stellar age spectrum of all star particles bound to the subhalo. Additional properties such as specific mass accretion gradient and a proxy for virial velocity have been calculated from these data. These quantities have been compared with those in TNG-Dark to ensure that none are substantially biased by the lack of baryons in dark simulations, making the model unsuitable for application to dark matter simulations. Uchuu is chosen as a gigaparsec-scale N-body simulation on which to test the model, due to the similarity of its halo properties and identical cosmological parameters to TNG.

2. The neural network is designed with a semi-recurrent architecture, containing both a recurrent input for temporal quantities and a dense input for static variables. The choice of ELU activation of the network is optimal for avoiding gradient saturation while simultaneously avoiding dying neurons and gradient discontinuity. We also require an exponentially decreasing learning rate to achieve swift and accurate model convergence, which allows the training phase to be terminated when the learning rate becomes negligible in updating the model parameters.

3. Due to the diversity and sparsity of most input variables, we normalise these quantities by applying a quantile transformation to a Gaussian distribution, which is a

suitable choice of distribution as it does not diverge under small differences in the original data. For temporal quantities, we apply two methods of this transformation: scalar and vector normalisation, allowing temporal quantities to be implemented as a set of unique variables or a single entity which can be integrated or differentiated over time, respectively. Overdensity histories are scalar normalised due to significant changes in their value over time, whereas other temporal quantities are vector normalised to implement their temporal geometry into the model.

4. Due to the differences in resolution between the simultaneously used TNG100-1 and TNG300-1 simulations, the coarser TNG300-1 has lower calculated star formation rates, and therefore, summary statistics such as the SHMR are offset. To obtain a constant relation in our data, we apply a multiplicative correction to the stellar masses, metallicities and galaxy formation histories in TNG300-1, inspired by a method to correct the SHMR introduced in Pillepich et al. (2017a). This method has fruitfully produced consistent formation histories and galaxy-halo relations between the original TNG100-1 data and modified TNG300-1 data.

In the following chapter, we evaluate the accuracy and physical properties of the galaxy formation histories produced by the neural network, comparing the relations between synthetic galaxy properties and halo properties with those in the TNG simulations. Additionally, we present a method of modification to these predictions to correct for a discrepancy between the original and predicted datasets, and investigate the input parameters with the strongest influence on galaxy-halo statistics. This is followed by the comparison of synthetic observables in chapter 4 and the application of the model to N-body simulations in chapter 5.

This chapter is predominantly based on the results of Chittenden & Tojeiro (2022), with the exception of section 3.3, which describes the methods and physical results of Behera, Chittenden, & Tojeiro (in prep.).

3

Neural Network Predictions

3.1 Introduction

In the previous chapter, the design of a semi-recurrent neural network, constructed to simultaneously implement time-dependent and time-independent variables relating to the histories of dark matter halos and their environment in the TNG simulations. The network is designed to relate these quantities with the star formation and metallicity histories of the galaxies bound to these halos, such that these galaxy properties can be predicted accordingly. Preparing this data has involved the acquisition of SubLink merger tree data, and required the calculation of secondary quantities (e.g. specific mass accretion gradient); as well as the scaling of galaxy quantities according to simulation resolution, and two distinct methods of quantile transformation of temporal variables.

In this chapter, we present the galaxy formation histories predicted by the neural network when applied to the TNG simulation data, in contrast with the hydrodynamical data with which the network was originally trained. Utilising the original and predicted

datasets, we compare properties of the directly predicted star formation and metallicity histories, and relate these to the derived summary statistics, such as the stellar-halo mass relation, which indicate the aspects of the galaxy-halo connection which are recognised by the network.

We show that the model is capable of predicting the stellar mass, metallicity and SFH and ZH geometry of galaxies in different evolutionary regimes to discernible accuracy, recovering comparable correlations between galaxy and halo properties as seen in the simulation data; and show that the semi-recurrent design of the model prevails over a conventional neural network in making these predictions. Despite this, we present a loss of information regarding SFH and ZH features acting on short timescales. We present a method of modifying the neural network predictions based on similarly predicted Fourier transforms, with which some degree of accuracy of the derived summary statistics is recovered.

Finally, we assess the relationships between input and output variables learned by the neural network, by performing a test similar to a permutation importance test for feature importance. By replacing groups of input variables with random data, we recalculate key relations between galaxy and halo properties with the signals of mass accretion history, cosmic environment, etc. removed, and compare these with the fiducial predictions. Similarly, we randomise individual members of these groups to demonstrate the utility of these variables to the network. We find that certain variables are important to predicting the scatter of these galaxy-halo relations, yet the relations themselves are impervious to randomisation, implying that the relations can be inferred from multiple groups.

3.2 Predicted Galaxy Properties

3.2.1 Evolutionary History

A typical prediction of the neural network is given in fig. 3.1, simultaneously showing the original and predicted star formation and metallicity histories of a single satellite galaxy. The predicted results show a similar amplitude and shape in both the SFH and ZH of this galaxy, and for most central and satellite galaxies in the TNG data. However, while the smooth trends of star formation and chemical enrichment with time are well-matched, the network fails to predict the variability of these evolutionary histories on short timescales.

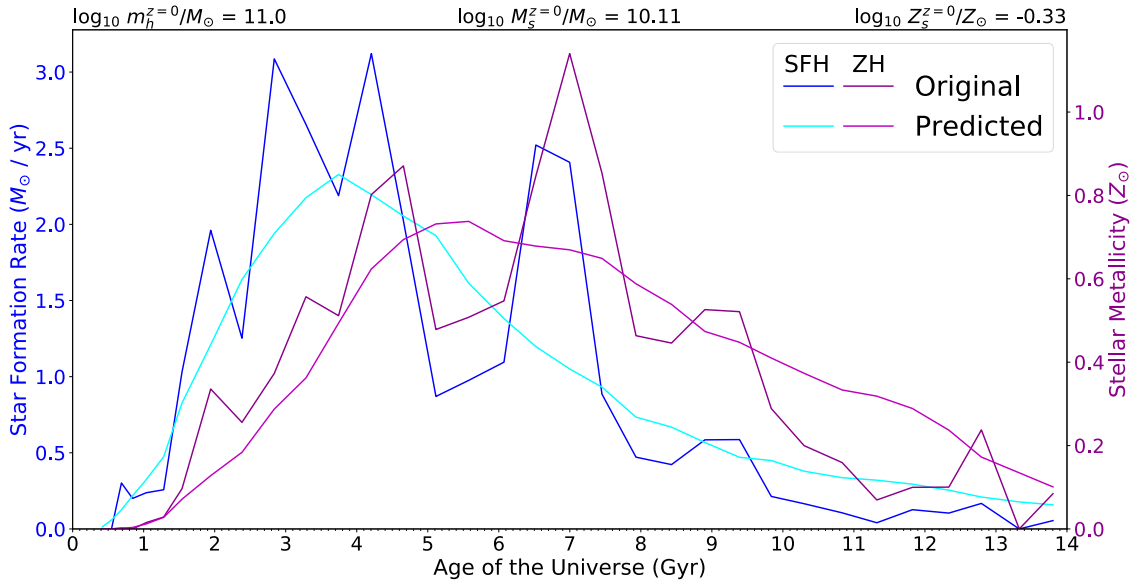


Figure 3.1: This figure illustrates the evolutionary history of a satellite galaxy in IllustrisTNG with intermediate mass. It displays the original star formation history in blue, and its corresponding prediction by the neural network in cyan; as well as the true stellar metallicity history (purple) and the predicted metallicity history (magenta), indicating the time-dependent metallicity of stars formed according to the corresponding star formation rate. The sample's subhalo mass and predicted stellar mass and metallicity values are presented in the header of the figure. While the sample shows a decent match to the shapes of the star formation and metallicity histories, it fails to replicate the fluctuations on short time scales.

This is illustrated in fig. 3.2, where we stack the amplitudes of the Fourier transforms, equivalent to the square root of their power spectra, of SFHs and ZHs in a narrow mass bin, which shows information loss at high frequencies in the predicted data.[1]

The implications of this result are that the network can be used to accurately derive results which depend on the full star formation and metallicity histories, such as stellar mass, luminosity and mass-weighted metallicity. On the contrary, it is less suited to predicting results which depend on the high frequency variability of the predictions, such as the luminosity of emission lines.

3.2.2 Galaxy-Halo Relationships At $z = 0$

Stellar-Halo Mass Relation

Numerically integrating the original and predicted star formation histories results in similar stellar-halo mass relations, shown in fig. 3.3. This is a result which shows the success

[1]The apparent noise of these stacked Fourier amplitudes at high frequencies owes to the fluctuative behaviour of individual Fourier transforms (for examples see fig. 3.11), which does not strictly correlate with mass, and is more apparent in stacks of relatively low sample size. For higher mass galaxies, there is less high frequency information in the predicted data which leads to larger uncertainties in the Fourier transform.

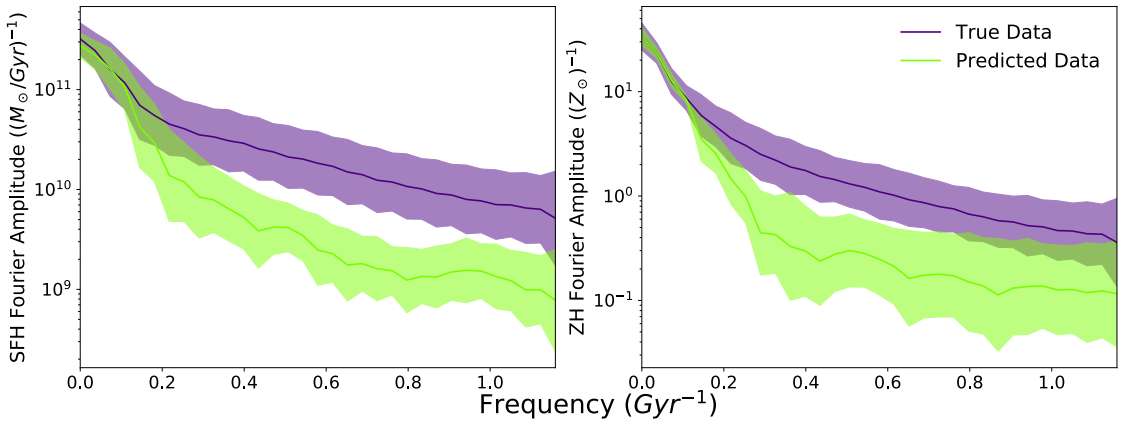


Figure 3.2: This figure shows the mean and standard deviation in the Fourier amplitudes of the star formation histories (left) and metallicity histories (right) of central galaxies in the halo mass range $10^{12.4} - 10^{12.6} M_\odot$, containing 638 samples. The Fourier transforms of the predicted data (green) show a clear decline in amplitude with respect to the original data (purple) for frequencies around 0.15Gyr^{-1} and higher. This indicates the lack of high frequency data in the neural network predictions. These Fourier transforms are plotted up to the Nyquist frequency of approximately 1.1954Gyr^{-1}.

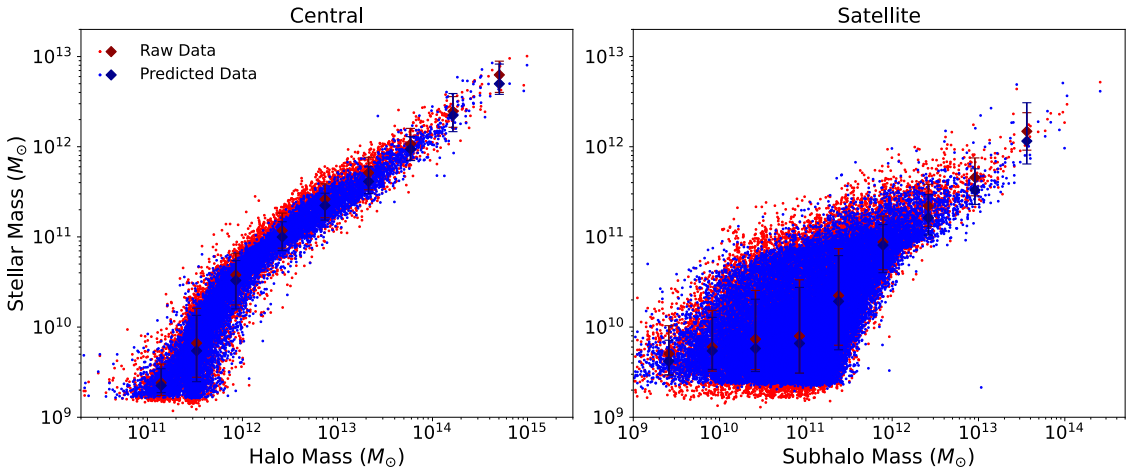


Figure 3.3: The numerical stellar-halo mass relation assessed with the fiducial and predicted star formation rates is shown in this figure for central galaxies (left) and satellite galaxies (right). For the former, this is depicted as a function of halo mass, and for the latter, as a function of subhalo mass. The original TNG dataset's datapoints are presented in red and predictions of the networks are depicted in blue. Red and blue errorbars display the median and 15^{th} and 85^{th} percentiles of stellar mass in halo mass bins, while blue and red datapoints represent individual galaxies. The similarities between the shape and scatter of the two SHMRs suggest that the star formation histories are predicted similarly overall.

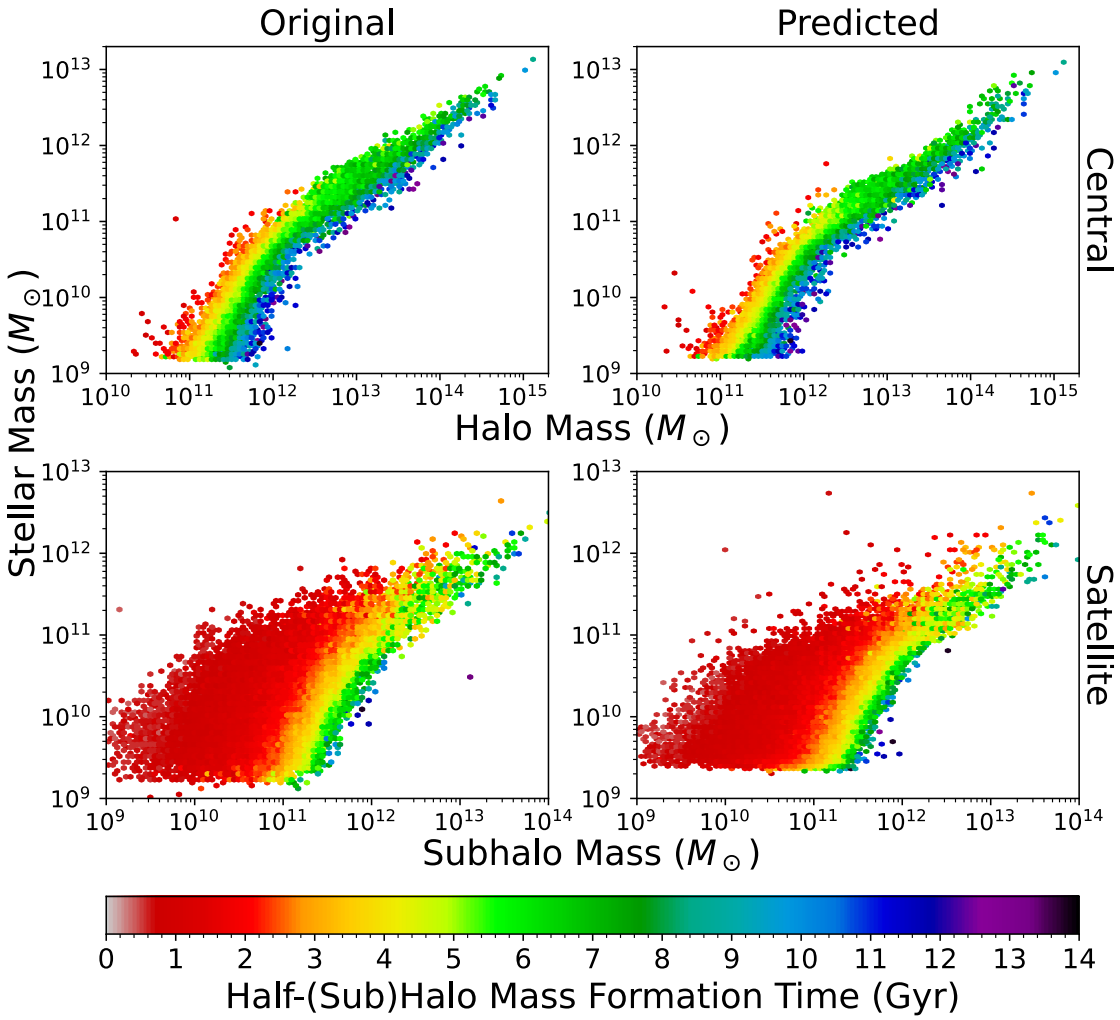


Figure 3.4: Stellar-halo mass relations for central galaxies (top panels) and satellites (bottom panels) are presented in 2D histograms which are coloured in accordance with the mean redshift per bin at which the central or satellite halo's final mass is produced. Plotting according to the individual star masses allows each SHMR to be seen independently for both the original TNG data and the predictions made by the networks. The network's predictions for this dependence of halo formation redshift on the SHMR are consistent with the original data for both central and satellite galaxies. The reader should be aware that the low occupancy of bins on the borders of the SHMRs makes them susceptible to slight variations in scatter, which deceptively gives the SHMRs an apparent distortion.

of the neural networks in recovering the total mass of stars formed in a galaxy's history. The accuracy is reflected in a median absolute residual between original and predicted stellar masses of 0.079 dex for central galaxies and 0.094 dex for satellite galaxies.

We show in fig. 3.4 that these stellar masses are predicted with respect to different halo mass accretion histories. These show the central and satellite SHMRs as 2D histograms, where samples are weighted according to the redshift at which half of their final mass was formed for the first time. This is a key property of the SHMR which signifies the tendency

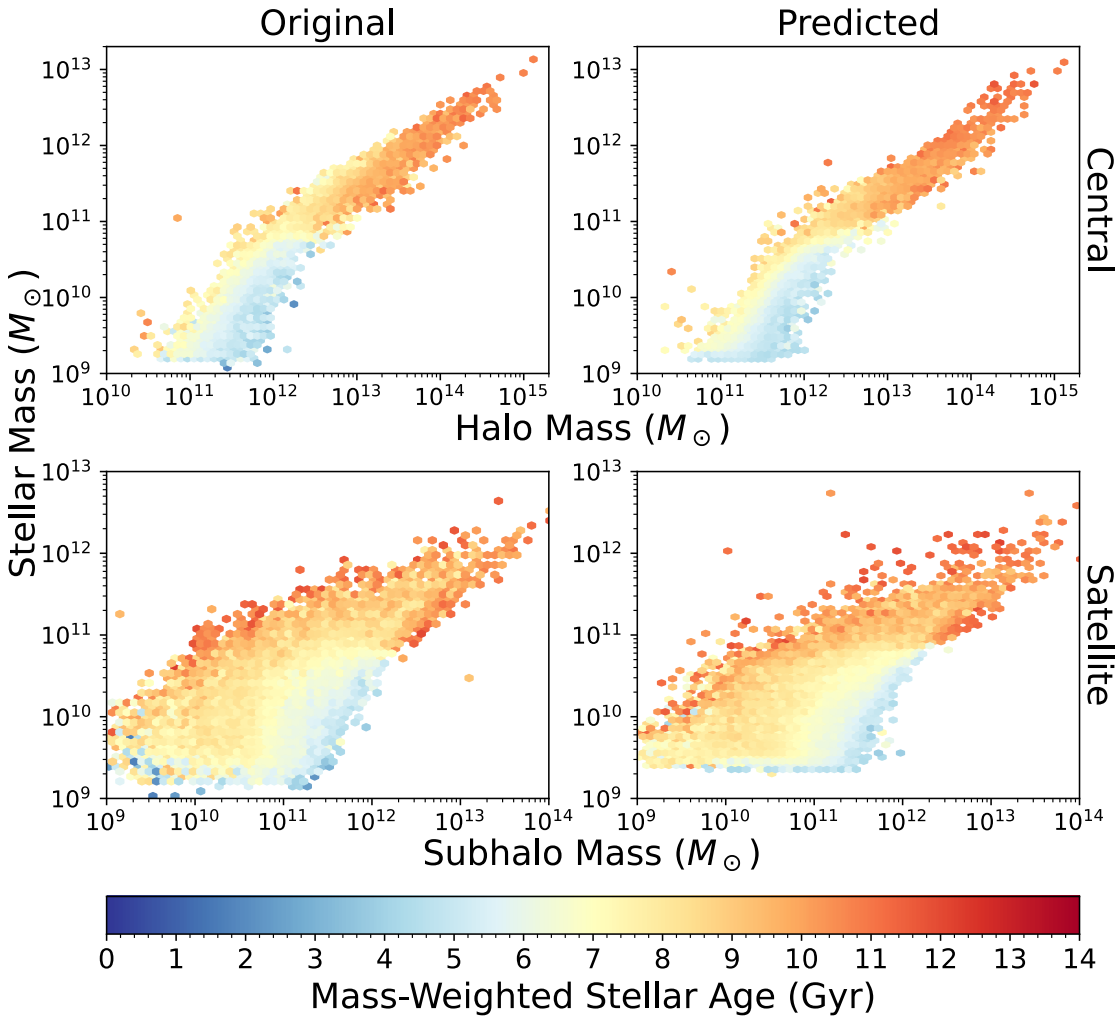


Figure 3.5: Similar figure to fig. 3.4, instead showing the dependence of the mass-weighted stellar ages of galaxies of the SHMR. These results show that this relationship, in which the galaxy age correlates positively with both mass and scatter, is captured by the neural networks and shows in the models' predictions.

for larger galaxies to manifest in rapidly growing halos (Tojeiro et al., 2017; Artale et al., 2018; Zehavi et al., 2019). It is also worth noting that the half-mass formation is not included as an input parameter in either network, signifying that this result has been derived independently from the mass accretion histories.

Star Formation Regimes

Similar to the halo mass formation redshift, we show in fig. 3.5 that the mass-weighted ages, defined in eq. (2.25), are also distributed similarly across the two SHMRs for both original and predicted datasets. The MWA is a measure of the geometry of the star formation history and of the time at which most of the stellar mass is acquired, and is indicative of the different growth regimes of galaxies of a given mass, e.g. merger-driven

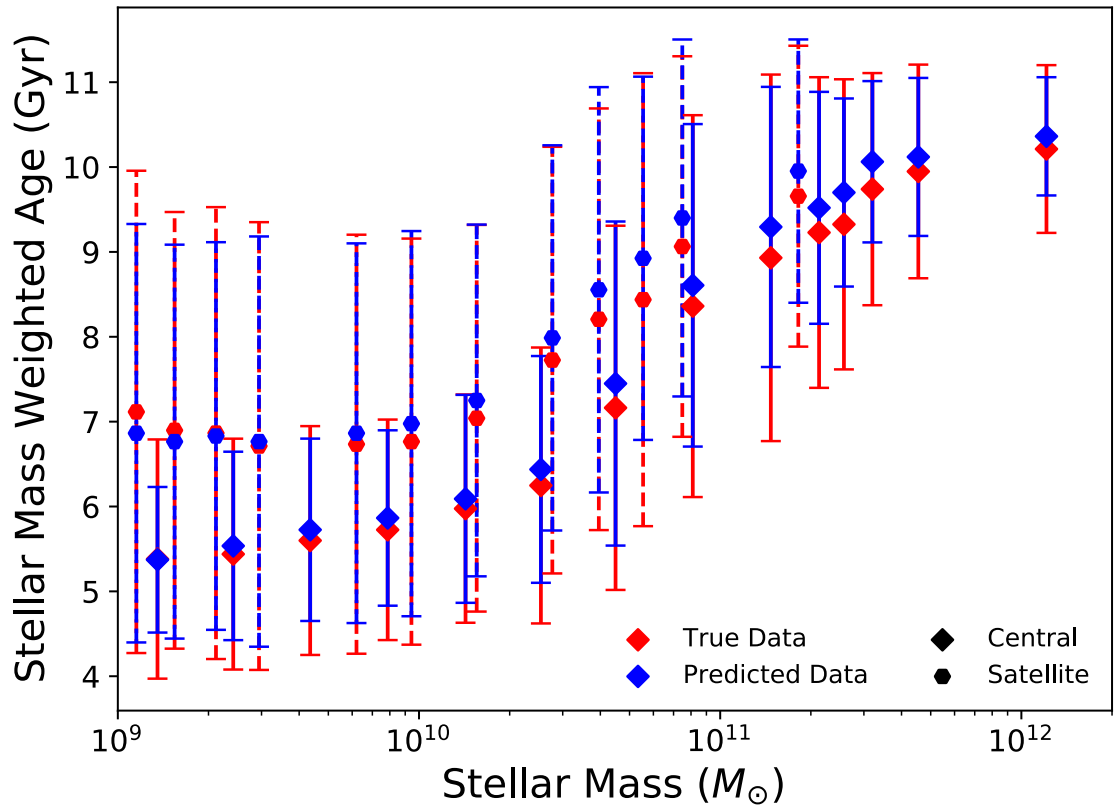


Figure 3.6: The graph displays the relationship between the mass-weighted age of galaxies and their mass, for both central galaxies (diamond points, solid lines) and satellite galaxies (hexagonal points, dashed lines). Each plotted point and error bar represents the median and interquartile range of age values for a specific range of masses. The ages estimated from the predicted history of stellar mass assembly (blue) match the overall trend of ages with respect to mass, as computed from the original data (red). However, some error bars have been shifted or reduced, indicating that some of the samples have significant differences in their mass assembly geometry.

growth vs. smooth accretion.

This figure shows the tendency for galaxy age to increase with halo and stellar mass and the scatter of the SHMR, again being predicted by the neural networks. In short, the results shown in figs. 3.4 and 3.5 indicate that the distinct evolutionary regimes of galaxies modulated by the historical halos or subhalos which host them are encoded in the machine learning models.

When we compare ranges of calculated mass-weighted ages in bins of stellar mass, shown in fig. 3.6, we affirm this correlation between mass and age, but we see that the network's predictions result in a mass-weighted age which is biased towards higher values, particularly at high mass. For high mass galaxies, star formation histories are typically largest at early times, and as these galaxies are quenched their star formation rate becomes insignificant. In the original data there are some small, secondary spikes in this declining

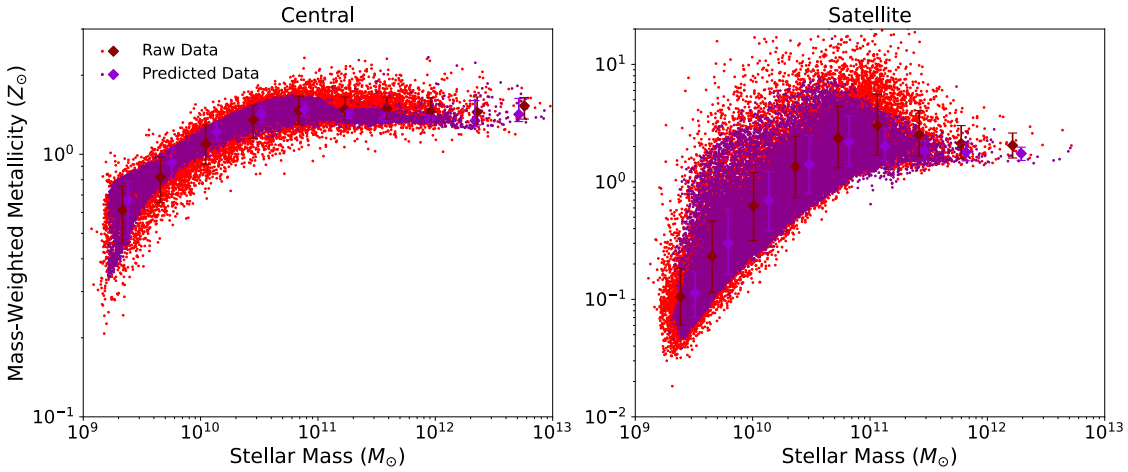


Figure 3.7: In a similar format to fig. 3.3, this figure shows the mass-metallicity relation for central galaxies (left) and satellite galaxies (right), where the original and predicted galaxies are shown in red and purple, respectively, and the errorbars indicate the median and 15$^{\text{th}}$ and 85$^{\text{th}}$ percentiles of stellar metallicity in a given bin of stellar mass. These results show that the neural networks predict a similar MZR shape in both datasets, but the scatter in metallicity is underpredicted, particularly at high mass.

SFH at later times, which as established in section 3.2.1 are seldom predicted by the neural network. Failing to predict these secondary features can serve to over-predict the age of quenched galaxies.

Mass-Metallicity Relation

When computing the mass-metallicity relations of the data, shown in fig. 3.7, we see that for both central and satellite galaxies, the correlation between stellar mass and metallicity is established in the predictions of the neural network. However, the scatter in metallicity is noticeably smaller in the predicted data, owing to the lack of variability on short timescales in both the metallicity history itself and the star formation history which sources these metals.

3.2.3 Comparing Network Designs

In section 2.2 we motivate the design of a semi-recurrent neural network for incorporating static and temporal input variables simultaneously, and including multiple variables with multiple timesteps in a framework which explicitly models causal effects, without adding many thousands of degrees of freedom to the model.

To demonstrate the performance of a semi-recurrent design over a conventional neural network, we train two simplified networks to predict star formation histories of central

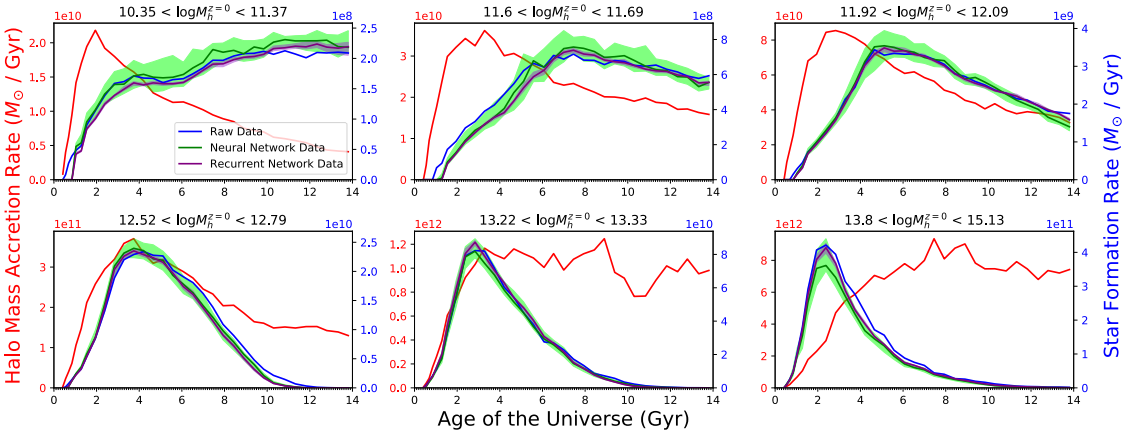


Figure 3.8: Comparison of the predictions for the mean star formation history of central galaxies across six halo mass bins. The predictions were generated by two neural network models: a basic dense neural network shown in green, and a semi-recurrent architecture shown in purple. The median and interquartile ranges of the predicted mean SFH are plotted for both networks, based on ten independent runs of each network. The true mean star formation history in each bin is shown in blue, while the averaged mass accretion history is shown in red. The difference in interquartile ranges between the predictions of the two networks demonstrates a significant superiority of the semi-recurrent model to converge accurately.

galaxies, using only the following input variables: $M_h^{z=0}$, $\dot{M}_h$, $t_{\rm start}$ and $\mid \dot{M}_h \mid$. One is a semi-recurrent network which includes the mass accretion historry in its temporal input layer, and others in its static input layer. The other is a simple neural network which includes all variables in its solitary input layer.

In fig. 3.8 we show the consistency of ten independent predictions of the two networks in six bins of halo mass. The networks both show similar correlations between the halo and galaxy formation histories, in terms of both amplitude and shape, representing total mass and formation times respectively. However, the variance in predictions of the simple neural network is clearly larger than that of the semi-recurrent network; in fact this is usually an order of magnitude or two larger. The semi-recurrent network therefore has much faster convergence speed as a result of its enforced causal connections, and this justifies our use of a semi-recurrent design, in addition to the ability to add multiple temporal variables without significantly increasing the multiplicity of the parameter space associated with the network.

3.2.4 Quality Of Predictions

Outcome

The direct predictions of galaxy formation history have proven to be fruitful in the synthesis of SFHs and ZHs of both central and satellite galaxies, in a manner which self-consistently replicates principal relationships between galaxies and halos which encapsulate the different regimes of galaxy-halo coevolution. For instance, the scatter of the SHMR is correlated with the halo formation time, indicating the importance of rapid dark matter accretion in growing galaxies, in both the original and predicted datasets. The network can therefore be used to reproduce this relationship in N-body applications, and may be used to make additional insights into the galaxy-halo connection in future.

Despite this outcome, the network scarcely predicts the details of these star formation and metallicity histories which take place on smaller timescales, which have adverse effects on the calculation of metallicity scatter and stellar age, and on derived observable quantities (see chapter 4). The model can therefore be used to predict the evolutionary regimes of galaxy formation across cosmic time, but not the properties of quick events such as starbursts or rapid quenching.

Despite similar results of the prediction quality of star formation and metallicity histories, the calculation of mass-weighted metallicity suffers when compared with integrated stellar mass. The chemical content of a massive galaxy can be enhanced by a combination of events occurring on short timescales, from merger events to rapid star formation or gas accretion. The overall lack of metallicity scatter may be attributed to the abscence of these high frequency features in the predicted data, to which high mass galaxies with large metallicity scatter may be particularly sensitive.

This illustrates that the shortcomings of the predictions of the neural network, namely the lack of variability in a single star formation or metallicity history, have negative impacts on the true evolution of the star formation over time, and on the final metallicity of the galaxy. These errors are reflected in our metrics of the quality of prediction, such as mass-weighted age; and will contribute to systematic errors in physical quantities which are derived from these results, such as the galaxy's stellar luminosity. Mitigating these results is therefore desirable for more accurate galaxy catalogues and mocks.

Centrals vs. Satellites

We find that the quality of the predictions of the satellite galaxy model are marginally better than those of the central galaxy model in the following respect. After examining the star formation histories of quenched galaxies, it appears that the decline in star formation and the rate of that decline match well for satellite galaxies, and not necessarily for central galaxies.

The satellite network specifically includes variables pertaining to their infall, which of course have notable effects on the star formation history, such as satellite quenching. This implies that the star formation history of satellites is naturally more constrained than central galaxies, and thus easier to predict. Though the infall parameters are themselves derived from the mass accretion histories of the satellite subhalo and host halo, the relationship between these quantities is potentially valuable to the network.

Specifically, the relationship between $a_{\rm max}$ and $a_{\rm infall}$ can indicate the time in the subhalo's growth history at which it is acquired by a larger halo. Where $a_{\rm max}$ is larger, for instance, the subhalo is continuing to accrete mass and form stars as it becomes bound; unlike quenched satellites which acquire most of their mass in the central phase. The mass ratio and velocity upon infall can provide a measure of the subsequent rate of mass loss, yet this is also dependent on the mode of accretion at the time, therefore their relationship with $a_{\rm max}$ and $a_{\rm infall}$ are potentially important to the network.

On the contrary, the deduction made by scaled time quantities may be inferred from the growth histories of the subhalo and host halo, suggesting that they are not so important to the model. Shi et al. (2020) concluded that star formation rates behave differently between rapidly and slowly growing subhalos, which can be derived from their mass accretion histories. The merger histories that correlate with their scaled formation time can be inferred from overdensity and skew histories. Thus, while the relationship between infall properties may be a useful constraint on galaxy evolution for the satellite model, it is not strictly a unique constraint.

Improving High Frequency Predictions

It is possible that the inclusion of additional variables may have improved the quality of these predictions. It is widely accepted that the properties of the gas in the progenitor

subhalos are crucial to subsequent star formation (Hani et al., 2020; Trevisan et al., 2021; Sorini et al., 2022). While these gas properties depend on the local environment, these studies show that the environment is not a distinguishing factor and that numerous other processes such as stellar feedback can influence the subhalos' affinity for enhanced star formation.

Alternatively, a direct measure of the merger history, such as the dimensionless merger and smooth accretion proxies used by Dhoke & Paranjape (2021) may have predicted the enhancement in star formation owing to major merger events. As written in section 2.4.1, static variables relating to merger activity have been difficult to implement and fruitless in improving predictions, motivating the use of skew as a temporal interaction parameter.

These dimensionless accretion parameters are direct measures of the mass acquired by major and minor collisions, and may be useful in predicting the star formation driven by these forms of accretion separately. However, these do not capture close interactions which trigger tidal distortions as skew does, and thus is not a complete replacement of the interaction history. These accretion rates are also implicitly dependent on dynamical time intervals between snapshots and on the matter power spectrum, therefore being potentially unsuitable for use in simulations of different cosmological parameters and spatial and temporal resolution.

In section 3.3 we outline how the successful prediction of the power spectra of galaxy formation histories has advocated stochastic amendments to the original predictions of the neural network, which replicate the noisy behaviour of the original star formation and metallicity histories.

3.3 Stochastic Corrections

3.3.1 Motivation

The fiducial predictions of the neural network have shown that predicting the features of star formation histories on short timescales is difficult. The processes which drive the fluctuations are numerous and dependent on effects ranging from cosmic ray and photoionisation feedback to mergers and stellar winds on timescales below the scales captured by the neural networks (Kannan et al., 2022; Robaina et al., 2010; Iyer et al., 2020). An

accurate model of this high frequency star formation variability over the galaxy's complete evolutionary history would be beneficial in applications to N-body simulations as it can be used to compute the effects of halos and environment on these events, and would result in more accurate observational features which are used to constrain these effects.

The same dark matter properties and an identical network design were used to predict the absolute amplitude of the Fourier transforms of star formation and metallicity histories, or the square root of their power spectra, for central and satellite galaxies. The stacked Fourier transforms shown in fig. 3.9 exhibit similar degrees of precision as the star formation and metallicity histories themselves are found in the structure of the power spectrum and correlation of its amplitude with halo mass. Therefore, a model that predicts these Fourier transforms and produces a stochastic signal could be created to reduce the errors in our SFH and ZH predictions, including in N-body applications. We use this method to recompute summary statistics such as the SHMR (see section 3.3.3) and observational statistics (see chapter 4).

3.3.2 Methodology

Consider an additive relationship between time-dependent signals denoted $h(t)$, representing a star formation or metallicity history:

$$h(t)_{\text{true}} = h(t)_{\text{pred}} + h(t)_{\text{stoc}} \tag{3.1}$$

where $h(t)_{\text{true}}$ represents the ideal temporal signal, $h(t)_{\text{pred}}$ is the signal predicted by the neural network, and $h(t)_{\text{stoc}}$ is the desired correction term.

We consider the Fourier transforms of the temporal signals from the TNG data and the neural network, respectively:

$$\begin{aligned} H(f)_{\text{true}} &\equiv \mathscr{F}\left[h(t)_{\text{true}}\right] \\ H(f)_{\text{pred}} &\equiv \mathscr{F}\left[h(t)_{\text{pred}}\right] \end{aligned} \tag{3.2}$$

where $f \equiv t^{-1}$, and we exploit the fact that the transformed signals are additive provided that the original signals are additive, as in eq. (3.1):

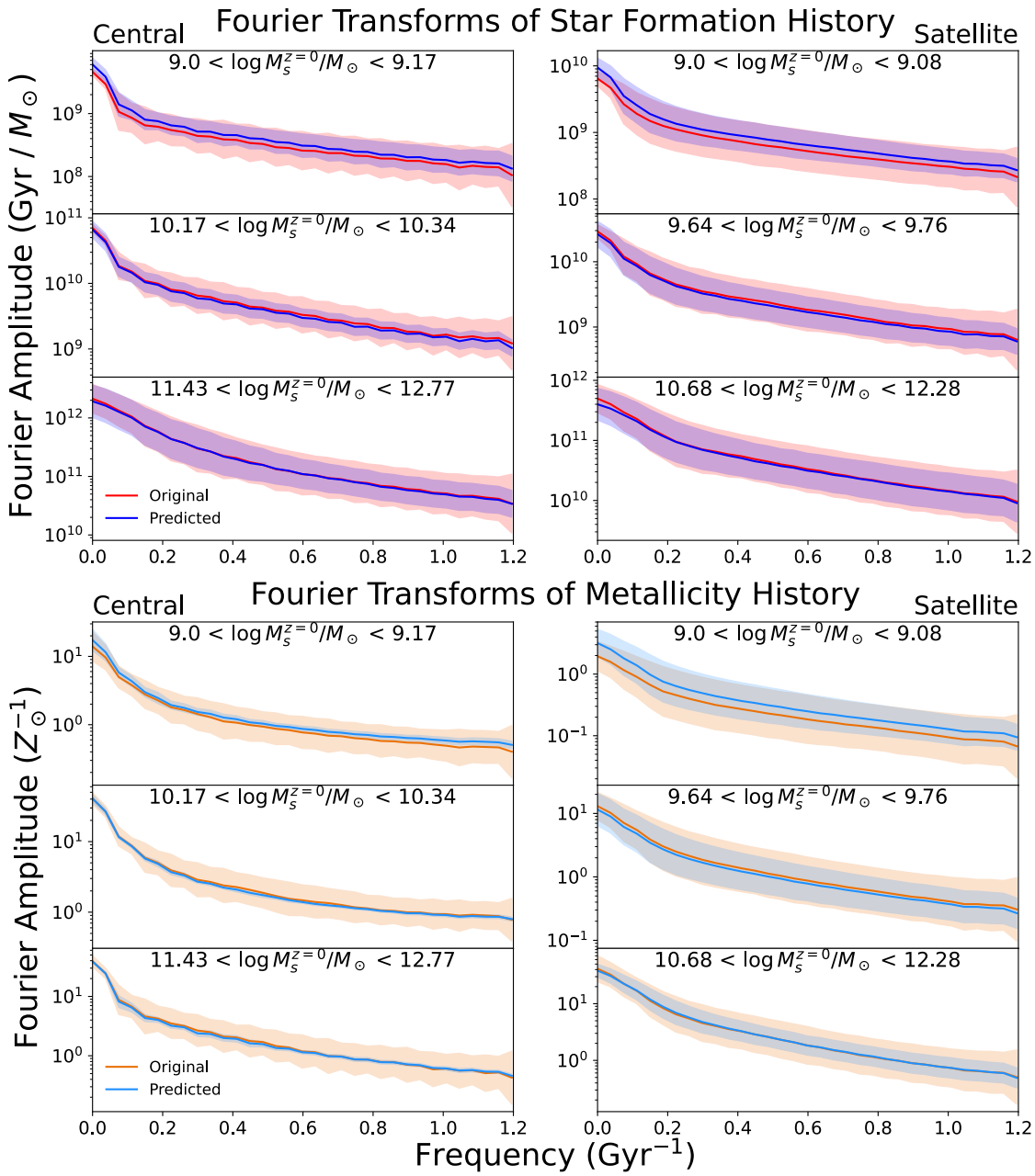


Figure 3.9: The amplitudes of the Fourier transforms of the star formation histories (top panels) and metallicity histories (bottom panels) for central galaxies (left column) and satellite galaxies (right column), in three narrow bins of stellar mass. This shows the Fourier amplitudes for the original TNG data (red) against the Fourier amplitudes predicted by the neural network when trained to fit said amplitudes (blue). The quality of the predicted Fourier transforms constitute a superior fit to the original data than the star formation or metallicity histories predicted directly by the network (see fig. 3.2), and therefore the network can be used to produce a stochastic amendment to its own predictions.

$$H(f)_{\text{true}} = H(f)_{\text{pred}} + H(f)_{\text{stoc}} \tag{3.3}$$

where $H(f)_{\text{true}}$ is the Fourier transform predicted by the neural network, $H(f)_{\text{pred}}$ is the Fourier transform of the predicted temporal quantity, and $H(f)_{\text{stoc}}$ is the Fourier transform of the desired stochastic signal.

Fourier tranforms are complex functions of real-valued amplitude and phase, and can represented as follows:

$$\begin{aligned} H(f)_{\text{true}} &= A_{\text{true}} e^{i\theta_{\text{true}}} = A_{\text{true}} \cos\theta_{\text{true}} + iA_{\text{true}} \sin\theta_{\text{true}} \\ H(f)_{\text{pred}} &= A_{\text{pred}} e^{i\theta_{\text{pred}}} = A_{\text{pred}} \cos\theta_{\text{pred}} + iA_{\text{pred}} \sin\theta_{\text{pred}} \\ H(f)_{\text{stoc}} &= A_{\text{stoc}} e^{i\theta_{\text{stoc}}} = A_{\text{stoc}} \cos\theta_{\text{stoc}} + iA_{\text{stoc}} \sin\theta_{\text{stoc}} \end{aligned} \tag{3.4}$$

By substituting the expanded terms of eq. (3.4) into eq. (3.3), we obtain expressions for the real and imaginary components of the true Fourier transform:

$$\begin{aligned} \Re\left[H(f)_{\text{true}}\right] &= A_{\text{true}} \cos\theta_{\text{true}} = A_{\text{pred}} \cos\theta_{\text{pred}} + A_{\text{stoc}} \cos\theta_{\text{stoc}} \\ \Im\left[H(f)_{\text{true}}\right] &= A_{\text{true}} \sin\theta_{\text{true}} = A_{\text{pred}} \sin\theta_{\text{pred}} + A_{\text{stoc}} \sin\theta_{\text{stoc}} \end{aligned} \tag{3.5}$$

which can be rearranged to solve for the variables A_{stoc} and θ_{stoc}:

$$\begin{aligned} A_{\text{stoc}} &= \sqrt{H_{\text{real}}^2 + H_{\text{img}}^2} \\ \theta_{\text{stoc}} &= \tan^{-1}\left(\frac{H_{\text{img}}}{H_{\text{real}}}\right) \end{aligned} \tag{3.6}$$

where,

$$\begin{aligned} H_{\text{real}} &= A_{\text{true}} \cos\theta_{\text{true}} - A_{\text{pred}} \cos\theta_{\text{pred}} \\ H_{\text{img}} &= A_{\text{true}} \sin\theta_{\text{true}} - A_{\text{pred}} \sin\theta_{\text{pred}} \end{aligned} \tag{3.7}$$

Thus, by computing the amplitude and phase of the stochastic component of the Fourier transform, as in eq. (3.6), we can obtain the temporal stochastic signal simply by applying an inverse Fourier transform:

$$h(t)_{\text{stoc}} = \mathscr{F}^{-1}\left[H(f)_{\text{stoc}}\right] = \mathscr{F}^{-1}\left[A_{\text{stoc}}e^{i\theta_{\text{stoc}}}\right] \tag{3.8}$$

We thus compute the desired stochastic signal by using the amplitude and phase of the transformed temporal predictions as A_{pred} and θ_{pred}, and the Fourier amplitudes predicted by the neural network as A_{true}. The network cannot be used to predict the phase information of the transforms, so we apply a small, semi-random shift to the true phases:

$$\theta_{\text{true}} = \theta_{\text{pred}} + \delta \tag{3.9}$$

where δ is sampled from a zero-centered Gaussian distribution of width 0.1 radians, restricted to a range of $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$ radians.

The star formation and metallicity histories are thus applied to this method to return a stochastic SFH or ZH component. These are then added to the predicted star formation and metallicity histories to produce a modified result. We then use these modified star formation and metallicity histories to reproduce key galaxy-halo relations to assess the validity of the stochastic amendment.

As we intend to retain the smooth shapes of the star formation and metallicity histories, which have been predicted accurately, we apply this correction only to high frequencies, where variability is lost. Specifically, the geometry is preserved up to a Fourier frequency of approximately 0.18Gyr^{-1}, corresponding to the approximate scale of information loss seen in fig. 3.2.

3.3.3 Results

In fig. 3.10, we show the star formation history and metallicity history of a single satellite galaxy, including the original TNG sample, the network's fiducial prediction and the modified result. This shows that the stochastic correction can be used to reproduce the fluctuative behaviour seen in the original star formation and metallicity histories. However, the SFH and ZH is not significantly reshaped by the stochastic correction. It does not contain a number of high amplitude features seen in the original sample, which will influence the predicted mass, optical spectrum and other key features of the galaxy.

As the neural network predicts the Fourier amplitudes in similar detail to the SFHs and

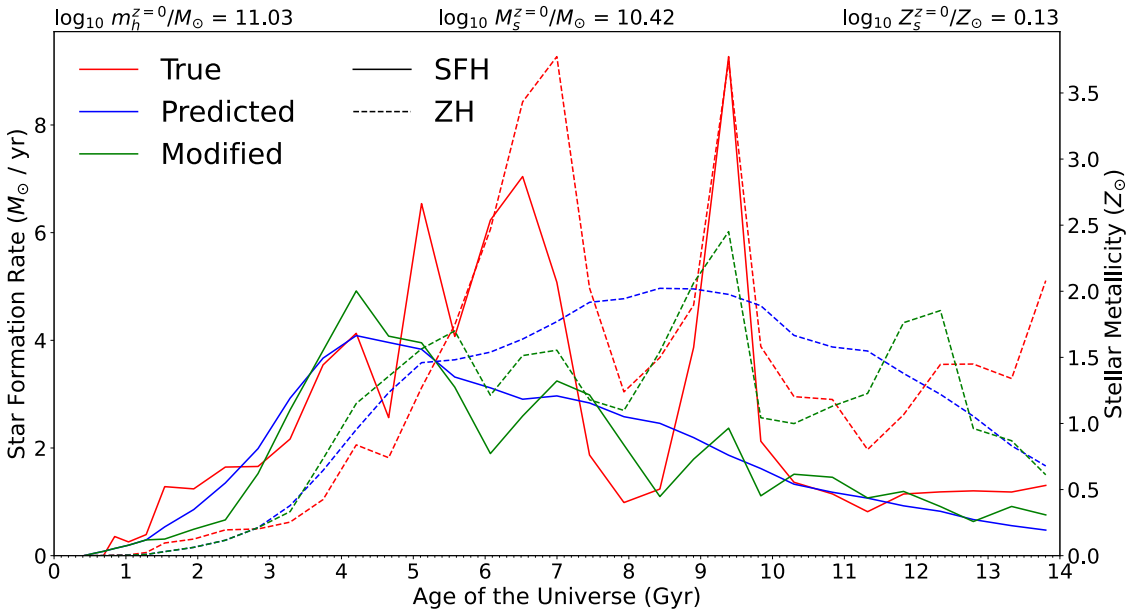


Figure 3.10: Star formation histories (solid lines) and metallicity histories (dashed lines) of an exemplary intermediate mass satellite galaxy, showing the original TNG sample (red), the prediction of the neural network (blue) and the stochastically modified result (green). This shows a typical result where the fluctuative behaviour of the original galaxy is reproduced by the stochastic amendment, yet the overall shape of the star formation and metallicity histories is not significantly changed to match the target SFH and ZH.

ZHs themselves, it is possible that missing features of the Fourier transforms correspond to critical information relating to the shape of predictions in the time domain. While the Fourier amplitudes have well-defined shapes, some samples have additional fluctuations in the Fourier spectra which are not necessarily captured by the model, which can be seen by example in fig. 3.11. Infrequent events such as massive starbursts may be entailed by such fluctuations and consequently may be absent from modified results. For this example, the power spectrum of the modified SFH is not fully recovered at the highest frequencies, which would explain the lack of short peaks in the SFH itself.

Comparing the SHMRs of the modified data with predictions and true data in fig. 3.12, we see that the stochastic amendment has made some modest improvements. The median points and errorbars in fig. 3.12 show that the modified SHMR is closer in amplitude and scatter to the original data than the network predictions, for most of the central and satellite data. The exceptions are high mass objects in both datasets, where sample size is low; and low mass satellites, where modified masses are overpredicted. Figure 3.9 shows that predicted Fourier transforms are offset for the lowest mass galaxies, which contributes to both centrals and satellites, but the effect is greater for the latter.

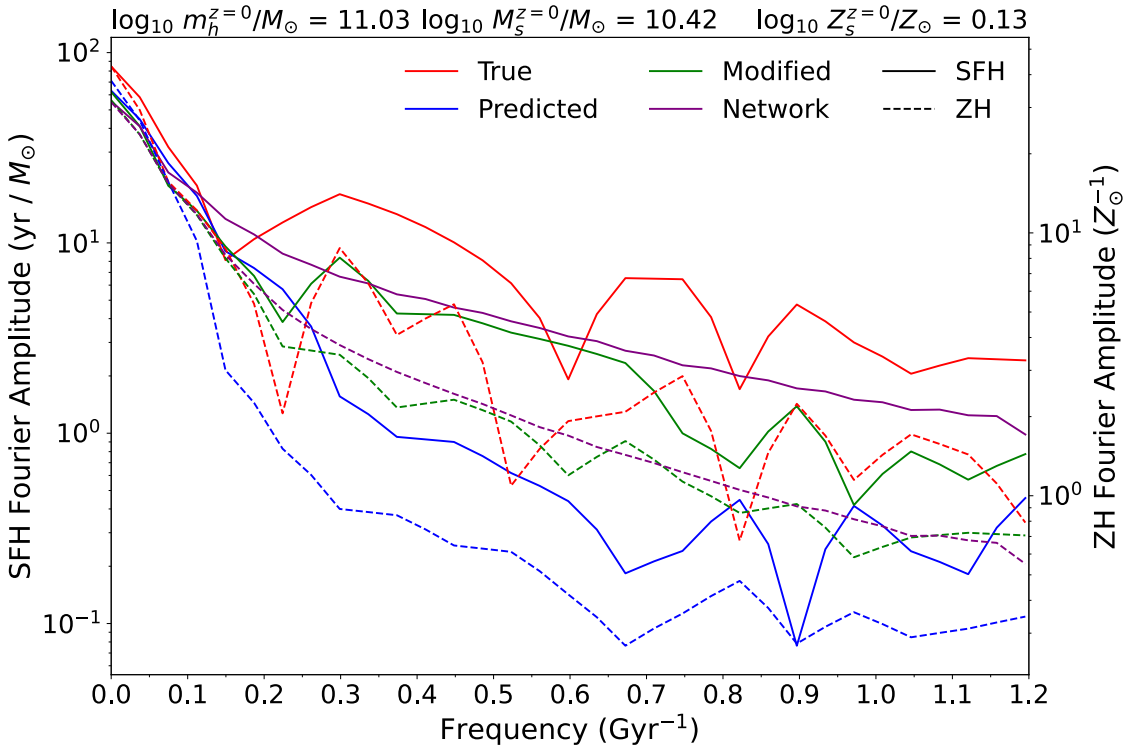


Figure 3.11: Absolute-valued Fourier transforms of the data in fig. 3.10, adopting the same choice of colours and linestyles, with the additional field of the Fourier amplitude predicted by the neural network shown in purple. This shows that the stochastic modification brings the Fourier transforms of the star formation and metallicity histories of this sample close to that of the TNG data, yet the predicted transform itself does not include distinctive features seen in the original transform, which in turn is absent from the modified result.

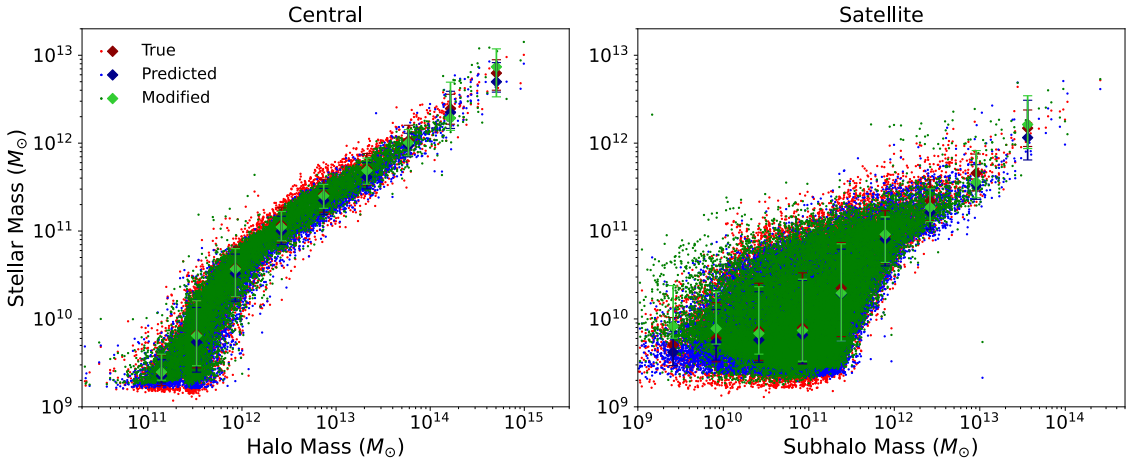


Figure 3.12: Stellar-halo mass relations, showing the original TNG data (red), the predictions of the neural network (blue) and the stochastically modified results (green). For each dataset, individual points represent samples, whereas errorbars indicate the median and $15^{\text{th}} - 85^{\text{th}}$ percentile ranges of stellar mass in bins of halo mass. The stochastic correction shows a modest improvement to the amplitude and scatter of the precdicted SHMR.

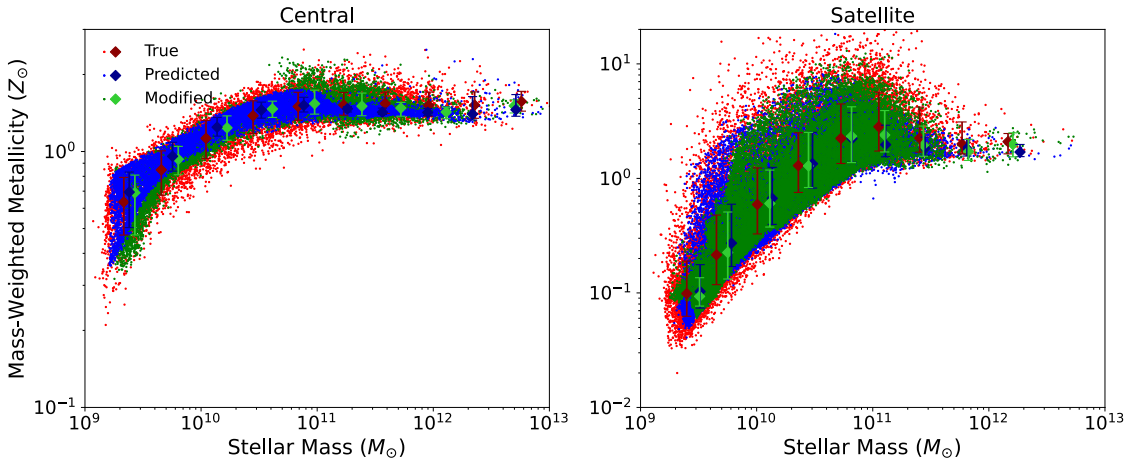


Figure 3.13: Stellar mass-metallicity relations, showing the original TNG data (red), the predictions of the neural network (blue) and the stochastically modified results (green). For each dataset, individual points represent samples, whereas errorbars indicate the median and $15^{\text{th}} - 85^{\text{th}}$ percentile ranges of metallicity in bins of stellar mass. The stochastic correction provides a significant improvement to the scatter of the MZR for central galaxies, yet it performs poorly for low mass satellite galaxies.

The scatter of the SHMRs may have been visibly improved by the stochastic correction, yet the scatter of the target SHMR is not fully reproduced by the modified data. As established, the correction does not completely recover the original star formation histories, and does not include important features such as high mass spikes in the star formation history. These, of course, influence the final stellar mass of the galaxy, and therefore contribute significantly to the SHMR scatter.

Inspecting mass-metallicity relations in fig. 3.13, we see improved results for central galaxies, where the scatter in metallicity, which was poorly predicted by the neural network, is much more reminiscent of the scatter of the TNG simulation. Satellite metallicities, on the contrary, are not significantly improved by the modification, and some subpopulations of the data have highly distorted metallicities which separate from the MZR. This applies predominantly to low mass galaxies with a high stellar to subhalo mass ratio, which correspond to samples with high predicted Fourier amplitudes, and have fluctuations larger than those of the original TNG data. The majority of satellite samples, however, are not disrupted this way.

For low mass satellite galaxies, however, the mass-metallicity relations do not differ substantially between the TNG and network results, and so the ideal stochastic correction may only make minor improvements to these galaxies. For high mass satellites, on the other hand, the metallicity scatter is visibly improved, but only by a small margin. The

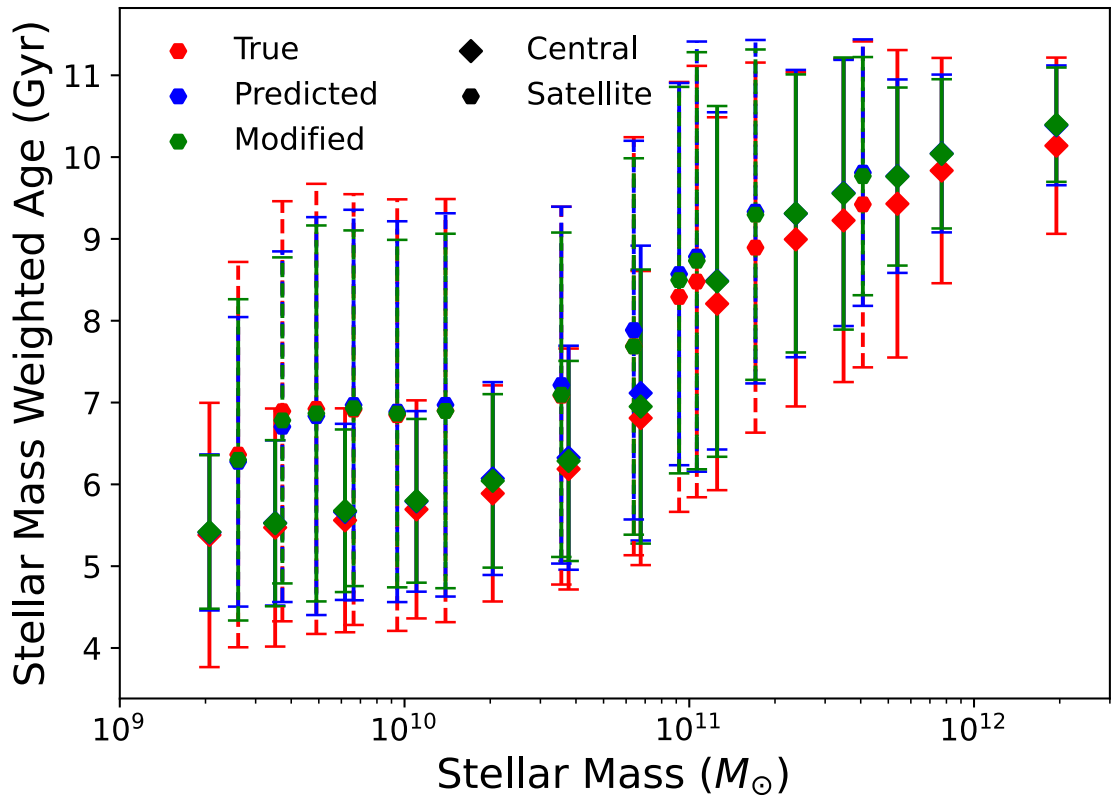


Figure 3.14: Median and interquartile ranges of mass-weighted ages as a function of stellar mass for central and satellite galaxies, showing the original TNG data (red), the predictions of the neural network (blue) and the stochastically modified results (green). Median stochastic mass-weighted ages are closer to that of the original data for most samples, showing that the correction improves the inferred geometry of the star formation histories; yet its smaller ranges show that these shapes are too generalised.

method therefore has the potential to refine the satellite mass-metallicity relation as well.

Like the SHMRs, however, neither of the two MZRs are completely recovered by the stochastic amendment, due to the lack of large transient features in the modified star formation and metallicity histories. The size and adequacy of the adjustment of the MZR scatter, however, differs for galaxies of different masses. This suggests that the optimal frequency range over which to apply the correction varies according to the mass and evolutionary regime of the galaxy.

Finally, in fig. 3.14 we compare the mass-weighted ages of central and satellite galaxies before and after modification, in order to measure the effect of the modification on SFH geometry. This shows that for most stellar mass bins, the stochastic amendment brings the median ages closer to those of the TNG data, suggesting that the adjustment is capable of recovering star formation features at specific times accurately. However, this is, for the most part, a small effect. The ages of galaxies in a given mass bin are more likely to be

influenced by large features which the stochastic correction does not predict. Photometric quantities may not be visibly adjusted due to the galaxy ages being mostly unaffected by the correction.

3.3.4 Interpretation

These results effectively demonstrate that the variability in the amplitudes of various frequency components is necessary to obtain histories of the same accuracy and detail of the TNG data. The level of detail of the predicted Fourier transforms and the rarity of features have profound effects on the SFHs and ZHs.

The correction may be improved by the inclusion of input variables which may influence events on certain timescales, such as the inclusion of merger ratios to predict star formation events of certain timescales irrespective of their phase. However, with the size and timescale of some star formation events depending on baryonic factors such as the gas richness of mergers and accretion, it is possible that constraining such SFH features cannot be achieved by this stochastic method, particularly for events within the phase error margin. An alternative correction may be based on wavelet transforms, which unlike Fourier transforms capture local frequency information, thus capturing transient signals which correspond to isolated events which shape the SFHs and ZHs.

This project is a work in progress, and in recent months, efforts have been made to correlate the phases of the stochastic component with the mass accretion history of the galaxy's halo, as well as simultaneously sampling phases from star formation and metallicity histories to preserve any correlation which exists between them, and drawing from the phase distributions associated with bins of mass, metallicity and age. This approach has thus far introduced some star formation history features which are closer in time to those of the TNG galaxy, making small improvements for intermediate-mass star-forming galaxies. However, it struggles to constrain the amplitude of high mass galaxy spectra, and biases the mass-weighted ages of these galaxies toward lower ages. While this strategy may not be complete, it may in future replicate high frequency SFH and ZH features with a realistic time domain, which may be used to compute age-dependent galaxy features such as emission line luminosities more accurately.

The stochastic correction has nonetheless shown that missing variability in the fiducial

predictions of the neural network can be partially recovered by means of a stochastic component generated from the Fourier spectra predicted by a network of the same design and input data. This variability has been shown here to make valuable improvements to the statistics of physical galaxy properties, and in chapter 4 we show the influence that this has on observational quantities.

3.4 Feature Importance

Having developed a machine learning model capable of predicting galaxy evolution via the galaxy-halo connection, we wish to investigate the properties which control these predictions. This section describes the methodology used to identify the input features of each neural network that have the most predictive power over star formation and metallicity histories, in effect modulating the scatter of the relations of galaxy properties with halo mass.

Using a common algorithm to determine feature importance, such as a Random Forest Regressor (RFR), is inadequate for this task because of the complex interplay between the variables and the final result; specifically, the multidimensional influence of temporal and static variables is impractical to characterise with a simple, scalar quantity. Instead, we conduct a test where the network is trained multiple times, while groups of similar features are scrambled to eliminate their signal.

This method is similar to permutation importance, but it differs in that the summary statistics are derived from predictions of the network when trained with scrambled data, and compared to those from the predictions of the complete model. This allows one to measure physical properties of the data after scrambling and identify the importance of randomised quantities on these properties, regardless of whether they are explicitly given as model parameters.

3.4.1 Connected Input Properties

Quantities used for predictions in the neural network are usually related in some way; the mass accretion rate integrates to give the final halo mass, and overdensities evaluated on different scales are inevitably correlated. In order to measure the effects of related groups of variables, we retrain the neural networks where physically connected parameters are

simultaneously scrambled, whether temporal or static. We refer to these sets of shuffled variables as "shuffle groups".

For each input variable in the neural networks, their shuffle groups are indicated by the "shuffle" column in table 2.2, which includes full details of the shuffle groups. The five shuffle groups are as follows:

- **Group 1** Mass and growth history of the main halo. Includes mass accretion history, final halo mass, accretion gradient, etc.
- **Group 1a** Defined only for the satellite data. Akin to shuffle group 1 for the satellite subhalo, while group 1 refers to its host halo. Includes mass accretion history for the subhalo, as well as infall variables which derive from this mass accretion history. For central galaxies the halo and subhalo accretion histories are congruent and infall variables do not apply, making this group redundant.
- **Group 2** Environment group. Includes overdensity histories and cosmic web distances.
- **Group 3** Halo substructure group. Includes half-mass radius and circular velocity.
- **Group 4** Interaction history group. Consists of radial skew and distance to closest external subhalo.

3.4.2 Methodology

Diverging Scatter

For each shuffle group, the neural networks are retrained where one of the shuffle groups is individually randomised. Specifically, as the data is normalised by quantile transformation to a normal distribution (see section 2.5.1), the input data is replaced with Gaussian random noise prior to training.

In order to measure the effect that these shuffle groups have on the predictions of the neural network, for each trained network we compute the scatter in stellar mass, mass-weighted metallicity and mean metallicity history as a function of halo or subhalo mass, and compare this with the fiducial result. By computing these three quantities, we are effectively measuring the difference in overall accuracy of the star formation and metallicity histories, as well as the mass-weighted metallicity value which depends on them

both, thereby measuring the importance of these quantities for predicting star formation, full chemical enrichment and metal synthesis independently of mass.

We conduct ten independent runs for each data randomisation, and assess the significance of each shuffle group using the median and interquartile ranges in the difference in scatter. Specifically, we define the quantity Ξ:

$$\Xi \equiv \left\langle \log_{10} \left| 1 - \frac{\sigma_x}{\sigma_0} \right| \right\rangle \tag{3.10}$$

where σ_x is the characteristic scatter of the baryonic variable in dex when scrambling shuffle group x, and σ_0 is the scatter of the fiducial relation. σ_0 is calculated using the predictions of the model and not the original simulation data, such that Ξ qualifies as a measure of the influence of specific features, not a measure of the irrelevant systematic errors in the neural network, such as the lack of short-timescale information.

The scatter in a target quantity is computed using a running standard deviation filter with a window size of 500 samples. Ξ is then calculated by binning the scatter ratio σ_x/σ_0 in 80 loguniform bins of halo mass, and computing the average "local Ξ" value, weighted by the occupancy of each bin in order to minimise the bias from bins of low sample size. This operation is indicated by the angled brackets in eq. (3.10). Similarly, local Ξ values are discarded when they are smaller than -2, as this can bias the results. This typically happens when the difference in scatter is both small and regularly fluctuates to either side of the fiducial value.

The value of Ξ indicates the characteristic size of the divergence in predicted stellar mass, metallicity, etc. when the information from the given shuffle group is lost. Where the scatter in fiducial and disrupted models are identical, Ξ assumes a value of negative infinity, which would indicate that the shuffle group has no influence on the model whatsoever. As Ξ increases in value, it indicates larger differences between the average scatter of the two relations, showing that the scatter is supported by a parameter or multiple parameters in the given shuffle group.

However, differences in this scatter are small, and some differences from a given run may be a spurious result of random errors in the network's coefficients. This is why we run each network ten times, and compute the median and interquartile range of Ξ values.

A small interquartile range compared with the median suggests that this shift in scatter is significant, while a large interquartile range suggests it is spurious.

To determine the important variables of a particular shuffle group, one can repeat this methodology where only the quantity of interest is randomised, which determines whether the variable has a significant effect on predictions, despite being physically related to other members of the shuffle group. We demonstrate this in the results subsection with two examples; where we consider the role of overdensity and the cosmic web in the central galaxy data, and infall properties of satellite halos in contrast with their full accretion history.

Diverging Distributions

As well as measuring the effect of the shuffle groups on the scatter, we measure their influence on the fits to the relations themselves. To do this, we calculate the median filter of each relation and replace the scatter in eq. (3.10) with the logarithmic median of each baryonic variable. Unlike the scatter offsets, these offsets in the amplitudes of the relations are not significant for any shuffle group, implying that no single input parameter has a unique influence on the shapes of the relations themselves.

Despite this, the differences in the scatter illustrated by Ξ ultimately serve to shift the distribution of the baryonic properties. If this distortion is closer to the distribution of the original training data, this implies that the neural network performs better without this shuffle group, and that one or several of its parameters are negatively influencing the model. On the contrary, if the fiducial prediction is closer to the original data than the disrupted prediction, it implies that the data contained in the shuffle group is critical to accurate prediction of this output quantity.

To assess the importance of the shuffle groups in the accuracy of predictions, we compare distributions of the median predictions of each network with those of the TNG data. Specifically, we compare the distributions of calculated stellar mass, mass-weighted metallicity and mass-weighted age for central and satellite galaxies. The underpredicted scatter in mass and metallicity results in subtly narrower and offset predicted distributions for all of these quantities, and the distortions of each disrupted network can indicate whether any of the input variables are responsible.

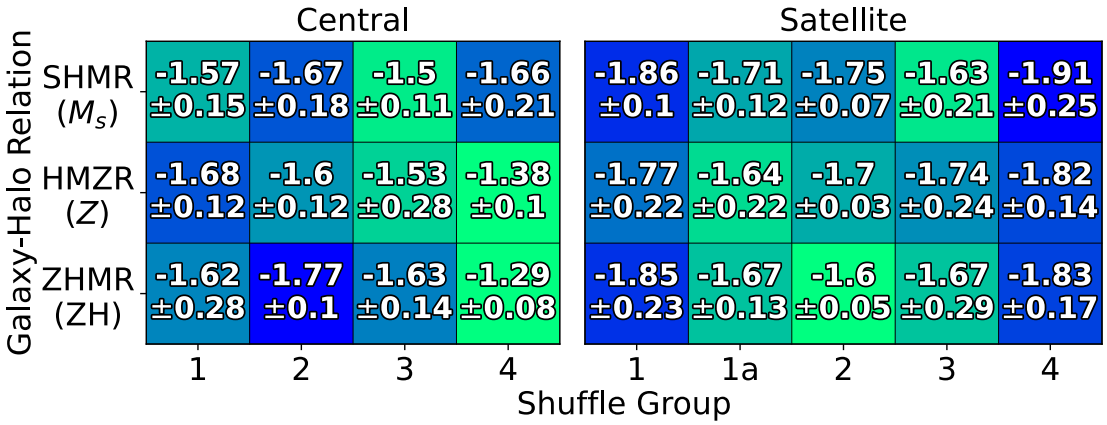


Figure 3.15: The Ξ values of the stellar-halo mass relation, mass-metallicity relation, and mass-metallicity history relation for each shuffle group are presented in the top, middle, and bottom rows, respectively; for both the central model (left) and satellite model (right). The median and interquartile range of Ξ values obtained from ten independent runs of each network are displayed in each cell's text. Illustrating the most critical shuffle groups, grid cells with smaller Ξ values are shaded in dark blue, transitioning to bright green as the median Ξ becomes larger. Higher Ξ values indicate that the given shuffle group is more important as the model determines the galaxy-halo relation, while a smaller interquartile range indicates greater significance of this result.

3.4.3 Results & Discussion

Scatter In Stellar Mass

Tabulated values of the median and interquartile range of Ξ for each shuffle group in the central and satellite datasets are shown in fig. 3.15, showing results for the stellar mass, mass-weighted metallicity and metallicity history as a function of halo mass for central galaxies and subhalo mass for satellites. Cells which are greener in colour have larger median Ξ, while smaller Ξ values are shown in blue.

For central galaxies, the Ξ values indicate that shuffle groups 1 and 3, entailing variables relating to mass accreton history and halo substructure, are the groups with the most significant effect on the SHMR. This can be explained by the tendency for the early forming structure of a halo to influence the initial growth of its galaxy. In particular, the frequency and size of merger and accretion events will have governed the mass, size and shape of the newly formed halo to different extents, driving the dynamics of star forming gas. As the final mass of galaxies is strongly connected to their mass at early times, this likely has significant effects on the trajectory of galaxy evolution. The continued growth of internal structure would continue to influence the galaxy over time, constraining their stellar mass. McGibbon & Khochfar (2022) show that measures such as circular velocity and velocity dispersion for $z < 2$ do indeed have such an effect on the stellar mass at

$z = 0$.

For satellite galaxies, the results are similar for the SHMR, implying that the mass accretion and internal structure are equally important for star formation over time. However, satellites have two key differences from centrals. First, shuffle group 1a, describing the history of the satellite subhalo, is considered important for the satellite SHMR, unlike shuffle group 1, describing the history of the central halo. Second, there is an additional importance of shuffle group 2, i.e. overdensity history, which can be reconciled with satellite quenching and other environmental effects on satellite star formation.

The satellite network therefore favours properties of the satellite subhalo and its surroundings over those of its host, yet it may prove in a future study that properties other than the host halo's mass history, such as its size and centre of mass relative to the subhalo, will offer additional constraints on the galaxy's satellite phase evolution. The fact that the halo's interior environment is influenced by the abundance of satellites (Bose et al., 2019) would suggest the use of assembly bias in satellite predictions.

Scatter In Stellar Metallicity

The Ξ values for metallicity and metallicity history show that shuffle group 4, relating to interaction history, is the group with the strongest influence by far. The skew parameter is believed to influence chemical enrichment by tracing the infall of subhalos which may be metal-rich or gas-rich, which in relation to the halo itself influences the metal content acquired by the galaxy during accretion or mergers.

The distinction between the influence of skew and overdensity, encompassed by shuffle groups 4 and 2 respectively, is that the skew is important for both the metallicity history and mass-weighted metallicity, whereas overdensity is important only for the latter. It is likely that the overdensity is measuring the most major interaction events by tracing the concentration of mass in the vicinity of the target halo, which are expectedly more frequent in denser regions (L'Huillier et al., 2015). The skew parameter is independent of the mass of the surrounding subhalos, and measures the concentration of subhalos around the target halo over time, regardless of their mass. As minor interaction events are more frequent, it is feasible that these events are contributing significantly to the galaxy's metallicity, as well as secondary, smaller chemical enrichment effects which show in the unweighted

metallicity history.

In the satellite data, the influence of overdensity is more significant, while skews are no longer significant. This can be reconciled with a tight correlation between the gas phase metallicity of satellites with their local densities, potentially owing to the satellites enriching their circumgalactic medium (Peng & Maiolino, 2013; Genel, 2016). The effect may be likened to environmental quenching, restricting the continued in situ synthesis of metals, which is not an effect which is common to central galaxies. On the contrary, these density-driven quenching effects are strongest for low mass satellites (Bluck et al., 2020), and their metallicities and star formation rates may be influenced by their mass.

High mass galaxies are in general quenched by AGN feedback. The mass and growth rates of AGN are tightly correlated with circular velocity and velocity dispersion (Davies et al., 2019; Bluck et al., 2020), and in fact the third shuffle group, relating to internal dynamics, has a great effect on stellar mass and a modest effect on metallicity in both datasets.

We argue that it is the circular velocity and not the half-mass radius in this shuffle group which affects the predictions of metallicity. While the physical size of the halo has been shown to influence quantities such as gas and black hole mass, which can influence the growth of a galaxy in future (Lovell et al., 2021), these quantites can also be determined by halo mass and environmental properties. As well as influencing the scatter in metallicity, the circular velocity can distort the distribution of metallicities on sample scales; something which no other quantity does. It is unclear whether this effect improves or degrades the quality of predictions, but it does suggest that the circular velocity has an effect which is noticeably independent of halo mass.

The under-prediction of high metallicity objects is decreased and the MZR is fit more accurately if the network is run using the baryon-sensitive maximum circular velocity instead of our virialised proxy. Our dark matter variables therefore cannot fully account for the metal enrichment of TNG galaxies, which also depend on purely baryonic phenomena, and the metallicity scatter in a pure dark matter simulation will be undermined by our model. If our networks were trained on a semi-analytic model simulation, or a hydrodynamical model with an alternative prescription of chemical enrichment which the network can more accurately identify, and this model provides appropriate metallicities despite the

dark matter controlled rotation curve in such a simulation, then this may justify the use of an alternative model in refining future predictions.

To summarise these two subsections on distorted scatter, it is the shuffle groups which relate to the mass accretion history of the target subhalo and the internal structure as a function of time which have the greatest influence on the scatter in stellar mass, independently of halo mass. This is most likely a measure of the quantity of accreted star-forming gas and the conditions within the galaxy to promote star formation. Environmental quantities control the scatter in metallicity at fixed halo mass: in the form of interaction rates for central galaxies and local mass density for satellite galaxies, illustrating the importance of subhalo infall and flybys, and the concentration of massive, metal-rich subhalos, respectively.

It should be stressed, however, that no single shuffle group or input quantity has made a unique effect on the baryonic outputs, and that some mechanisms in which these variables play a role can, in principle, be inferred from other model parameters; e.g. the mass-concentration relation of halos is strongly affected by the local environment and location of halos in the cosmic web (Hellwing et al., 2021). Nevertheless, the presence of shuffle groups with significant Ξ values in fig. 3.15 shows that the neural network models rely on the variables of these shuffle groups to constrain baryonic predictions, and that the input quantities which are members of said shuffle groups are considered to play a physical role in the galaxy-halo connection.

Subgroup Quantities

As described above, shuffle groups are arranged in accordance with physically related quantities, such as mass accretion history and final halo mass. While they are physically related, their relationship is not necessarily straightforward, and so the influence of a shuffle group on the network's predictions may be dominated by just one or few of its member parameters. We show this with two examples in fig. 3.16, where we scramble one input variable before training the network, do this for the remainder of the shuffle group and compare their Ξ values, illustrating the importance of these parameters in the shuffle group. As before, we conduct ten independent runs per shuffle and characterise their significance by the median and interquartile range of the Ξ distortions.

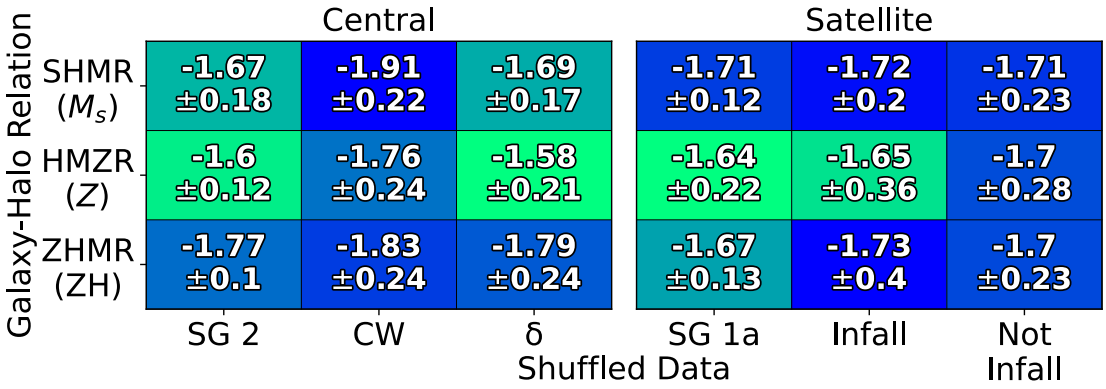


Figure 3.16: The tables presented in this section are similar to fig. 3.15, but instead highlight the effects of shuffling particular subsets within a shuffle group. Each table has three columns: the left column displays the results for the entire shuffle group, as shown in fig. 3.15. The center column shows the results for the input parameters being analysed, and the right column displays the results for the remaining components of the shuffle group. These tables indicate that overdensity is the primary factor in the second shuffle group for central galaxies, while the infall parameters in group 1a have a discernible impact on the satellite galaxy model.

The mass-independent scatter of star formation histories and the amount of metals present in gas within the TNG model are influenced by gas inflows which depend significantly on the galaxy's position in the cosmic web (Torrey et al., 2019; Hellwing et al., 2021; Van Loon et al., 2021; Donnan et al., 2022). In the central model, we randomised the distances between points in the cosmic web and compared the predicted results with those obtained by scrambling overdensities. The Ξ values from this test are shown in the left panel of fig. 3.16.

Although the there is a subtle correlation not shown by these tables between cosmic web distances and high-mass galaxies, the overall conclusion is that the cosmic web has little impact on the metal content of stars. In contrast, randomising the overdensity components has had a more noticeable effect on the metallicity scatter; in fact the Ξ values for the full shuffle group 2 are close to those from overdensity alone, while the cosmic web produces very small deviations in scatter. Therefore, the machine learning model for central galaxies appears to give more weight to overdensity components than to the cosmic web when predicting the mass weighted metallicity of central galaxies.

The lack of influence of the cosmic web on our predictions may be physical or simply numerical. On the one hand, the lack of sampling of halos of a given mass or the cosmic web features defined by the DisPerSE algorithm may affect their importance in the network. On the other hand, the size of the filaments or the number of small filaments can have effects which either enhance or suppress star formation and chemical enrichment, depending very

specifically on the scale and geometry of the local cosmic web (Galárraga-Espinosa et al., 2023). It may prove useful in a future study to include further properties of the local cosmic web in relation to the target halo in order to truly characterise its influence, however, these effects are also correlated with halo mass and merger activity (Hellwing et al., 2021), which is measured using environmental quantities such as overdensity and skew. These variables include temporal information, and therefore may be a better constraint on the structure growth history which the zero-redshift cosmic web distances effectively generalise.

In addition to assessing the influence of the cosmic web in the central galaxy model, we conducted experiments on the satellite network to examine the impact of infall parameters; namely the scaled infall time, scaled formation time and infall mass ratio. While these factors may be significant for satellite star formation histories, the right-hand panel of fig. 3.16 shows that the difference in Ξ values between these infall parameters and the rest of the data in shuffle group 1a is minimal, indicating that they have little effect on the performance of the satellite network. This could be due to the target's transition from a central to a satellite subhalo being inferred from the growth histories of the subhalo and its host.

Despite the apparent insignificance of the infall parameters, the deviation from scrambling infall parameters is more significant, suggesting that explicitly utilising these parameters can be useful. Additionally, while the metallicity deviation associated with infall-only is poorly constrained, it is larger than other factors, which may indicate some significance of infall quantities. In fact we stress in section 3.2.4 that the quality of prediction of the star formation histories of satellite galaxies is marginally superior to that of central galaxies, probably due to the inclusion of infall parameters constraining their quenching timescale.

Distorted Distributions

By evaluating Ξ, we have evaluated the effect of removing predictive information on the output of the network relative to the fiducial result, yet this does not indicate whether it is closer to, or further from, the original simulation data, and therefore a more or less physically accurate result. In fig. 3.17, we compare the distributions of mass, metallicity, and mass-weighted metallicity in an intermediate halo mass range for each network and each shuffle group, alongside the median of ten fiducial predictions and the original simulation

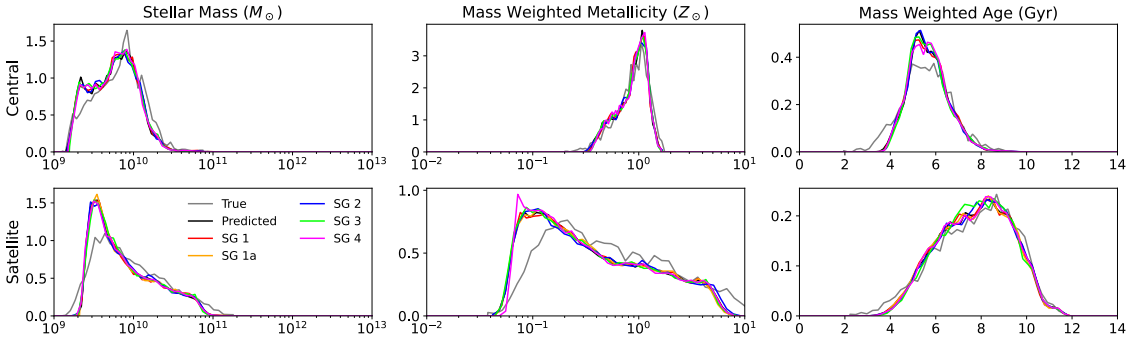


Figure 3.17: Smoothed probability density functions of the baryonic data in an intermediate halo/subhalo mass bin; for stellar mass, metallicity, and mass-weighted age; based on the predicted star formation and metallicity histories. The distributions are obtained from the original TNG data (grey), the median of ten standard predictions (black), and the median of each randomisation (coloured lines). The distributions resulting from each randomisation do not exhibit a clear improvement over the fiducial data, nor do they bring the network any closer to the true distributions from the TNG simulations. If the horizontal axis is logarithmic, the PDF is a function of the logarithmic value labelled on the figure.

data. As established in section 3.2.2, we observe that the network somewhat underestimates the scatter in stellar masses and metallicities, resulting in a discrepancy between the true distributions and our predictions. Specifically, our predictions yield histograms of stellar mass which are slightly narrower than those from TNG, and metallicity histograms which are offset from the desired result. If any shuffle group presents a distribution which is closer to the true distribution, it may indicate that a member of this shuffle group is misleading the network, whereas larger differences may suggest that the group contains essential information for predicting accurate distributions.

The clear similarity between the distributions in fig. 3.17 illustrates our conclusion: we have not discovered any evidence of a shuffle group which enhances or derails the network's performance when randomised, as indicated by our evaluation of a Ξ analogue for median data values. However, we did observe a slight variation in the shape of satellite MWA distributions when shuffle group 3 was scrambled. While the scatter ratio shows the dependence of other shuffle groups in different halo mass regimes, our analysis of data in a narrow halo mass range reveals that shuffle group 3 is useful for distinguishing between similar samples. Nonetheless, it is unclear whether this dissimilarity deforms the predictions, indicating that the shuffle group contains an essential detail of the galaxy-halo connection, or whether these features are correlated with the TNG distributions, which would imply that shuffle group 3 misleads the satellite network.

This difference between true and predicted data, regardless of any scrambled data,

implies that the inaccuracies of the model cannot be rectified by the exclusion of any of the existing input parameters. Including alternative prescriptions of the galaxy-halo connection, such as including temporal measures of merger history or progenitor data, may result in a closer match to the original data. It may prove, however, that these results represent the limit of what can be derived concerning the baryonic evolution of galaxies from dark matter alone, and improvements may constitute numerical methods such as the stochastic corrections discussed in section 3.3, which have shown to accurately model the variability of SFH and ZH which was missing from the fiducial model, can be used to improve the fit to the original GHC. Alternatively, the discrepancies between simulations and predictions may be a numerical artifact of the model, and perhaps an alternative machine learning design will improve upon these predictions.

3.5 Summary

In this chapter, we have applied the neural networks described in detail in chapter 2 to the hydrodynamical TNG simulations, comparing the physical properties of the output star formation and metallicity histories of the trained model with the original simulation data. We additionally apply a stochastic model to improve the quality of predicted galaxy properties, and evaluate the importance of different input variables by retraining the networks with scrambled subsets of the input data. This has led to the following conclusions:

1. Accurate predictions can be made for the full range of star formation histories in both central and satellite galaxy datasets. In section 3.2.1, it is demonstrated that the geometries of the predicted star formation histories match those of the original simulation, from continually growing star-forming galaxies to quenched satellites and high-mass galaxies. The integrated stellar mass derived from these star formation histories correlates well with their values in the TNG data, as does the stellar mass taken from their merger trees. It is shown by fig. 3.3 that our predicted star formation histories fully recover the SHMR.

2. Although the network model has yielded positive results, it has limitations in predicting star formation events that occur rapidly, such as star formation bursts or rapid quenching. The lack of high-frequency information is reflected by the predicted features of star formation and metallicity histories such as those shown in fig. 3.1, and

the declining Fourier transforms of these predictions in fig. 3.2. The absence of these features can lead to numerical stellar mass estimates which are systematically lower than their actual value, and impacts the scatter of the MZR shown in fig. 3.7. This is a more prevalent shortcoming for central galaxies, where such features are more frequent.

3. As the network is capable of predicting the Fourier amplitudes of these historical properties, as shown by fig. 3.9, these predicted Fourier transforms were used to construct a corrective stochastic component to add to the predicted star formation and metallicity histories, which is discussed in section 3.3. This has improved estimates of the stellar mass of most samples, and recovered the previously compromised scatter in the MZR. However, this correction does not recover the full star formation and metallicity histories, still failing to produce short-timescale events containing a high rate of star formation. The quality of the modified results depends on the size of the missing features, which depends on the halo mass regime, and therefore the correction may be improved if the frequency range to which it is applied is optimal.

4. In section 3.4 we assess the importance of the input variables of the neural networks in replicating key galaxy-halo statistics, by grouping variables according to physical relationships with one another, and retraining and testing the neural networks where these groups are replaced with random Gaussian noise, effectively removing their signal from the model. This has shown that for both central and satellite galaxies, variables relating to the halo's mass and substructure, such as mass accretion history and half-mass radius, are most important to predicting the SHMR, i.e. the star formation histories; while environmental quantities, such as skewness and overdensity, are correlated with metallicity histories and the MZR. Previous studies concerning a single redshift have struggled to measure the effects of these variables on galaxy properties, demonstrating the value of historical input data.

5. While the deviations in scatter which manifest from scrambling input data highlight the data which influences certain results, these deviations are usually small. The summary statistics, particularly the SHMR, can be inferred by the neural network via the connection between scrambled and non-scrambled datasets. When we scramble subsets of a single group, however, we find that an individual quantity, such as infall

parameters or overdensities, can dominate the influence of the given group. These results show that the model may be used to identify the most significant parameters in the galaxy-halo connection, which will be supported by much larger datasets when applied to high fidelity N-body simulations.

The direct predictions of the neural network indicate that it is possible to develop a predictive machine learning model which models the full evolutionary history of galaxies using historical halo and environmental data, and can be used in future studies to investigate the temporal nature of the galaxy-halo connection for galaxies of assorted properties and evolutionary regimes. However, the predictions of the model are subject to inaccuracies, which require a robust correction to produce accurate statistics and observables. Plus, the low sample size makes it difficult to thoroughly investigate the various regimes of halo-galaxy coevolution with TNG predictions alone.

The practical limitations of the model as it stands are discussed further in chapter 4, where we discuss how the model may complement observational studies of the galaxy-halo connection; and in chapter 5, where we evaluate the effects of applying the model to pure dark matter simulations, with a lower mass resolution and alternatively defined halo properties.

This chapter is based on the observational results of Chittenden & Tojeiro (2022) and Behera, Chittenden, & Tojeiro (in prep.).

4

Observables

4.1 Introduction

In observations of galaxies, the spectroscopic and photometric features of a galaxy, i.e. the luminous flux received from the galaxy decomposed as a wavelength spectrum and in different wavelength bands, trace the properties of the stellar, gas and dust components of the galaxy, and can inform a plethora of information relating to its evolutionary past (Tinsley, 1980; Kennicutt & Evans, 2012; Sánchez Almeida et al., 2012). The luminosity of a galaxy, for instance, owes to the sum of luminosities from all of its stars, which may be used to derive the galaxy's total stellar mass and current star formation rate (Curtis-Lake et al., 2012; Johnson et al., 2013b). The difference in luminosities between wavelength bands, interpreted as colour, can infer the abundance of cool, red stars and hot, blue stars (Liao & Cooper, 2022; Whitler et al., 2023), as well as the abundance of gas and dust (Berta et al., 2016; Wang et al., 2017). Ionising radiation which permeates the gas reservoir also creates emission lines in the galaxy's spectrum, whose relative fluxes can

inform the metallicity of the gas (Kobulnicky & Phillips, 2003; Mouhcine et al., 2005; Nagao et al., 2011), while the width of these lines can be used to trace the kinematics of the gas (Maseda et al., 2014; Kewley et al., 2019). These are all observational features of galaxies which can be combined to construct a more complete picture of the galaxy's evolutionary history; for example, a bright, blue galaxy with high velocity dispersion likely underwent a merger which accelerated its star formation.

Mock surveys are a powerful tool with which to test theoretical models of galaxy evolution on cosmological scales, providing a simulated dataset of galaxies of assorted masses, colours and other properties at different times in the universe's history (for examples see Kauffmann et al., 2020; Ferrero et al., 2021; Snyder et al., 2022; More et al., 2023). As described in section 1.2.1, these galaxy evolution models may be based on empirical relations between galaxies and N-body simulations of large scale structure, or by the propagation of a set of equations of motion governing the coevolution of galaxies and halos. The properties of the simulated galaxy population can be used to comprehensively investigate the prevalence of the physical processes which influence them, and by comparing these galaxies with those observed in real galaxy surveys, astronomers can test the validity of different theoretical models and gain insights into the physical processes which drive galaxy evolution. The advent of larger, more sensitive surveys (proposed or ongoing examples including Euclid Collaboration, 2013; DESI Collaboration, 2016; Dunlop et al., 2021; Malkan et al., 2021) has allowed for these galaxy statistics to be studied in greater detail, motivating the need for equally large, detailed mocks to complement them, which is one of the key motivations for the predictive model outlined in this thesis.

In order to produce accurate mocks, the machine learning model must predict the baryonic properties from which observational features are derived to an adequate level of precision. This is because the physical properties of galaxies influence these observables in various ways. Galaxies which formed most of their mass at early times, for instance, will have a greater population of cold stars with long lifespans such as red giant and red dwarf stars, and thus the galaxy is more likely to appear redder in colour and have stronger absorption features. Galaxies with ongoing star formation, on the contrary, will have an abundance of young stars, including hot, short-lived, UV-luminous stars emitting ionising radiation, such that the galaxy appears bluer in colour and exhibits strong emission lines, where the radiation from these UV-luminous stars has ionised the surrounding gas. With

thorough modelling of the spectroscopic and photometric signatures of galaxy evolution, one can produce a catalogue of spectra, magnitudes, colours and emission lines from an existing catalogue of star formation and metallicity histories, which can be used to compare the observational statistics of the neural networks with the hydrodynamical simulations.

The Flexible Stellar Population Synthesis (FSPS) code (Conroy et al., 2009; Conroy & Gunn, 2010) is a software which generates a spectral energy distribution from a stellar population by encoding empirical parameterisations of the distribution of stellar masses, abundance of ionising radiation and other features which determine the spectral properties of the population, which has been used in multiple studies to investigate the influence of galaxy evolution on its observational features (Chaves-Montero & Hearin, 2021; Pan et al., 2023; Pfeffer et al., 2023). In this chapter, we discuss how we use FSPS to emulate spectroscopic and photometric data, from the star formation and metallicity histories from TNG, and from the predictions of the artificial neural network models, which we use to assess the model's capability of producing data for mock surveys. Having discussed various successes and shortcomings of the predictive power of the neural networks in chapter 3, we discuss here the consequent effects on synthetic spectra and photometry, and the benefits of improving the model for observational purposes. This chapter entails the methods of constructing and manipulating zero-redshift spectra in section 4.2, comparing observables between original, predicted and stochastically modified data in section 4.3, and summarising our findings in section 4.4.

4.2 Calculated Observables

As has been established in section 4.1, the emission spectrum of a galaxy encodes plenty of information relating to its evolutionary properties, and so simulated galaxy properties can be used to produce model spectra. We evaluate such spectral energy distributions from the predicted SFHs and ZHs of our neural network, using the Python FSPS wrapper for the FSPS Fortran code (Johnson et al., 2013a), and compare these spectra with those of the original TNG simulations. We also show improvements to these results from the stochastic corrections outlined in section 3.3.

4.2.1 Spectral Energy Distributions

In the FSPS code, a galaxy spectrum is constructed as a composite stellar population (CSP), in which the histories of star formation and chemical enrichment quantify a weighted sum of a set of basic spectra corresponding to stars of identical age and metallicity, known as a simple stellar population (SSP) (Conroy, 2013). This spectrum can then be used to calculate photometric magnitudes, emission line luminosities and other observational quantities used in galaxy surveys to evaluate the physical and evolutionary properties of galaxies.

The SSP modelling in FSPS consists of a summation of stellar spectra of different masses, weighted according to an initial mass function (IMF). These stellar spectra are considered at a fixed isochrone: a population along the Hertzsprung-Russell diagram of controlled age and metallicity, whose mass, surface temperature and self-gravity are tightly related (Tinsley, 1980). The core ingredients of a SSP spectral synthesis model are therefore the choice of IMF, isochrones and stellar spectra as a function of mass (Conroy, 2013). Specifically, the SSP spectrum of a given isochrone is given by integrating mass over the given isochrone:

$$f(\lambda;\ \mathcal{Z},\ t) = \int_{M_{\min}}^{M_{\max}} f_*(\lambda;\ M,\ \mathcal{Z},\ t)\ \Phi(M)\ dM \tag{4.1}$$

where Φ is the IMF of the zero-age main sequence population of stars, f_* is the stellar spectrum of stars of a certain mass and isochrone, and $M_{\min}$ and $M_{\max}$ are the lowest and highest star masses along the stellar evolutionary tracks which define the isochrones; usually ranging from the hydrogen burning limit of $\sim 0.1 M_\odot$ to a maximum of $\sim 100 M_\odot$ (Conroy, 2013).

Adding SSP spectra from successive stellar ages while parameterising each spectrum according to metallicity and star formation rate is typically used to construct a CSP: a more complex ensemble of stars of varying ages, temperatures, chemistries and luminosities; ranging from stellar clusters, to galactic disk and bulge components, to the full evolution of galaxies.

For each time step in our SFHs and ZHs, we emulate a SSP spectrum assuming an IMF corresponding to the Chabrier (2003) model, the MILES spectral library (Falcón-Barroso

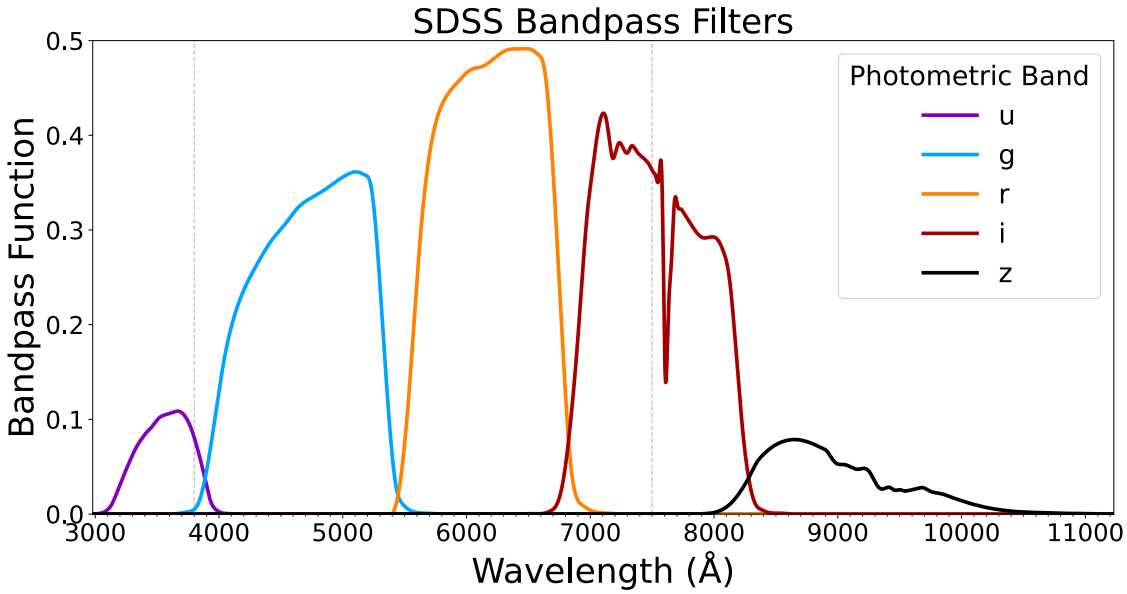


Figure 4.1: The response functions of the five optical filters used in the SDSS project. These are the weighting functions used to compute band fluxes from the spectra computed from our star formation histories. These filters are adjusted for atmospheric transmission with a typical airmass of 1.2 (Fukugita et al., 1996). The grey, vertical lines indicate the approximate range of optical wavelengths, showing that some of these these filters are sensitive to ultraviolet and infrared wavelengths. Each line is coloured according to the approximate perceived monochromatic colour of the band's effective wavelength, with the exception of the z band, which has no optical wavelengths.

et al., 2011) and MIST isochrone model (Choi et al., 2016). The SSPs are parameterised by the current age and metallicity of the galaxy, and weighted according to the total stellar mass formed in the time interval between the present and previous time steps (see eq. (2.26)); which produces a full CSP spectrum, which we associate with the galaxy at $z = 0$. Given a set of SSP spectra f_j, defined in eq. (4.1), the full spectrum $\mathcal{F}$ is calculated as follows:

$$\mathcal{F}(\lambda) = \sum_{j=1}^{N_{\text{snap}}} \mathcal{M}_j \; f_j(\lambda;\; \mathcal{Z}_j,\; t_j) \tag{4.2}$$

These CSP spectra are parameterised by the complete star formation and metallicity histories of their galaxies, and therefore this calculation is a self-consistent method with which to test the suitability of the predicted galaxy formation histories for accurate mock spectroscopy. Furthermore, the following observational quantities are calculated from these SEDs, illustrating the effects of the quality of the network predictions on photometry.

Band	Region	Range (Å)	$\lambda_{\text{effective}}$ (Å)
u	UV	2980 - 4130	3551
g	Green	3630 - 5830	4686
r	Red	5390 - 7230	6166
i	Near-IR	6430 - 8630	7480
z	IR	7750 - 11230	8932

Table 4.1: Characteristics of the five SDSS bands for which we calculate photometric magnitudes from our galaxy spectra (Fukugita et al., 1996; SDSS Collaboration, 2002).

4.2.2 Photometric Luminosity And Colour

Using the SEDs whose derivation is outlined above, we calculate the magnitudes in the photometric bands used by the SDSS Collaboration (2002). We mimic the flux passing through the five bandpass filters used in the survey (Fukugita et al., 1996), by integrating the spectrum over the response functions shown in fig. 4.1. Specifically, if a given band has a dimensionless filter function $\eta(\lambda)$, the band flux density F from the spectrum $\mathcal{F}(\lambda)$ is defined:

$$F = \frac{1}{4\pi D_L^2} \int_0^\infty \eta(\lambda)\ \mathcal{F}(\lambda)\ d\lambda \tag{4.3}$$

where we assume a luminosity distance $D_L = 10\text{pc}$, abiding by the definition of absolute magnitude. The absolute band magnitude m is defined:

$$m = -\frac{5}{2} \log_{10} F \tag{4.4}$$

where F is converted to the *maggie* unit of flux density, which calibrates to the AB system filters in SDSS. The key properties of these five band filters are shown in table 4.1.

Using these derived magnitudes, we evaluate differences between magnitudes in separate bands to compute colours, and compute galaxy colour-mass diagrams. The distribution of colours from several combinations of bands is bimodal, where "blue" galaxies have ongoing star formation, "red" galaxies are quenched, and galaxies in the transition phase from blue to red are members of the "green valley" (Nelson et al., 2017). A predictive model which recovers this property therefore recognises the photometric distinction between star-forming and quiescent galaxies.

4.2.3 Emission Line Luminosity

A feature of the stellar libraries used to define SSP spectra is the empirical calculation of emission line luminosities, which can be achieved by correlating the density of ionising photons with galaxy properties such as stellar mass and star formation rate (Baugh et al., 2021). In the FSPS code, nebular line and continuum emission are implemented by using template SSP spectra to ionise the gas clouds whose mass and metallicity are empirically inferred. The resulting nebular emission is added to the SSP spectra themselves (Byler et al., 2017).

The Python FSPS library includes a function which returns various line luminosities of atomic and ionised transitions for each SSP spectrum. To assess the quality of emission lines derived from our predictions, we calculate the total Hα luminosity of each galaxy, typically the strongest emission line, by evaluating the sum of SSP luminosities weighted by their star formation history, in the same manner as the full spectrum:

$$L_{\rm gal}^{\rm H\alpha} = \sum_{j=1}^{N_{\rm snap}} \mathcal{M}_j \; L_j^{\rm H\alpha}(\mathcal{M}_j,\; \mathcal{Z}_j,\; t_j) \tag{4.5}$$

4.3 Results

4.3.1 Spectral Energy Distributions

In fig. 4.2 we display the mean and standard deviation of galaxy spectra for both central and satellite galaxies in bins of stellar mass. The shape and amplitude of the spectra in most bins are similar in both the original and predicted galaxies. In high mass galaxies, however, the networks underpredict the mean and variance in the high frequency luminosity, and at low masses, the amplitude of the spectra themselves are slightly underpredicted. This inaccuracy is somewhat more prevalent for central galaxies as satellite galaxies tend to have more well-defined, smooth star formation histories due to satellite quenching.

The mass-to-light ratio of an SSP is correlated with its metallicity and stellar age (Gallazzi & Bell, 2009), explaining inaccurate luminosities in samples of underpredicted variance in SFH and ZH. However, the stochastic corrections make a modest improvement to the spectra, particularly for intermediate mass central galaxies, and most satellites.

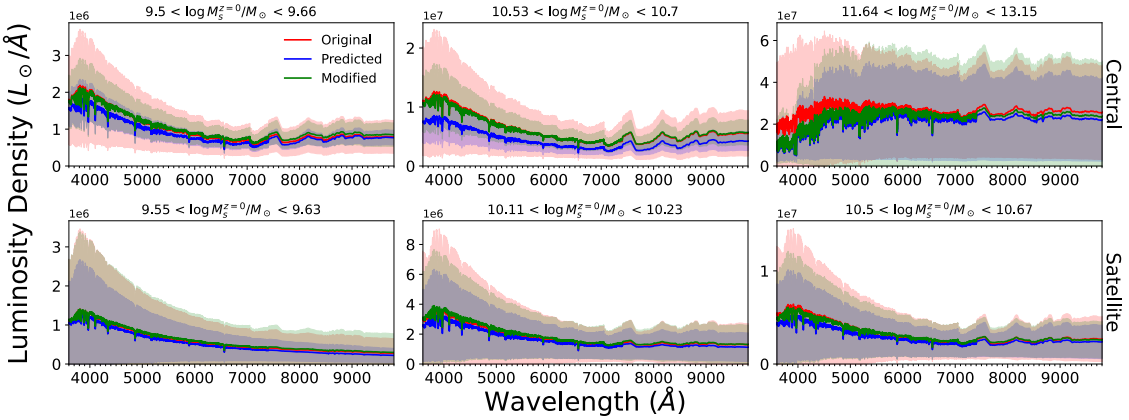


Figure 4.2: The mean and standard deviation for stacked central (top row) and satellite (bottom row) spectra in bins of stellar mass, shown for predicted star formation and metallicity histories in green, and TNG data in blue. Emission lines have been omitted from these plots for clarity. In the majority of samples, the continuum is generally well recovered, and is of similar amplitude. However, for high mass objects there is a reduced variance at short wavelengths, and lower mass galaxies have a smaller variance overall. This represents a poorer prediction of central galaxy spectra, with lower mean amplitudes and smaller variance than the spectra evaluated from TNG data.

The increase in the mean amplitude of the spectra can be likened to the increase in stellar mass providing a closer match to the original SHMR, and the increased variance in SEDs coinciding with increased scatter in stellar mass and metallicity.

The use of the stochastic correction method to rectify the total luminosity is motivated by fig. 4.3, which shows the correlation between the fractional errors in high frequency Fourier modes of central and satellite star formation histories, evaluated as the average amplitude of the absolute Fourier transform above a frequency of $\sim 0.3\text{Gyr}^{-1}$; and the total luminosity of their spectra, evaluated by integrating the SED over all wavelengths. For galaxies of high star formation rate at $z = 0$, there is a clear correlation between these residuals, which shows that a star formation history with a stronger power spectrum at high frequencies provides necessary constraints on the spectrum of an individual galaxy. When the stochastic correction is applied, the Spearman correlation between these residuals with respect to the original TNG data is reduced from 0.527 to 0.159 for central galaxies and 0.614 to 0.294 for satellites, while the luminosity residuals themselves are reduced by up to an order of magnitude, signifying the improvement made.

Despite the success of the stochastic amendment in improving the accuracy of the SEDs, where the median amplitudes of the spectra in low and intermediate mass bins are closer to the original result after modification, fig. 4.2 shows that the correction does not fully recover the variance in spectra of the original TNG data, particularly for central

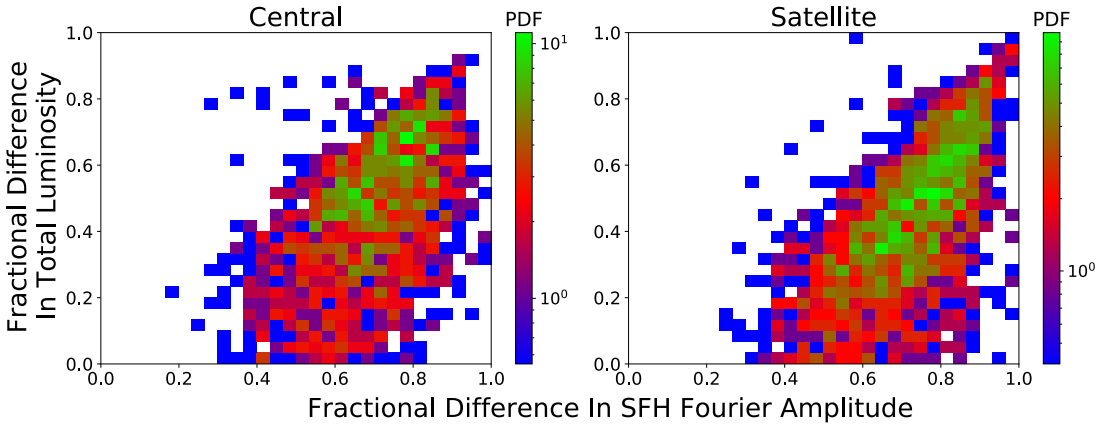


Figure 4.3: 2D histograms of the fractional difference between true and predicted total galaxy luminosity and the mean high-frequency Fourier amplitudes of their star formation histories. These are shown for galaxies between the 75^{th} and 95^{th} percentiles of $z = 0$ star formation rate, and shows data within a frequency range of 0.3-1.2 Gyr^{-1}, i.e. a timescale range of 0.8-3.3 Gyr. This correlation between the two residuals indicates the dependence of the calculated luminosity on high-frequency star formation events.

galaxies. This can be likened to the incomplete star formation histories even after modification, which were shown in section 3.3.3. The remaining correlation between residuals signifies that the remaining features will introduce further correction to these observables.

The modified star formation histories, despite showing similar fluctuations to the original TNG data, do not contain fluctuations which include a large quantity of stellar mass, contributing significantly to the total luminosity. As discussed in section 3.3.4, the lack of high amplitude features in the modified SFH may owe to the features of the Fourier transforms which can be predicted by the neural network, or the frequencies of the correction required for different samples, and may be improved by implementing variables which affect galaxy evolution on different timescales.

4.3.2 Hα Line Luminosity

In fig. 4.4, we show that the general trend of the total Hα line luminosity with stellar mass is recovered by the network for both central and satellite galaxies, however the variance in this line luminosity is small by comparison with the original data. However, the range of Hα luminosities is visibly increased by the stochastic amendment, closely matching some of the contours of the TNG data.

Unlike the total luminosity of the galaxy, Balmer line luminosities are particularly dependent on SSPs of a low age and metallicity (Bruzual & Charlot, 2003; Byler et al., 2017); therefore it would be practical to constrain the star formation history at late times

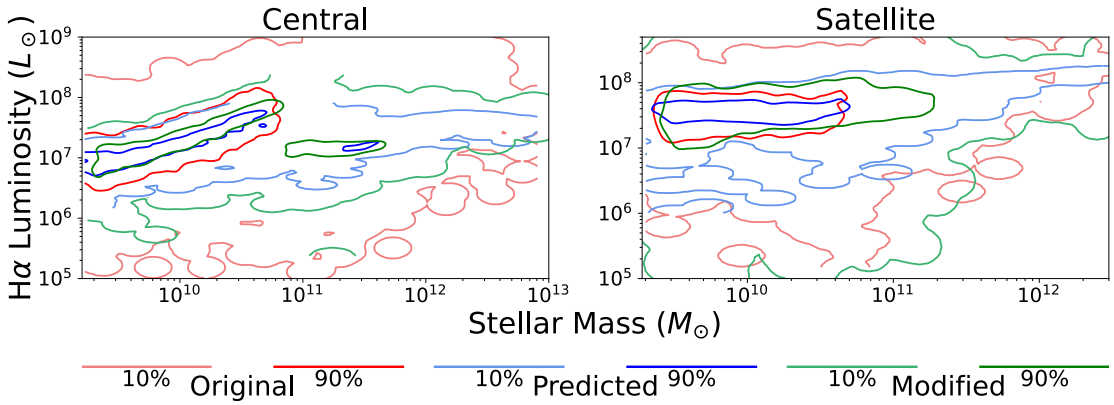


Figure 4.4: Distribution of positive Hα luminosities evaluated from the original and predicted spectra, shown in relation to stellar mass, with contour lines indicating the tenth and ninetieth percentiles of the distribution of data, indicated by the legend at the bottom of the figure. Only galaxies with positive Hα fluxes are shown. Predictions show a modest fit to the mass-Hα luminosity relations for central and satellite galaxies, yet the scatter in both is underpredicted due to missing star formation points.

to gain better predictions of the Hα luminosity. This late star formation rate is not strictly constrained by the neural network, nor is it explicitly modelled by the stochastic correction. Predictions of emission line luminosities would therefore be improved if the network were trained to model this late star formation more explicitly, either by weighting the star formation history at late times, or rebinning the star formation histories to include more samples of young stars, if the objective were to model ongoing photoionisation more accurately.

Despite the recent star formation not being modelled by the stochastic correction, fig. 4.5 shows that, similarly to fig. 4.3, residuals in high frequency Fourier modes are correlated with residuals in this line luminosity. The Spearman coefficient between residuals as described in section 4.3.1 is reduced from 0.233 to 0.112 for centrals, and 0.189 to 0.055 for satellites; with the residuals in Hα luminosity reduced by a factor of 2.5 on average. This reflects the utility of the stochastic method in improving Hα predictions. Improved metallicity histories may also be used to improve predictions of metal emission lines, and despite the particular sensitivity of Hα to recent star formation, historical formation histories may be used to trace emission lines which are more valuable probes of high-redshift star formation (Stark et al., 2010; Suzuki et al., 2016; Lagache et al., 2018; Madden et al., 2020).

It should be stressed nonetheless that the stochastic amendment has both positive and negative consequences on the Hα data. On the one hand, the modelling of fluctuations

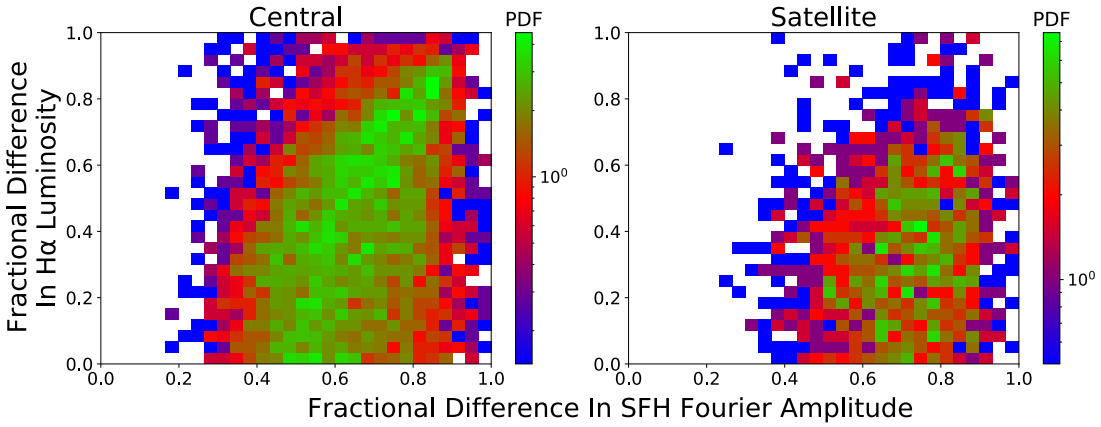


Figure 4.5: For the same galaxies as in fig. 4.3, this figure shows the correlations between residuals of their high frequency SFH data and their total Hα line luminosity. This indicates the importance of measuring short-timescale star formation events as in fig. 4.3, in particular at low stellar ages with the largest contribution of ionising photons.

contributing to line emission has rectified the lack of such emission in the predicted data, and reproduced the distribution of line luminosities, highlighting its utility in computing emission lines in future work. However, some predicted Fourier transforms will lack the necessary properties to accurately constrain the phases of such fluctuations, and this can introduce unwanted line emission, or lack thereof. Thus, the stochastic amendment is a practical tool for recovering line emission, but is subject to errors originating from the predicted Fourier spectra. A more robust method of stochastic modelling, potentially one which computes the phases of fluctuative signals, would therefore be of benefit to these predictions.

4.3.3 Band Magnitudes

Evaluating the SDSS band magnitudes using the formulae and filters in section 4.2.2, we show the dependence of absolute magnitudes in the five bands on the stellar mass of the galaxies in fig. 4.6. The challenge of predicting brightness at shorter wavelengths, as seen in fig. 4.2, causes noticeable deviations in the bluer photometric bands such as u or g. In the network-predicted data, the lack of scatter in stellar mass and metallicity, as well as the lack of variance in SEDs in different mass bins, is reflected in the smaller range in their contour diagrams.

Despite this, the adjustment to the spectra at these wavelengths through stochastic correction brings the contours of mass-magnitude distributions marginally closer to those of the TNG data, corresponding to samples with better estimates of stellar mass and

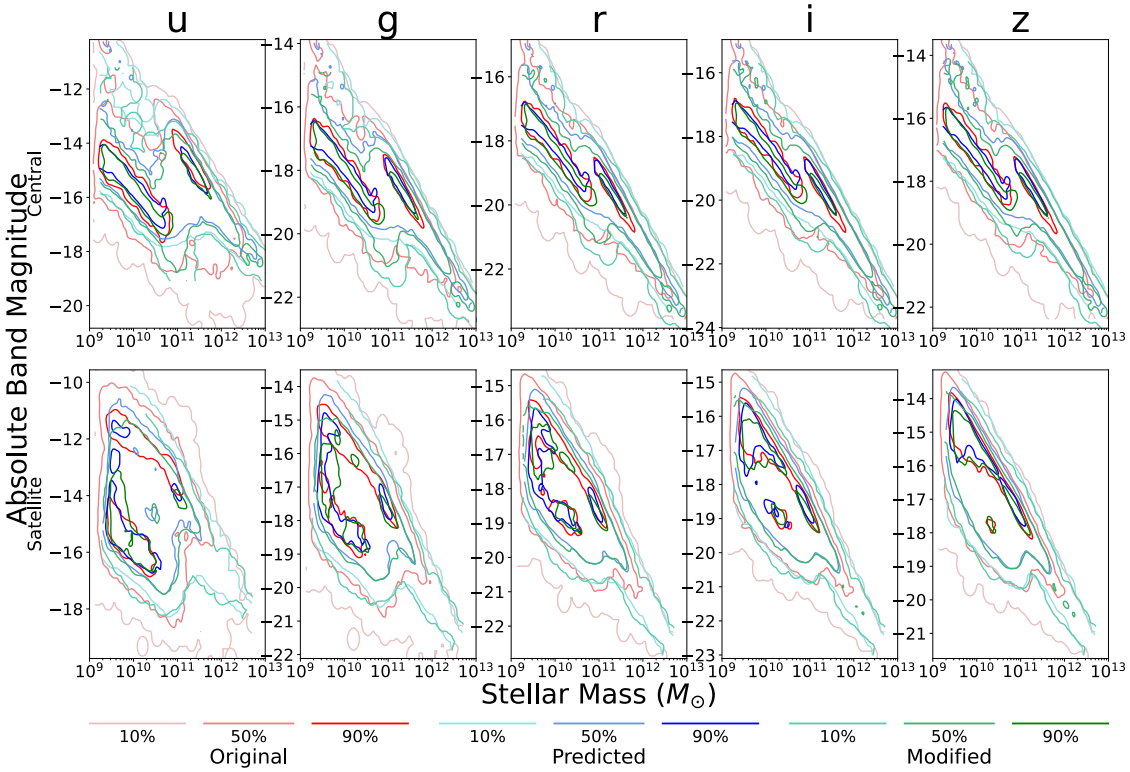


Figure 4.6: Estimates of the five SDSS band magnitudes from the true and predicted spectra of both central and satellite galaxies, shown as a function of stellar mass, with contour lines indicating the tenth, fiftieth and nineteenth percentiles of their 2D distribution. These show a reasonable similarity in all bands despite a slight reduction in the variance of magnitudes in the predicted data. In both central and satellite data, the bimodal distribution of magnitudes can be seen in relation to mass.

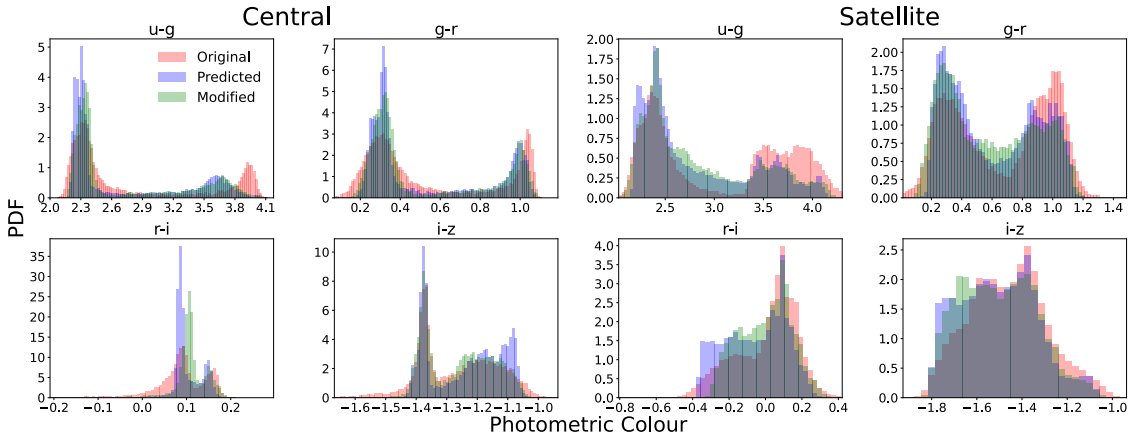


Figure 4.7: Photometric colour distributions across the five bands, showing the differences between two consecutive bands. The distributions, mostly bimodal, are in rough agreement between datasets, however there are clear offsets in some of the data, such as bluer red galaxies in $g-r$, and significantly smaller predicted ranges.

metallicity. However, the edges of the modified mass-magnitude distribution remain far from the TNG target, again emphasising that the stochastic method does not completely rectify the lack of variance in galaxy brightness, having a distinct lack of large fluctuations, and that more direct implementation of dark matter quantities which influence star formation events and their timescales more precisely may be necessary for accurate photometry in mocks.

4.3.4 Photometric Colours

We display the colour distributions of the three galaxy datasets in fig. 4.7, which exhibit the anticipated two-peak pattern in various band differences, for central and satellite galaxies. As a result, the network models differentiate between "blue" galaxies that are actively forming stars and "red" galaxies with quenched star formation. The tendency for high mass galaxies to be quiescent and low mass galaxies to be star-forming is shown by the colour-mass diagrams in fig. 4.8.

The SEDs in fig. 4.2 reveal the challenge of predicting luminosity at shorter wavelengths, which leads to systematic deviations in the bluer photometric bands in the network output data. This is evident in fig. 4.7 with colour distributions such as $u-g$ and $g-r$, where numerous red galaxies are shifted towards bluer colours. The stochastic correction makes a small improvement to the distribution of colours of central galaxies, such as increasing the "redness" of quiescent galaxies; but on the contrary, the distribution of satellite colours is distorted. However, the change in the colours with respect to the

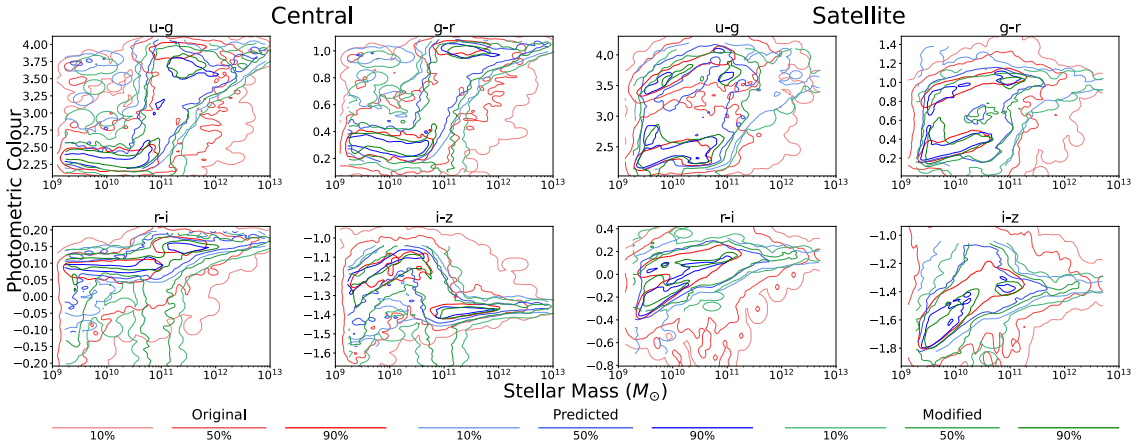


Figure 4.8: Colour-mass distributions of the three galaxy datasets, shown for central and satellite galaxies, and for four colours evaluated from neighbouring band magnitudes. This shows the distinction between low mass "blue" galaxies and high mass "red" galaxies in all datasets, showing that the network predicts this relationship. However, there is a smaller range of colours in the predicted data which is not noticeably improved by the stochastic amendment. Contour lines indicate the tenth, fiftieth and nineteenth percentiles of the 2D distribution of each dataset.

network's predictions are not substantial, and this is most likely due to the correction's lack of constraint on recent or high mass features, which can have a profound effect on these colours. As mentioned in section 3.3.3, the lack of reshaping of the SFHs and ZHs and small change in mass weighted ages per stellar mass bin can explain this small change in photometry.

In fig. 4.9, we illustrate that errors in the $g - r$ colour, like the total luminosity of the galaxy, are proportional to residual Fourier modes. However, unlike the errors in luminosity, this correlation is also evident for the Fourier modes in metallicity history. Deviations in absolute colour residuals are similar in magnitude to visible deformations of the $g - r$ distributions shown in fig. 4.7, which may be explained by the absence of short timescale events, resulting in narrower peaks of the distribution. Nonetheless, high-frequency metallicity history features are uncommon in metal-rich galaxies with under-estimated metallicities, and their colour error may be due to inaccurate star formation histories instead.

The stochastic correction has nonetheless failed to rectify the errors in photometry to the same extent as spectroscopy or physical properties of the galaxies. While the Spearman coefficients between SFH and colour residuals for centrals and satellites are reduced from 0.408/0.400 to 0.122/0.133 respectively, the changes to the residuals are not significant, indicating that the colours cannot be rectified using only this high frequency component

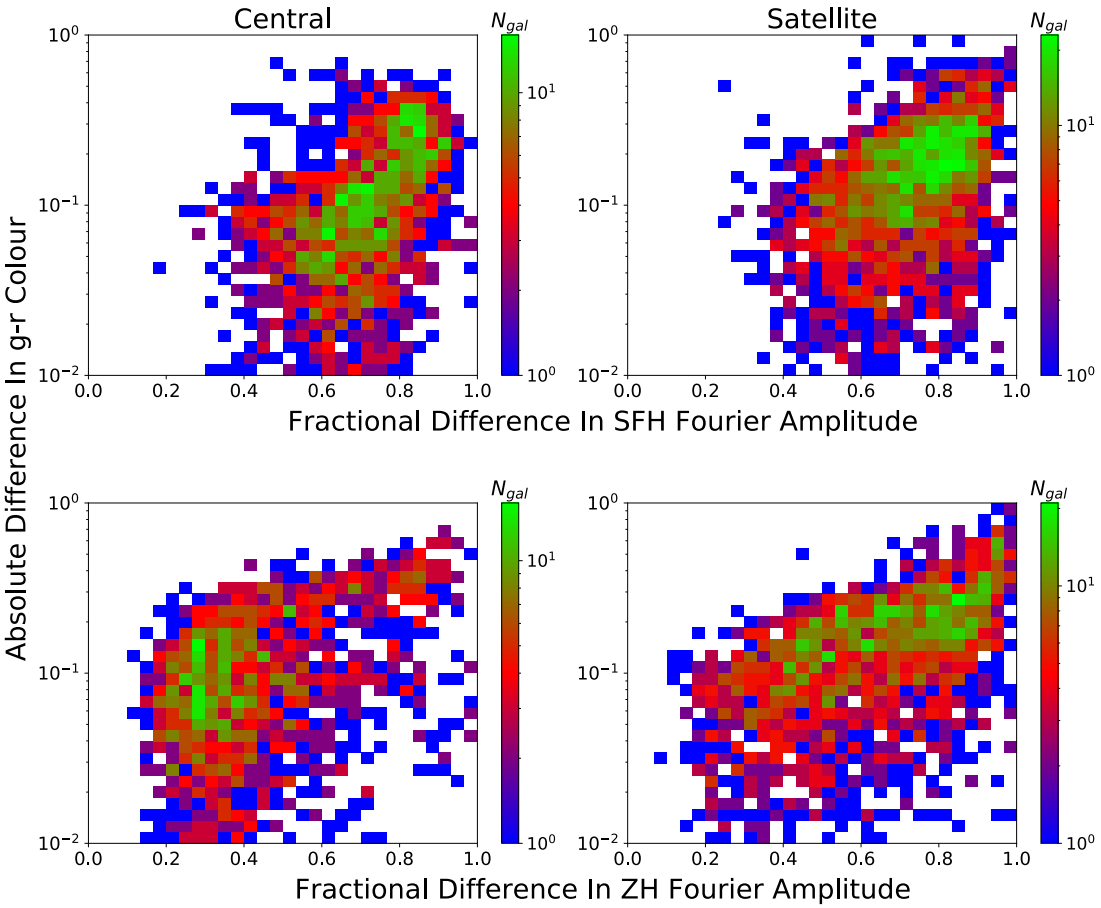


Figure 4.9: For the data shown in figs. 4.3 and 4.5, this figure shows 2D histograms of the absolute difference between true and predicted $g-r$ colours and the mean high-frequency Fourier amplitudes of their star formation histories (top row) and metallicity histories (bottom row). This clear correlation indicates the importance of measuring short-timescale star formation and chemical enrichment events in the aim to calculate accurate colours.

of the star formation history.

The short-lived bursts of star formation which we aim to model have a considerable impact on the accuracy of photometric colours, especially when they occur recently (Chaves-Montero & Hearin, 2021; Fraser, Tojeiro, & Chittenden, 2022). However, the stochastic model does not predict large fluctuative events in the formation histories, nor does it strictly constrain late-time features, which contribute significantly to the spectral shape, and thus the colours derived from the spectrum.

While the inclusion of large spikes which the correction fails to reproduce would improve luminosity calculations, the timing of such features would be imperative to accurately determining colours. If a robust Fourier-based correction does not accurately amend photometric colours, a method with explicit use of time information, such as a wavelet transform, may prove necessary.

4.4 Summary

This chapter details our investigation of the observational properties which were computed by applying the FSPS code to three sets of star formation and metallicity histories: those from the TNG predicted by our neural network model, and those adjusted by the stochastic model described in section 3.3. Specifically, we examined the spectral energy distributions of these galaxies from near-ultraviolet to near-infrared wavelengths, and from these derived photometric band magnitudes and colours, and Hα emission line luminosities. Based on these analyses, we have drawn the following conclusions:

1. Predicted spectral energy distributions for galaxies of a given mass are similar in their average luminosity to their targets, but slightly lower in the mean and variance of this amplitude. The stochastic amendment is able to rectify this to a varying extent, being significant for intermediate mass but not high mass galaxies. These results correspond to the physical properties described in detail in chapter 3, which, while recovering similar summary statistics to the original simulation, are marginally improved by the stochastic correction.

2. Central galaxies are more pervious to underpredicted luminosities than satellite galaxies. We argue that this is due at least in part to the well-defined timescales

of satellite quenching, and the prevalence of quenched satellites effectively eliminating errors due to missing SFH and ZH features. The underpredicted luminosities of most galaxies are partially rectified by the stochastic correction, suggesting that these missing features are important to the calculated spectrum.

3. Inaccuracies in the network's predictions, namely the abscence of short-timescale SFH and ZH features, results in a reduced range of band magnitudes, emission line luminosities and photometric colours; as well as misaligned UV band photometry. The residuals in many of these quantities between the TNG and network-predicted values are strongly correlated with high frequency Fourier modes, illustrating that this is in fact a consequence of the lack of high frequency information in the network's predictions. Despite this, the stochastic correction has only a small effect on photometric quantities. The high mass features which have profound effects on the total luminosity and thus photometry are not well modelled by the stochastic amendment, whereas smaller features which typically influence emission line luminosities are modelled more accurately.

We have shown that our neural network which incorporates a causal model of the galaxy-halo connection can be used to indirectly predict spectroscopic and photometric galaxy properties, and recover important observational statistics, such as the colour bimodality of galaxies as a function of mass. The predictions of the model when applied to high volume N-body simulations, which we investigate in chapter 5, can therefore be used to predict observations which reflect the underlying galaxy-halo connection, and by complementing large, deep galaxy surveys can be used to gain insight into the GHC in the real universe.

Nevertheless, the SEDs corresponding to neural network predictions contain inaccuracies derived from errors in the prediction of physical properties. The stochastic correction has made some noticeable improvements to the predicted observables, yet this too is subject to systematic inaccuracies, which result in limited improvements to photometry. This suggests that a more vigorous correction algorithm is required for comprehensive galaxy mocks, if not an updated machine learning model.

This chapter is based on the methodology and results presented in Chittenden, Tojeiro, & Kraljic (in prep.).

5 Dark Simulations

5.1 Introduction

In previous chapters, the machine learning model predicting galaxy formation histories in the TNG simulations is described and tested, replicating important physical and observational galaxy-halo statistics. Having computed these results solely from the dark matter component of these galaxies, it is in principle possible to reproduce similar galaxy properties using pure dark matter simulation data, which, as discussed in section 1.3.1, is a means to producing a highly detailed galaxy formation catalogue on scales beyond the computational limits of full physics simulations; from which numerous enterprises would benefit (Habouzit et al., 2022; Johnston et al., 2023; Yuan et al., 2023). In practice, however, the N-body simulation data differs from its equivalent hydrodynamical model due to the abscence of baryons (Castro et al., 2020; Anbajagane et al., 2021; Haggar et al., 2021; Mansfield & Avestruz, 2021; Riggs et al., 2022). Furthermore, alternative models and parameterisations of the N-body simulation in question can affect the properties of their

halos, including the simulation resolution (Knebe et al., 2000; Rau et al., 2013; Angulo et al., 2014), the choice of cosmological parameters (Dooley et al., 2014; Villaescusa-Navarro et al., 2021), and the algorithms identifying halo and cosmic structures (Onions et al., 2012; García & Rozo, 2019; Zhang et al., 2022).

In this chapter, we investigate the effects of applying the machine learning model to data from dark simulations, in order to assess the suitability of the existing model for applications involving the scales of high fidelity N-body simulations. Specifically, we examine the impact of the absence of baryonic physics on the TNG simulations' dark counterparts to the hydrodynamical data utilised thus far (hereafter TNG-Dark), and we investigate the effects of lower resolution or different measures of halo properties on the pure dark matter simulation Uchuu (Ishiyama et al., 2021), which assumes the same Planck Collaboration (2016) cosmological model as all TNG simulations, thus eliminating any discrepancies relating to matter density, growth timescales or cosmic expansion. For specifics regarding the simulation parameters and further details of the simulation suites, see table 2.1.

By computing suitably accurate data in the Uchuu simulation, we effectively demonstrate that the neural network can be applied to existing high volume N-body simulations, and reproduce detailed galaxy formation histories reflecting the galaxy-halo connection in TNG, in volumes exceeding the largest present-day cosmohydrodynamical simulations. However, a number of logistical challenges are presented. Beside the lack of baryons, the Uchuu simulation has lower mass resolution than the TNG simulations, and halo variables including mass and half-mass radius are calculated using different methods to those used in TNG. Accounting for alternative data when applying novel methods is common in galactic astrophysics; yet for a machine learning model trained on a specific dataset, the failures of the model when applied to another simulation reflect the true versatility of the model, and provide insights into the necessary adjustments to advocate its practicality in independent applications.

In this chapter, section 5.2 covers the definitions of the dark matter data in the TNG-Dark and Uchuu simulations, as well as our data acquisition techniques. We compare and discuss the input data properties in section 5.3, while section 5.4 focuses on the baryonic output of the neural networks as well as the derived observational results, highlighting

differences between the original TNG data and the network's predictions in the hydrodynamical simulations (TNG-Hydro), and the TNG-Dark and Uchuu data. The model's ability to effectively model the GHC on gigaparsec scales are debated in section 5.5, with its successes and failures presented alongside potential changes to the model design. Finally, we summarise our findings in section 5.6.

Note that the stochastic correction introduced in section 3.3 and evaluated using observational data in section 4.3 is not discussed in this chapter. As the objective of this chapter is to compare direct predictions of the neural network under different simulations and isolate the causes of resulting discrepancies, adjustments which may add further effects to the data, potentially depending on the simulation in question, are intentionally excluded.

5.2 Dark Simulation Data

5.2.1 TNG-Dark

The TNG-Dark simulations have a resolution comparable to their hydrodynamical counterparts, as shown in table 2.1; implying that resolution-related effects can be largely ruled out. Additionally, the domain of cosmic redshifts at which their snapshots are defined is identical in both hydrodynamical and dark simulations. The primary distinction between the two simulations is therefore the lack of baryonic effects on the halos in the TNG-Dark simulations.

To evaluate the variations between the dark and hydrodynamical TNG simulations, samples in the dark simulation are obtained by cross-matching SubLink trees (Nelson et al., 2015; Rodriguez-Gomez et al., 2015) with the hydrodynamical dataset discussed in previous chapters. Samples in TNG-Hydro without a cross-matched subhalo in TNG-Dark are disregarded. This selection process has a negligible effect on the $z = 0$ halo mass distribution of our data for central halos, yet for satellite objects, the sample count begins to decline rapidly, below approximately $4 \times 10^{10} M_{\odot}$.

In general, there are two reasons why objects in one simulation may not be accurately paired with objects in the other simulation. First of all, halos close to significantly larger halos may move within the larger halo's virial radius in one of the two simulations, leading to central halos being paired with satellites, and vice versa. Second, and especially for

lower mass objects, the SubFind algorithm may combine two separate subhalos into a single entity, resulting in one of the low mass subhalos being undefined in the lower resolution simulation; that being TNG-Dark. These effects collectively result in the removal of 39% of our satellite subhalos, compared with only 1% of central halos.

The majority of TNG-Dark variables were calculated using the same methods as those used in the hydrodynamical simulations, i.e. as discussed in section 2.4. However, there is one exception: the distances to points in the cosmic web at redshift zero, defined using DisPerSE. Because of the geometric relationship between the hydrodynamical and dark simulations, these distances are identical to their cross-matched counterparts. Consequently, we attribute the cosmic web characteristics of TNG-Hydro samples to their cross-matched equivalents in TNG-Dark, as there is no publicly available data for the latter.

5.2.2 Uchuu

The merger trees within the Uchuu simulation are divided into 2,000 "Forests", each of which consists of a collection of merger trees containing all halos which have interacted with any member of that forest. These forests are situated in a distinct volume of space, separate from all other forests. As a result, each forest can be analysed separately, as the Consistent-Trees algorithm (Behroozi et al., 2012b) used to create the merger trees in Uchuu is executed autonomously in groups occupying a fixed volume. If they interact or come within 25 Mpc/h of one another, the groups are concatenated to form a forest (Ishiyama et al., 2021).

For this study, we focus on forest 1411, which is the largest forest in the Uchuu simulation. We obtain comparable samples to the TNG dataset by selecting from the TNG-Dark halo mass distribution and sampling the Uchuu forest at $z = 0$ accordingly, as shown in fig. 5.1. This yields approximately 30,000 central and satellite halos each from the Uchuu forest.

Given that the selected Uchuu forest has a geometric mean side length of about 234 Mpc, and hence a volume similar to that of the TNG simulations, sampling the (sub)halo mass distributions effectively approximates the (sub)halo mass functions of Uchuu for high mass halos. At lower masses, sampling Uchuu in accordance with the TNG-Dark

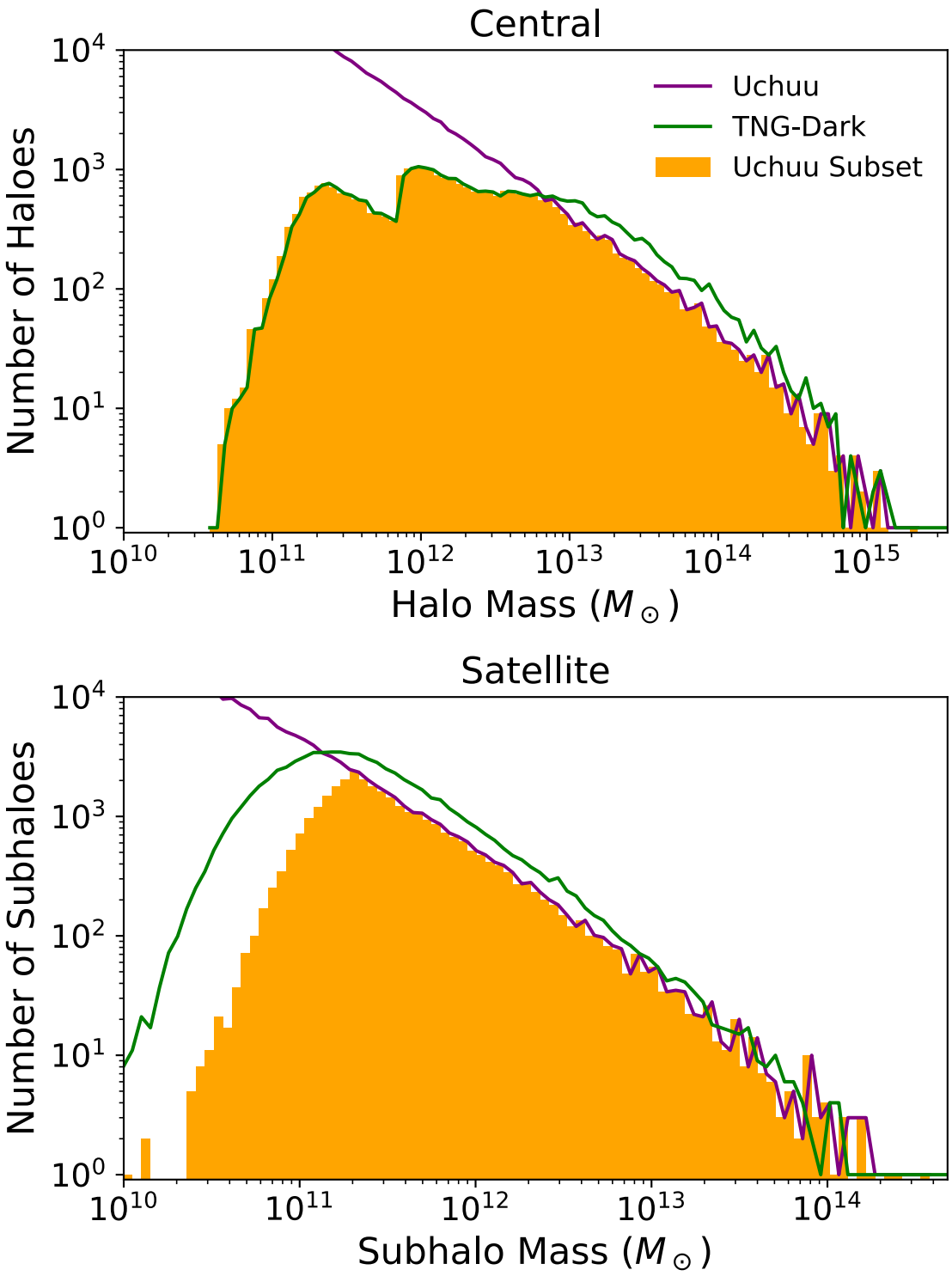


Figure 5.1: This graph depicts the distributions of central halo and satellite subhalo masses in the entire Uchuu forest, shown in purple, and our cross-matched TNG-Dark sample, shown in green. By drawing samples from the former distribution according to the latter, we derive the distribution of Uchuu halos used in our study, represented by the orange histogram. The distribution of central halos closely resembles the TNG-Dark data, but the lack of well-defined satellite subhalos at low mass results in a skewed distribution of satellite subhalos in our Uchuu dataset.

distribution makes it possible to compare attributes across multiple mass ranges without being influenced by sample size. While this selection becomes our test data for the neural network, environmental quantities are computed using the entire Uchuu forest and are considered unbiased due to the TNG and Uchuu simulations having identical cosmic matter density parameters.

We utilise the YTree Python package to extract the MPBs from the Uchuu forest. From the forest, we directly acquire information on halo mass, half mass radius, positions, and peculiar velocities. We obtain cosmic web distances by employing the DisPerSE algorithm on the forest. As the forest is a self-consistent subset of Uchuu, the algorithm does not need to be executed upon the entire Uchuu simulation. We calculate all other quantities, such as overdensities, using the same methods as in section 2.4.

In contrast with TNG, there are no SubLink merger trees in Uchuu, and halos along with their substructures are defined according to the Rockstar algorithm (Behroozi et al., 2012a). Rockstar calculates a hierarchical series of halo structures which exist within a larger ensemble. In each Uchuu merger tree, there exists a flag indicating the ID of the halo which houses the target halo. A "first order" satellite halo is one that is hosted by a central "zeroth order" halo, a "second order" halo is hosted by a first-order halo, and so on[1]. First-order halos are therefore considered satellite halos in this work, while zeroth order halos are defined as central halos.

The data presented in table 2.1 indicates that the mass resolution of Uchuu is around ten times lower than that of TNG300 simulations. Consequently, Uchuu has poor resolution for low mass haloes, and due to the quality control measures discussed in section 2.4, the number of satellites below approximately $2 \times 10^{11} M_{\odot}$ is reduced. This is corroborated by fig. 5.1, showing a very different distribution of subhalo masses below this threshold.

A second aspect to consider is the variation in the time resolution of the Uchuu snapshots when compared to TNG, which is shown in fig. 5.2. Although Uchuu has half as many snapshots as TNG, the time interval between the snapshots is smaller than TNG for redshifts ranging approximately from 2 to 4, but larger for other redshifts. In both simulations, all temporal features are linearly interpolated over the same 33 snapshots in TNG. However, due to the infrequent nature of these snapshots in Uchuu, there is a

[1]See fig. 2.6 for an illustration and description of this hierarchical merger tree structure.

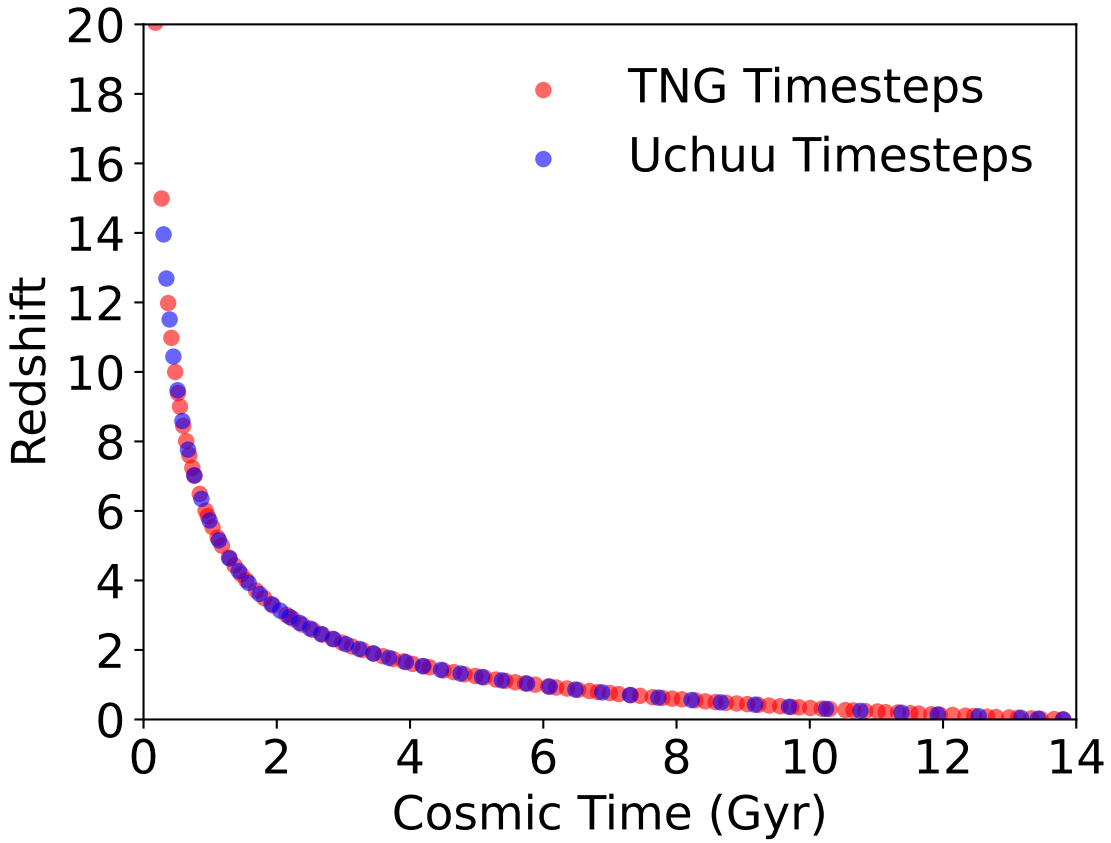


Figure 5.2: This schematic illustrates the redshifts and cosmic time of the snapshots for the TNG simulations represented in red, and for the Uchuu simulation in blue. Despite having fewer snapshots than TNG, Uchuu has a higher temporal resolution during early times and is more sparsely sampled for $z < 2$.

possibility that information pertaining to short time scales may not be captured.

5.3 Halo Variables

The procedures used to obtain input data for each simulation in this chapter are the same as those described in section 2.4. However, there exist notable differences in the definitions of properties and their statistics depending on the model, which are affected by factors such as the exclusion of baryons or the resolution of the data. In this section, we examine the similarities and differences in the input properties of the neural network in each simulation, and how these variations can impact the predictions made by the neural network.

To examine the differences in various evolutionary stages, we compare properties for central and satellite halos that are sorted into different mass and accretion history bins. For central halos, we categorise halos with distinct accretion histories based on the specific mass accretion gradient β, which Montero-Dorta et al. (2021) have demonstrated to be linked to gas fraction, quenching timescale, and assembly bias in TNG300. In section 2.4.1 we introduce this parameter both as an input variable to the model, and as a measure to establish quality cuts for both central and satellite datasets. For modeling satellite subhalo histories, it is not as effective due to the different modes of accretion in their central and satellite phases, hence we use the scaled accretion time $a_{\rm max}$, which was introduced as an input variable in section 2.4.1. Like β, $a_{\rm max}$ is explicitly derived from the subhalo mass accretion history and is linked to satellite galaxy characteristics, as shown by Shi et al. (2020).

The tabular figures used in this chapter compare different halos based on their histories. The subplots are arranged such that each column represents a quintile of halo or subhalo mass, with the larger masses being on the right side of the plot. Each row represents a quartile of accretion gradient, with the steepest accretion histories appearing on the topmost row. For the satellites, the higher $a_{\rm max}$ values are placed on the top row, whereas central galaxies have the smallest β values on the first row, such that the earliest half-mass formation times appear on the topmost row in both cases. The percentiles used in this arrangement are taken from the TNG-Hydro data. All figures of this kind follow this convention to ensure that the gas-poor halos or subhalos with the earliest formation times

appear on the top row.

5.3.1 Mass Accretion History

In TNG, the mass of halos and subhalos is determined by adding up the mass of all the dark matter particles bound to the group, as determined by the FoF and SubFind algorithms, respectively. While this specific field is not present in the Uchuu merger trees, there are several definitions of mass that are available. As established in section 2.4.1, the closest match to the TNG object masses which exists in both TNG and Uchuu is found to be M_{200c}, defined as the total mass enclosed within a region with an overdensity 200 times that of the critical density of the universe.

The mass accretion histories of halos in TNG-Dark and Uchuu exhibit similar patterns, leading to similar overall mass distributions, regardless of the chosen halo mass field. As a result, the neural network predictions are also similar when using M_{200c} to represent halo masses in TNG, as when using the SubFind mass. As this suggests that the differences in these halo mass properties has little effect on the predictions of the model, M_{200c} is chosen as the field to represent halo masses in Uchuu, and calculations which depend on halo mass, such as overdensities, are performed using this quantity.

Jiang et al. (2014) show that these mass definitions tend to agree well, provided that there is a unique bijection between FoF and Dhalo merger trees[2] for well-resolved halos; whereas the variance in the ratio of these masses owes predominantly to large substructures outside the virial radius: usually infalling or diffuse halos. In Millennium-II (Boylan-Kolchin et al., 2009), an N-body simulation of similar resolution to TNG, Jiang et al. (2014) show that this is the case for most halos above $10^9 M_\odot$, therefore this mass property can be assumed for the majority of TNG samples in this work, while the abscence of low mass Uchuu satellites owes to quality cuts of the mass accretion histories (see section 5.2.2), which likely removed samples with significant halo mass discrepancy.

Figures 5.3 and 5.4 illustrate the median and range between the fifteenth and eighty-fifth percentiles of the mass accretion histories, separated into bins based on the final (sub)halo mass and mass accretion gradient. In the majority of bins, the mass accretion

[2]The Dhalo algorithm developed by Jiang et al. (2014) is a method of constructing halo merger trees by linking subhalos over intervals of multiple snapshots, avoiding the false unification of distinct substructures, which is an issue with some halo finder algorithms.

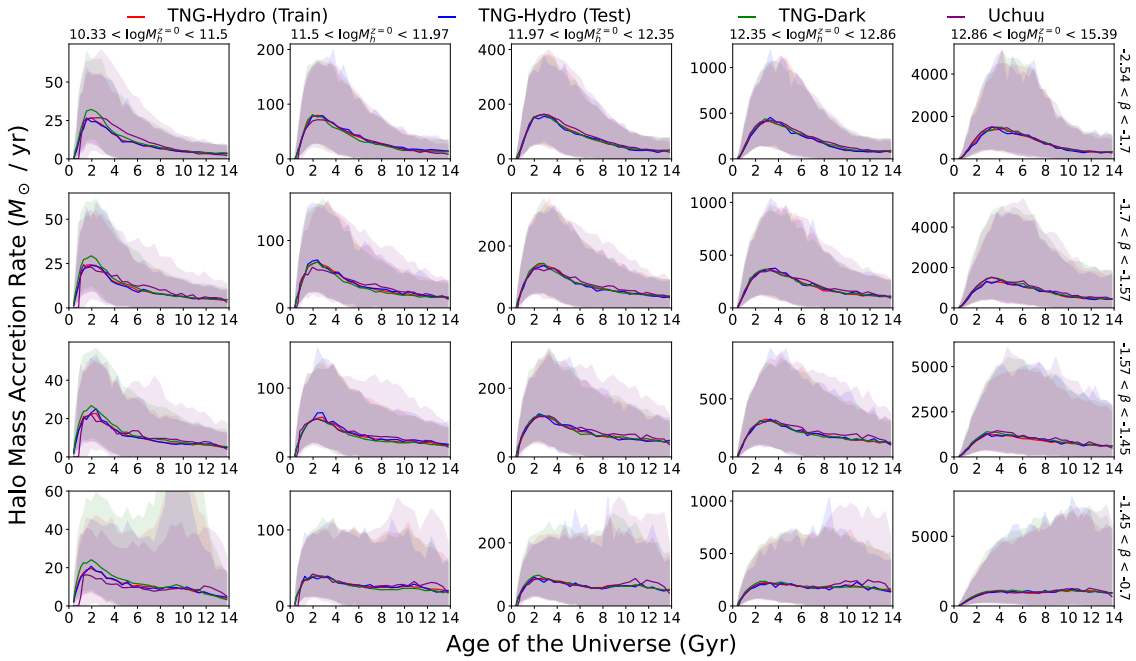


Figure 5.3: This schematic illustrates how the mass accretion histories of central halos are distributed according to the halo mass and specific mass accretion gradient. The horizontal axis represents the halo mass and increases in value from left to right, while the vertical axis shows the specific mass accretion gradient and decreases in steepness from top to bottom. The solid lines display the median mass accretion history for each bin, while the shaded regions represent the $15^{\text{th}} - 85^{\text{th}}$ percentile ranges of the binned data. The mass accretion histories for the TNG-Hydro simulations' training and testing datasets are presented in red and blue, respectively, while the green and purple data correspond respectively to the TNG-Dark and Uchuu simulations.

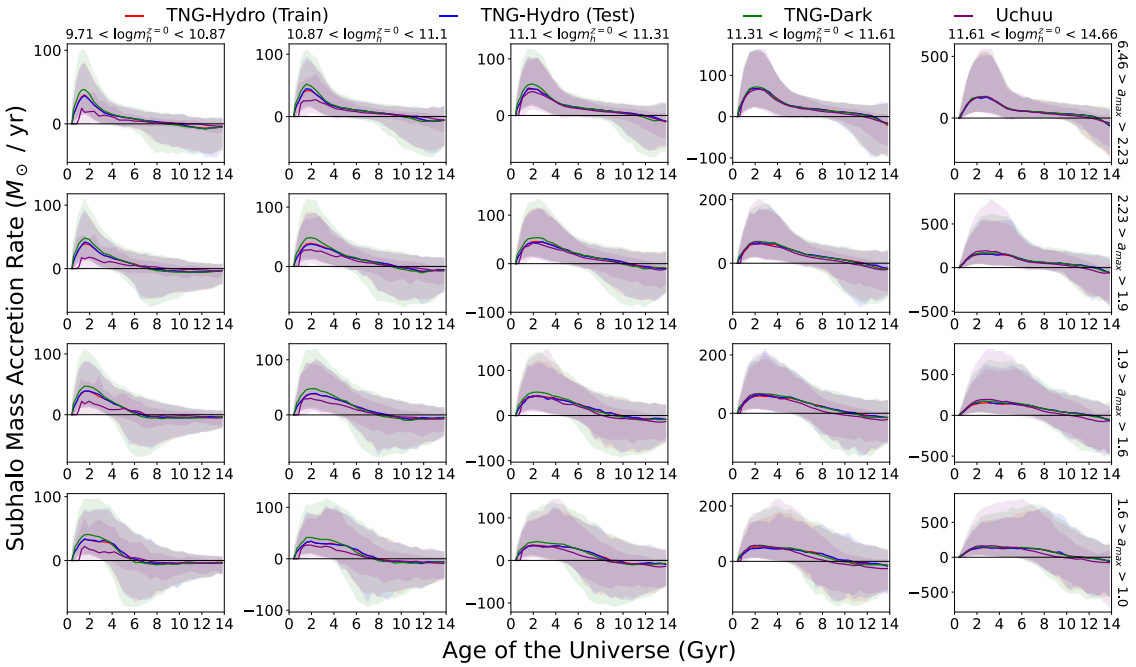


Figure 5.4: Mass accretion histories of satellite subhalos, categorised similarly to fig. 5.3, with the satellite subhalo mass m_h replacing the central halo mass M_h, and the scaled accretion time $a_{\max}$ replacing the specific mass accretion gradient β. The same colour and percentile schemes used in the previous figure are adopted here. One significant difference between these accretion histories and the previous ones is that the satellite subhalos' accretion approaches zero or becomes negative, which is uncommon to central halos. The various times at which the growth of the median subhalo terminates is shown in various growth regimes.

histories are alike between the four datasets. However, there are differences between the simulations, specifically in low mass and shallow gradient scenarios. In such cases, the amplitude of the Uchuu mass accretion histories is lowered, while TNG-Dark halos display increased mass accretion at early times in the low mass regime.

The increase in mass for low-mass halos in the early stages is most likely due to the lack of baryonic-driven outflows, which significantly impact objects of a high gas fraction. On the other hand, the lower accretion rates in the Uchuu simulation are attributed to its lower mass resolution. For shallow gradient bins, the Uchuu accretion histories display a flat profile, similar to the TNG samples which were discarded by existing quality cuts.

Since Uchuu has lower mass resolution than any of the TNG simulations, the resolution of low mass halos in Uchuu will be poorer. In the lowest mass quintile, predicting galaxy evolution accurately from the start may be difficult due to the sensitivity of TNG's star formation algorithm to the number of dark matter particles (Pillepich et al., 2017a). Montero-Dorta et al. (2021) show that halos with shallower accretion gradients tend to form later, as seen in our Uchuu data; therefore halos in any given mass bin will be of

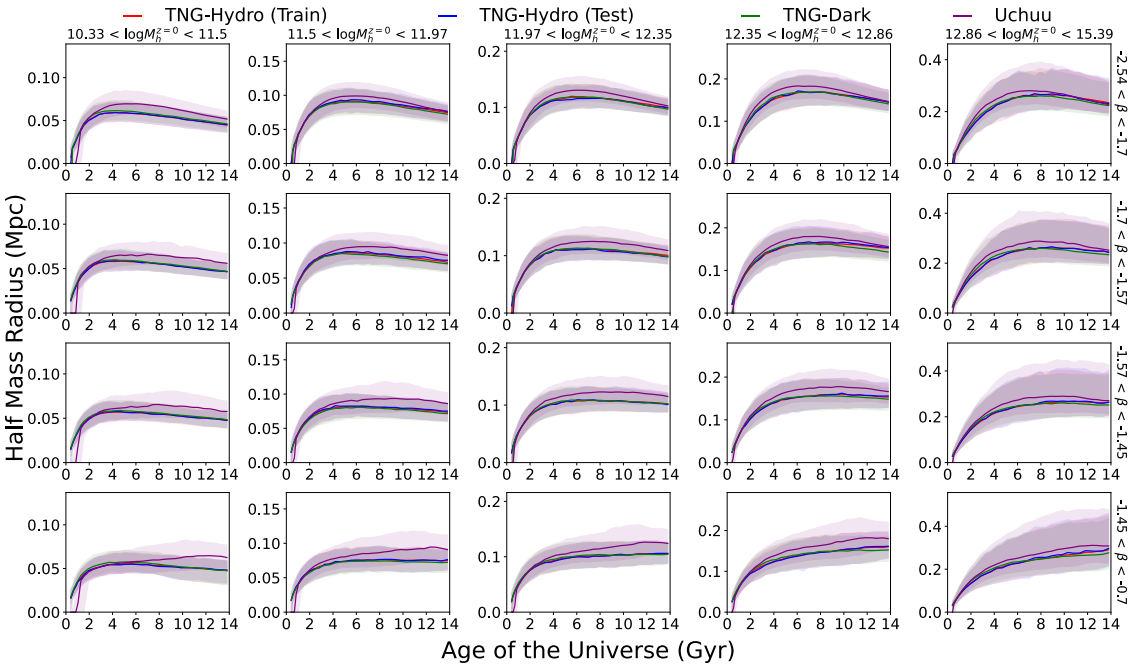


Figure 5.5: This figure illustrates the evolution of the half-mass radius of central halos over time. It is presented in the same tabular format as in fig. 5.3, with identical halo mass and accretion gradient bins.

lower mass at most nonzero redshifts, which means that they will be subject to similar resolution effects. For satellites with shallow accretion, there is a noticeable decrease in the median accretion rate compared to TNG, which is most likely due to the lack of low-mass objects being accreted onto the halo.

5.3.2 Half-Mass Radius

The most significant difference between Uchuu and TNG data is the half-mass radius of the halos. This is shown in figs. 5.5 and 5.6 for central and satellite haloes, respectively. In low mass and shallow gradient bins, there is a clear difference in the size evolution of haloes in Uchuu when compared with TNG. However, there are also small but noticeable differences in most other bins and at most times, with the haloes in Uchuu being slightly larger.

The central halos in TNG-Dark have a slightly smaller half-mass radius in high mass bins than the other simulations. This phenomenon could be elucidated by the findings of Haggar et al. (2021) and Riggs et al. (2022), showing that the number density of halos which are gravitationally bound to a galaxy group or cluster is underestimated in dark simulations relative to the hydrodynamical equivalent, within two virial radii of the cluster or galaxy group. The cause is due to the high density of baryons in the central region of

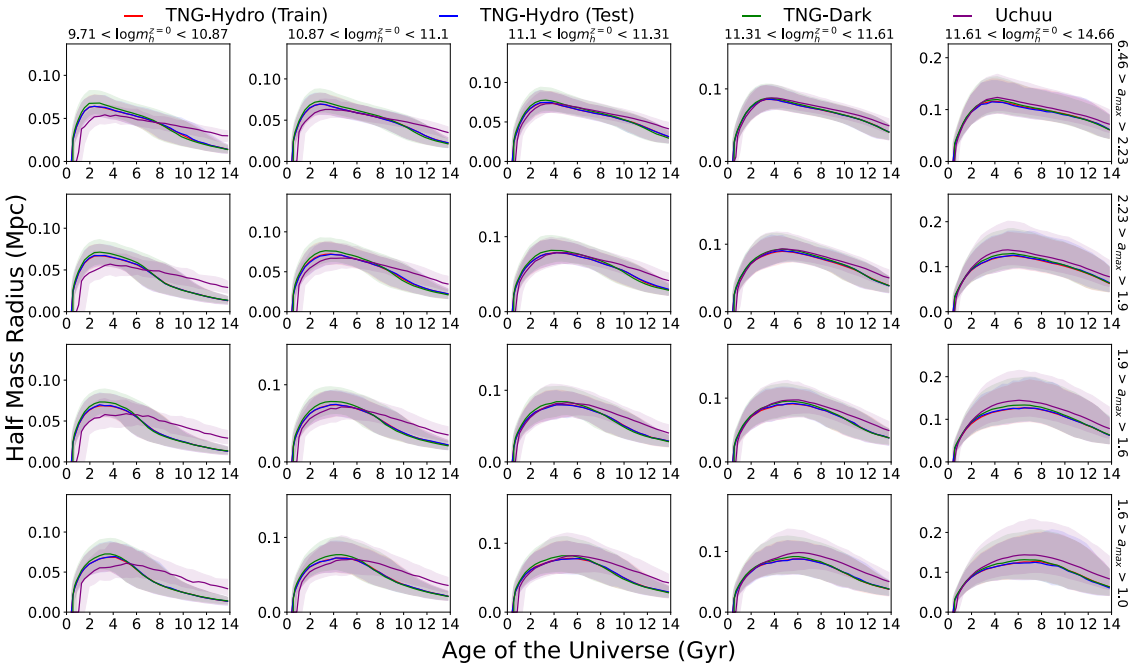


Figure 5.6: The half-mass radius growth of satellite subhalos, displayed in a tabular format similar to fig. 5.4, including the same bins for halo mass and scaled accretion time.

the large object, which subsequently concentrates its density profile.

Chua et al. (2019, 2022) find that the presence of baryons has a notable impact on the radial profiles and asymmetries of halos in the TNG and Illustris simulations. However, based on our findings, this effect does not affect the evolution of halo size in TNG-Dark as compared with the hydrodynamical simulation. To ensure that all variables used in our model were appropriate for the TNG-Dark simulation, they were compared with the full physics simulation at all times. However, in Uchuu, the half-mass radius is determined using the virial mass rather than the SubFind method used in TNG.

In the TNG simulation, the SubFind algorithm is used to determine subhalo boundaries by identifying a contour of constant density that meets a saddle point in the local density field (Springel et al., 2001). The half-mass radius is then calculated as the radius enclosing half of the mass within this boundary. On the other hand, in Uchuu, the half-mass radius is calculated based on a spherical region which encloses the virial mass (Ishiyama et al., 2021). This calculation depends on the density profile of the halo, which is derived from a direct fit of an NFW profile using the Rockstar halo finder (Behroozi et al., 2012a).

As stated in sections 2.4.1 and 5.3.1, the different algorithms for calculation of halo mass yield similar results, making little difference to the training and testing of the machine learning models. We find similarly, in TNG, that the radii of haloes established by

these means are also tightly correlated. However, the NFW fitting method used in Uchuu may prove to be an inaccurate measure of the halo's size and structure at a given mass, due to non-virial features, e.g. irregular shape owing to merger activity and tidal asymmetry; and the overlap of distinct halo structures, which can affect the agreement between virial and FoF mass, as established in section 5.3.1. The concentration of these haloes, scaling inversely with half-mass radius, may be exaggerated if the halo in question has not undergone sufficient internal gravitational collapse; yet figs. 5.5 and 5.6 appear to downplay the halo concentration instead. The concentration parameter is additionally sensitive to the resolution of the halo being fit, which arguably has a more substantial influence on the halo profile. We therefore argue that the simulation resolution has greater effect on the half-mass radius quantity than the halo finder; and due to the strong agreement between the FoF and virial masses of most data, the different mass definitions used in TNG and Uchuu are also considered less important than the resolution difference, yet this could also influence the half-mass radius calculation for low mass halos.

Zhao et al. (2009) demonstrated that the increase of the NFW concentration parameter over time in N-body simulations is dependent on the time of formation of 4% of the final halo mass, while Prada et al. (2012) find that the concentration is also sensitive to fluctuations in the linear density field on the scale of the halo's mass. These factors rely on the simulation's resolution, which can impact the growth of halo concentration by producing more extended mass distributions at late times and smaller radii of low mass haloes at early times. Ishiyama et al. (2021) show that the smaller, higher resolution Shin-Uchuu simulation exhibits different mass-concentration relations relative to the larger Uchuu model, indicating that structural differences between simulations are resolution-dependent. Additionally, morphological halo quantities such as virial velocity and axis ratios are affected by the gravitational softening scale, which is larger in Uchuu than in TNG, potentially resulting in a flatter $M_h - v_{\max}$ relation at low mass and a similarly altered mass-concentration relation (Mansfield & Avestruz, 2021).

The smaller, younger halos are most affected by the difference in halo concentration due to Uchuu's lower resolution, both for central and satellite halos. These halos are more likely to experience delayed growth in the Uchuu data because they are growing from low mass progenitors. The subhalo finder algorithm is unlikely to have a significant impact on the inferred halo shape, as long as the subhalo has a sufficient number of particles

(Hoffmann et al., 2014), which is not the case for low mass, low resolution halos. Thus, the delayed growth of the half-mass radius is primarily a resolution effect and is independent of the algorithm used to define the halo. However, Rockstar is believed to be superior to most algorithms in identifying halo substructures in dense halo centres (Onions et al., 2012), which can affect the value of the half-mass radius. The larger range of radii seen in some images from Uchuu for all times suggests that this is also an important factor to consider.

5.3.3 Maximum Orbital Velocity

Our model employs a proxy for the virial circular velocity of the halo, which is based on the two aforementioned quantities: halo mass and half-mass radius. The reason for the use of a proxy despite the definition of orbital velocities in the simulations is the significant differences between the orbital velocities observed in TNG-Hydro and TNG-Dark, particularly at high redshifts.

Owing to the displayed discrepancies in these quantities with the Uchuu data, the median values of the circular velocity are slightly underestimated. Nevertheless, the shape of the median curve remains consistent with its TNG counterparts. Conversely, the proxy tends to be overestimated in TNG-Dark for low mass bins, due to excess mass accretion. A similar trend is observed for high mass centrals, albeit to a lesser extent and as a result of smaller half-mass radii.

5.3.4 Local Environment

In our neural network model, we have assessed the environmental histories of central and satellite haloes by considering subhalo overdensities for each snapshot in the simulation. Additionally, we have introduced a radial skewness (skew) parameter to capture the interaction histories of these halos. In section 3.4.3 we argue that, although mass and structure features are crucial in predicting the star formation histories of the galaxies in these halos, these environmental properties have been demonstrated to be important in predicting the metallicity histories of TNG galaxies.

As mentioned previously, the Uchuu catalog defines halo substructure differently by using hierarchical, nested Rockstar halos in place of subhalos. Consequently, overdensity and skewness calculations are based on halo tracers.

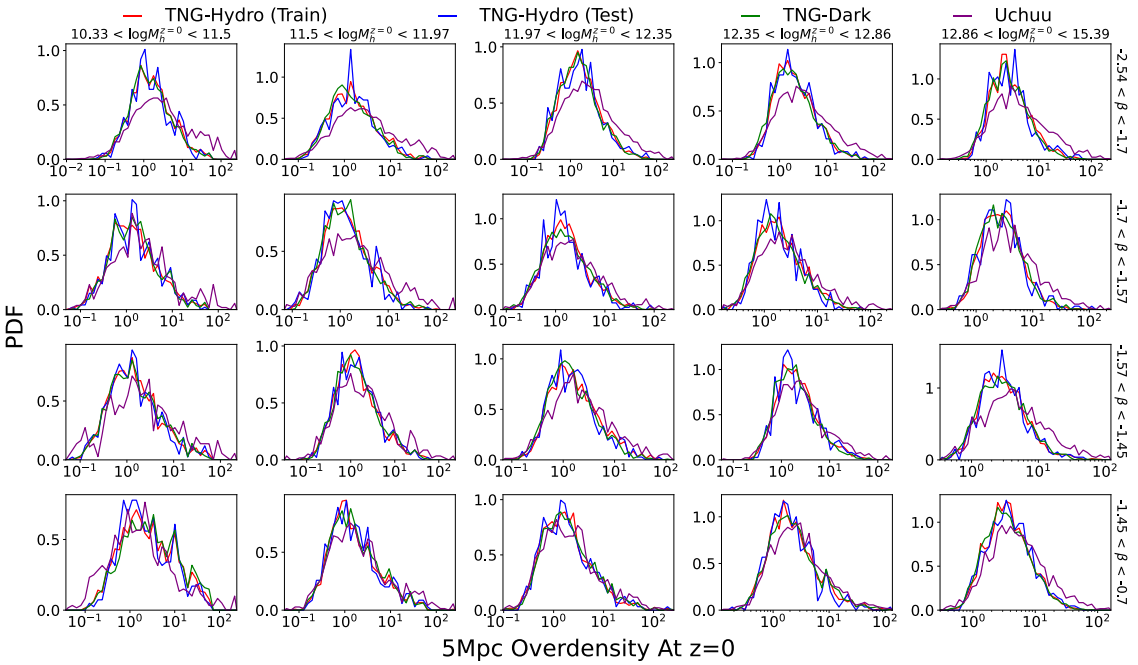


Figure 5.7: The histograms in each panel illustrate the probability density functions of logarithmic dark matter overdensities, contained in a spherical region of 5Mpc radius, surrounding halos at $z = 0$ in the four simulation datasets. The distributions of overdensity are comparable for most mass and accretion histories across the simulations, with the exception of the Uchuu dataset, which generally has higher densities due to the use of halo tracers.

Figure 5.7 shows that overdensities in Uchuu are marginally larger than those in TNG-Dark. In simulations with lower resolution, the calculation of environmental properties is more susceptible to edge effects of the calculation volume, and the calculation of the halo's centre of mass by the Rockstar algorithm being sensitive to the positions of the 0.1% most gravitationally bound particles (Behroozi et al., 2012a), which are of course influenced by the simulation resolution. These factors may contribute to the differences in overdensity.

However, the skew is not significantly affected by resolution differences. Since the skew is a mass-weighted quantity and independent of scaling, the difference in overdensities is most likely due to the mass content of the contributing density tracers. The invariance of the skew may imply that the ambiguity of the halo centres of mass is insignificant within the calculation volume, however this is more likely to influence the contribution of low mass objects, or tracers near the boundary of the calculation volume; which are more influential to the calculation of overdensity than they are to the calculation of skew.

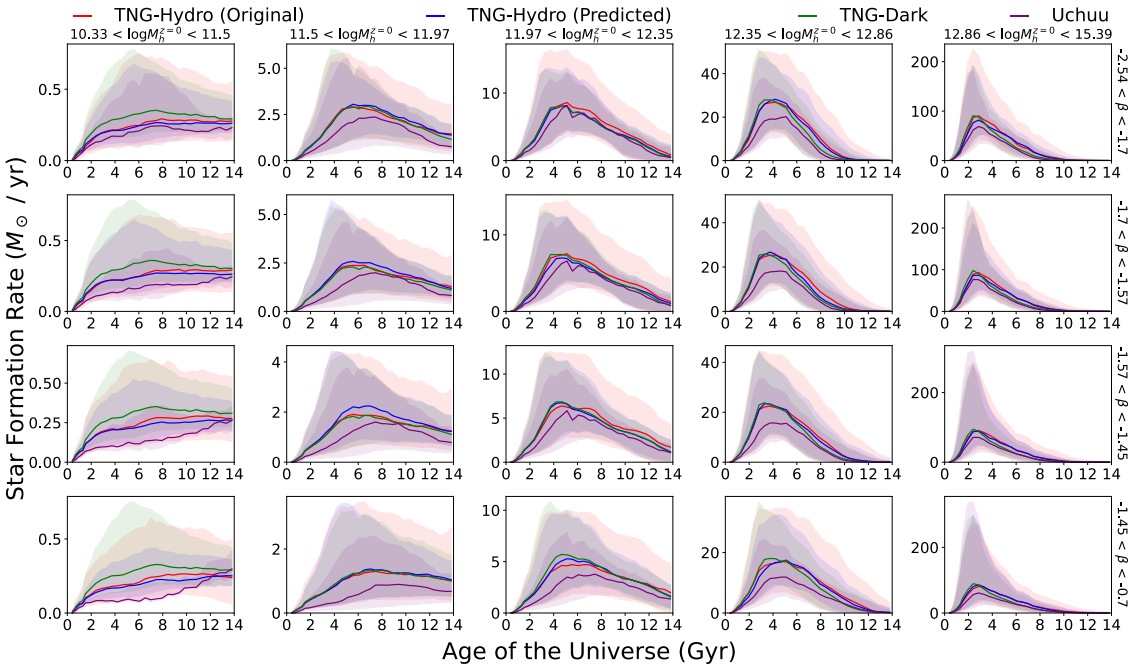


Figure 5.8: This figure depicts the star formation histories of central galaxies grouped by halo mass and specific mass accretion gradient. In low mass bins, TNG-Dark overestimates star formation rates, while Uchuu underestimates them. In higher mass bins, the difference between the two simulations becomes less pronounced.

5.4 Galaxy Quantities

In this section, we will describe how incorporating the various simulation variables into the neural network affects the quality of predictions. After outlining the causes of differences in halos and environments between simulations in section 5.3, we will examine how these differences impact both direct predictions and derived halo-galaxy relationships, while also offering a physical rationale for discrepancies in physical and observational results.

5.4.1 Star Formation History

Figures 5.8 and 5.9 illustrate the median and $15^{\text{th}} - 85^{\text{th}}$ percentile ranges of projected star formation histories, categorised by final (sub)halo mass and mass accretion gradient. The results demonstrate that galaxies in high mass and rapid accretion bins are generally well-matched, but in low mass bins, there are significant differences. Specifically, star formation rates in TNG-Dark are overestimated, while in Uchuu, they are underestimated.

Figure 5.9 may give the impression that most Uchuu satellite galaxies have a significantly lower predicted stellar mass. However, this is deceptive due to the lack of low-mass halos. The two lowest mass bins contain only 12% of the satellite galaxies in Uchuu, in

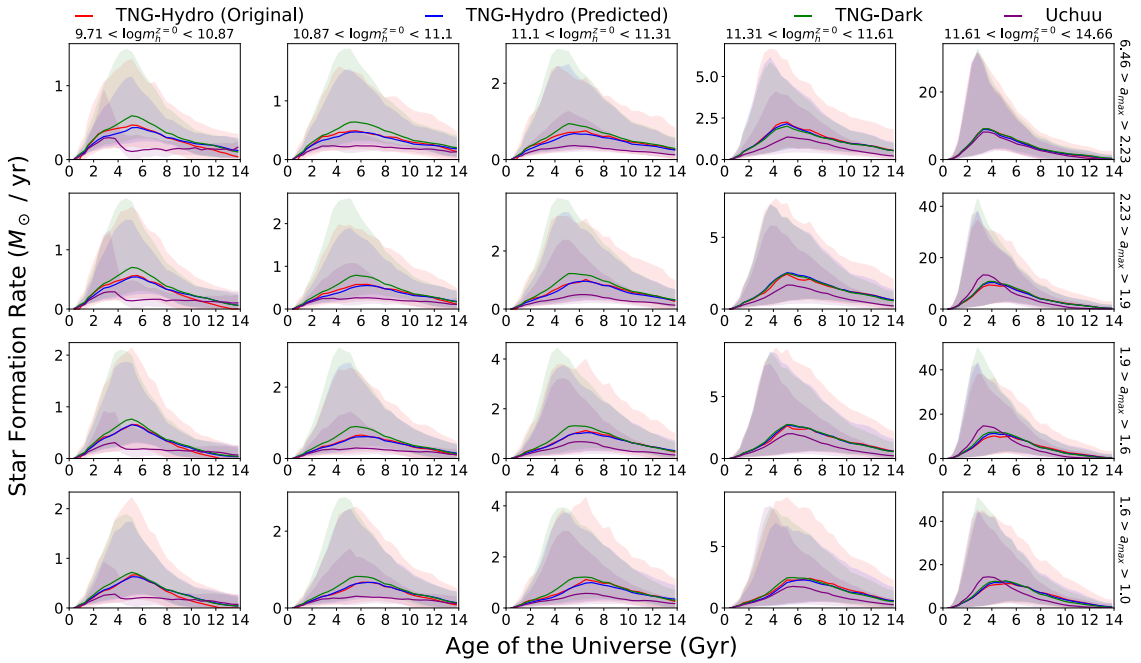


Figure 5.9: This figure displays the star formation histories of satellite galaxies grouped in the same manner as in fig. 5.8. The figure reveals seemingly inadequate predictions for the star formation histories of low mass galaxies in Uchuu. However, it should be noted that several of these bins in the Uchuu data have low population and are characterised by low-quality haloes.

contrast to 40% in TNG. Nevertheless, these star formation histories are still strongly underpredicted, and inadquate for use in Uchuu models. As their halo histories have demonstrated, these objects acquire their mass at a later time and at a slower rate than TNG halos of the same evolutionary regime. This is particularly evident for satellite galaxies, as shown by figs. 5.4 and 5.6.

Table 5.1 displays Spearman correlation coefficients for halo and galaxy properties in a narrow, low mass bin, for both central and satellite objects in Uchuu. The results indicate a strong correlation between final stellar mass and the circular velocity proxy in the Uchuu data, which is slightly smaller relative to the TNG data. There are also weaker correlations observed between final stellar mass and half-mass radius and overdensity, both of which exhibit differences in the Uchuu data.

However, the correlation between underpredicted star formation histories in fig. 5.9 and undermined mass accretion histories in fig. 5.4 is particularly clear for low mass satellites, and this would have had a causal effect on their galaxy growth from an early stage in their evolution. Acquiring star-forming gas is particularly crucial during early times, which would explain the lack of subsequent star formation in these predictions. Davies et al. (2019) show that in TNG simulations, low-mass halos exhibit a high gas fraction, hence

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

Table 5.1: Spearman correlation coefficients of various halo characteristics with metallicity and stellar mass, in narrow bins of relatively low (sub)halo mass. All the parameters are evaluated at the final snapshot of the Uchuu simulation. The centrals are considered in a halo mass range of $11.77 < \log M_h^{z=0} < 11.97$, whereas the satellites are taken from a mass range of $10.9 < \log m_h^{z=0} < 11.1$.

the star formation at early times is likely to be sensitive to the lack of accumulation of low-mass progenitors.

By comparing the neural network's performance on TNG-Dark with TNG-Hydro, we observe that the predicted star formation histories are boosted rather than reduced, in conjunction with excessive mass accretion rates. Sorini et al. (2022) have demonstrated that gas accretion and stellar feedback processes exert their greatest influence on the size and shape of halos and large-scale structures at higher redshifts. Furthermore, they suggest that stellar feedback is the primary cause of the suppression of star formation in low-mass objects. The absence of stellar feedback in dark simulations would have restricted the effects resulting in halo mass loss, which are especially prominent for low-mass objects. The overabundance of mass accretion in TNG-Dark would have led to an overestimation of the star formation rate.

Figure 5.10 shows the self-consistent and accurate stellar-halo mass relations obtained by numerically integrating the star formation histories of each simulation, as in eq. (2.24). There is little difference between the relations predicted by the TNG-Hydro and TNG-Dark simulations, except for a slight reduction in scatter in the dark predictions. However, the difference between the Uchuu simulations' stellar and halo mass distributions is noticeable, particularly at low masses. While there is a small underprediction of mass and scatter between intermediate to high masses, the Uchuu SHMR remains well-matched to the predictions in TNG-Hydro and TNG-Dark.

In the Uchuu simulations, reduced star formation rates lead to a significant bias towards lower stellar masses at low masses. Although low mass satellite halos have already been removed from our analysis, distorting the remaining halo mass distribution, there are still

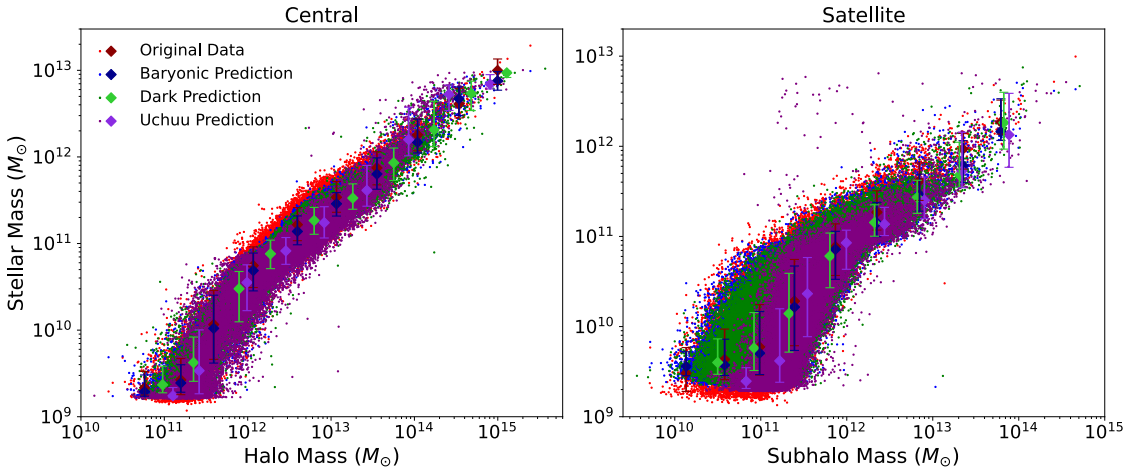


Figure 5.10: This figure depicts the quantitative SHMR of central galaxies (left) and satellite galaxies (right) based on numerical integration their star formation rates. Each individual galaxy is denoted by a data point, and the error bars denote the median and the fifteenth and eighty-fifth percentiles of stellar mass within a specific halo mass range. The similarity in the shapes of these relationships implies accurate prediction of the star formation histories in dark simulations.

a considerable number of objects with severely underestimated stellar masses for halo masses below approximately $2 \times 10^{11} M_\odot$, indicating that satellite objects in this mass range cannot be relied upon.

However, in the high mass bins, the majority of the stellar mass in both dark simulation datasets forms earlier than in TNG-Hydro. This is illustrated in fig. 5.11 by displaying the mass-weighted ages of galaxies in bins of stellar mass, indicating a bias towards older ages for high mass galaxies. This is partly due to a sharp increase in both halo and stellar mass at earlier times, followed by a more rapid decline in star formation rate, as shown in figs. 5.8 and 5.9. In TNG-Dark, the star formation histories initially align with TNG-Hydro, but the decline in star formation rates eventually matches the weaker star formation observed in Uchuu. In other words, the quenching of these galaxies is more efficient than in their hydrodynamical counterparts.

According to Davies et al. (2019), the expulsion of the circumgalactic medium in both TNG and Eagle simulations is strongly correlated with the central black hole mass of the galaxy, which in turn influences the specific star formation rate. Observationally, Bluck et al. (2020) found that the central velocity dispersion, which effectively measures the AGN mass, is a crucial factor in galaxy quenching in MANGA observations. This parameter is also correlated with the circular velocity and half-mass radius of the halo. Donnari et al. (2020) demonstrated that equivalently in TNG, internal feedback quenching mechanisms dominate central galaxy quenching, while environmental effects dominate

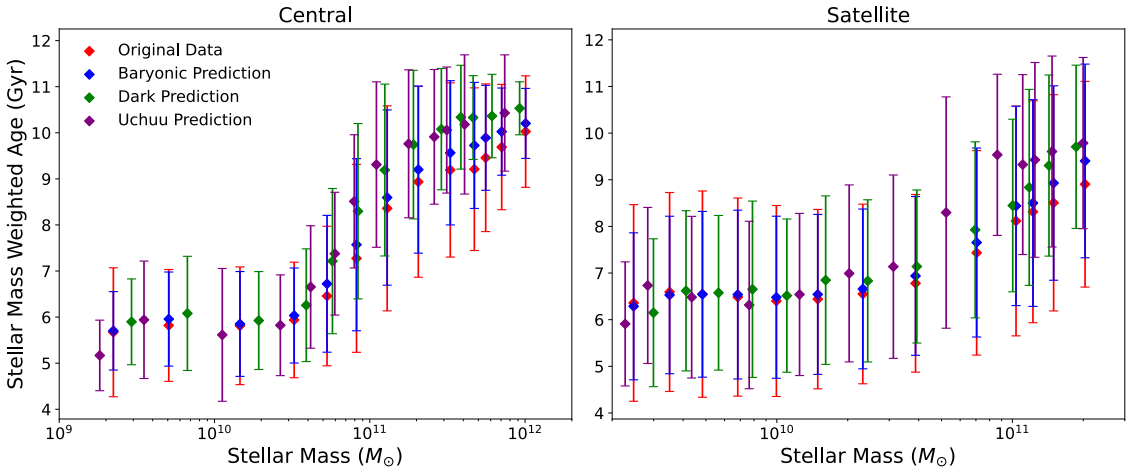


Figure 5.11: This figure shows the mass-weighted ages of central galaxies (left) and satellite galaxies (right) as a function of predicted stellar mass for the four simulation datasets, shown using the median and interquartile range of ages in different mass bins. This shows accurate recovery of the trend of age with mass in the dark simulations, yet there is a bias towards higher ages which increases with mass.

satellite quenching. These findings are in qualitative agreement with other hydrodynamical and semi-analytic models, as well as low-redshift SDSS data (Donnari et al., 2021).

Figure 5.5 demonstrates that in TNG-Dark, the half-mass radius of high mass halos is underestimated, which results in an overestimated halo concentration; closely related to the density and dynamics of the region surrounding the halo centre on sub-kiloparsec scales, as measured by the velocity dispersion. For galaxies of this mass, it is likely that an AGN dominates this region. Thus, overestimating the concentration can lead to an overprediction of AGN feedback, triggering early quenching of these galaxies.

5.4.2 Metallicity History

Figures 5.12 and 5.13 show the median and $15^{\text{th}} - 85^{\text{th}}$ percentile ranges of the metallicity histories of central and satellite galaxies in the four datasets. These demonstrate the inability to predict metallicity histories of low mass objects in Uchuu, exemplified by underpredicted chemical enrichment in the majority of satellite galaxies. Despite this, most intermediate to high mass central halos show good agreement between Uchuu and TNG.

When we compare the metallicity histories seen in TNG-Hydro, TNG-Dark, and Uchuu with their star formation histories, we find similarities. Given that the gas and metal composition of these objects' progenitors are crucial to early metal production, the suppression of chemical enrichment in low mass galaxies can be explained by the absence of early accre-

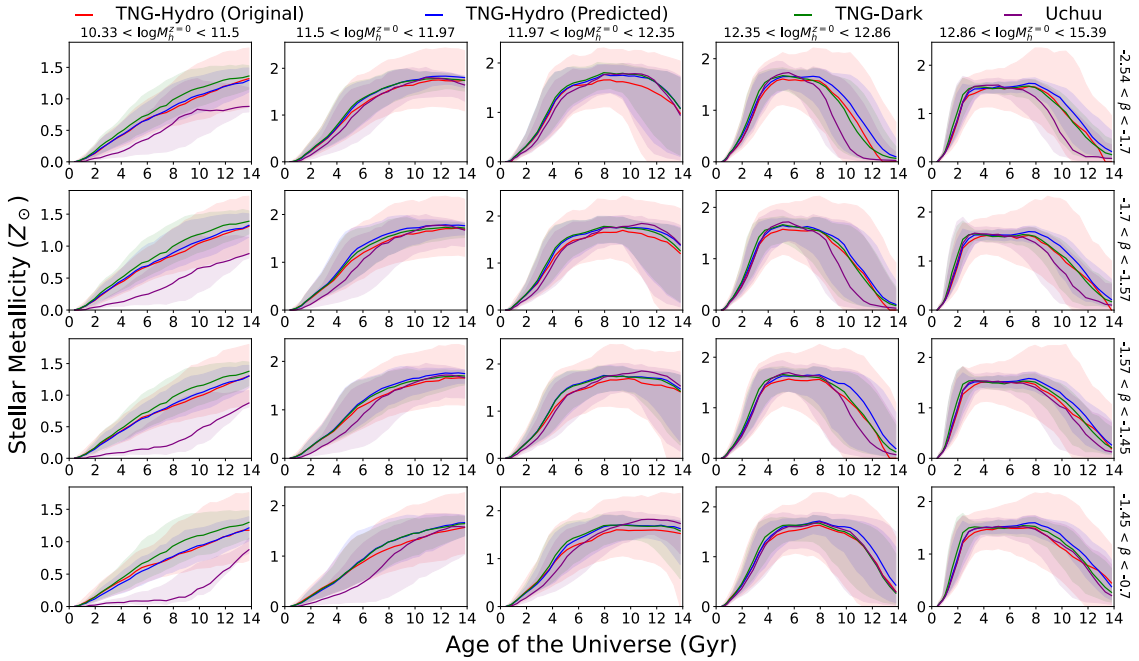


Figure 5.12: Metallicity histories of central galaxies, tabulated by halo mass and specific mass accretion gradient. This shows similar characteristics to the star formation histories in fig. 5.8, yet the low mass Uchuu samples are of particularly poor quality. However, the metallicity histories derived from dark simulations are generally very similar to the hydrodynamical predictions.

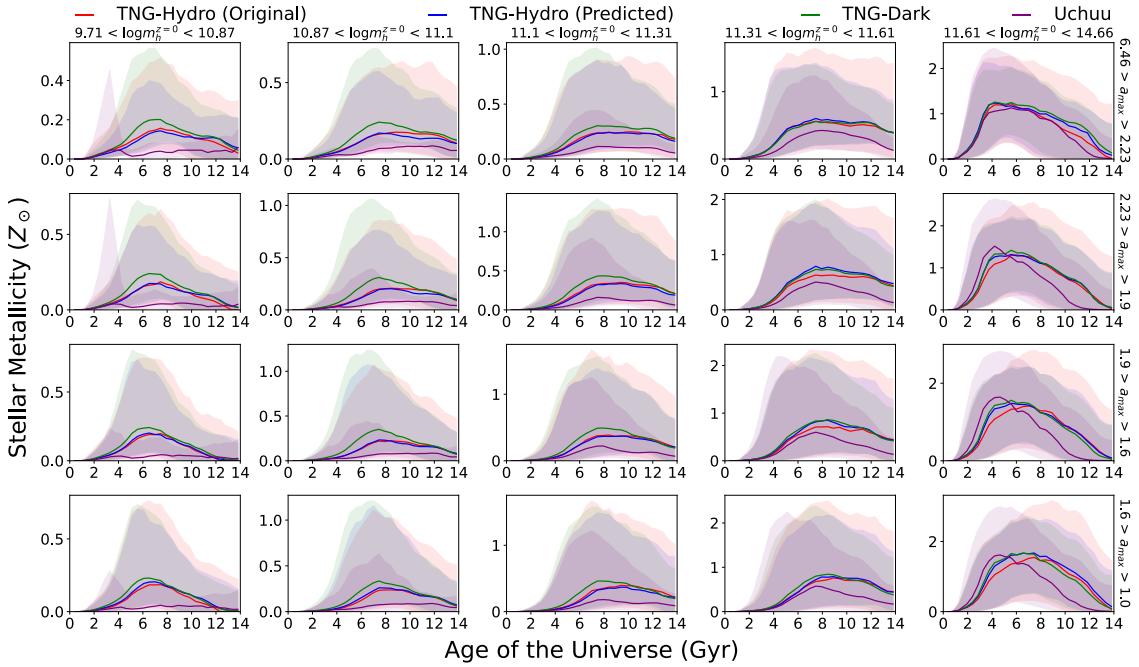


Figure 5.13: Metallicity histories of satellite galaxies, tabulated according to subhalo mass and scaled accretion time. These show similar characteristics to star formation histories in fig. 5.9, with overprediction in TNG-Dark and underprediction in Uchuu.

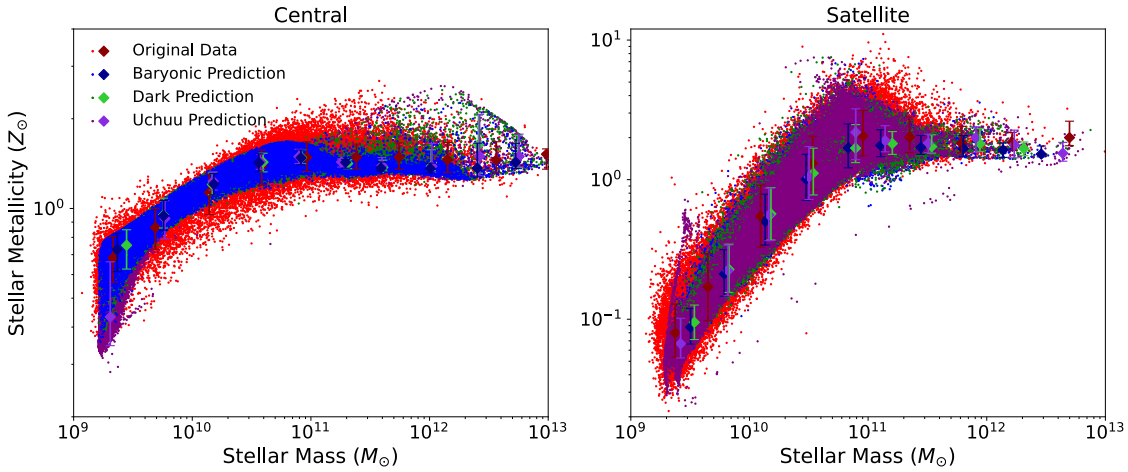


Figure 5.14: The numerical mass-metallicity relation is presented for central galaxies in the left panel and satellite galaxies in the right panel, with total stellar mass and mass-weighted metallicity values obtained by use of eqs. (2.24) and (2.27).

tion. It is also possible to explain the exaggerated metal synthesis in low mass TNG-Dark galaxies by the absence of stellar feedback, which would cause stars to retain more of their mass and synthesise metals more effectively.

Similar but weaker relationships between stellar metallicity and structure and density parameters are observed in small mass bins. However, our findings discussed in section 3.4 imply that chemical enrichment is influenced significantly by environmental history. We demonstrate that the calculated overdensities in Uchuu are slightly higher than in TNG, which can affect metallicities by foretelling an excess of mergers and flybys that redistribute the metals into high mass galaxies and thus contribute to quenching. Radial skews, which are used to track the anisotropic nature of these interactions, are demonstrated to have a significant impact on metallicity history. These are difficult to compare between simulations due to the lack of correlation with other halo properties, however the low number density of halos surrounding low mass objects can fail to produce the high skews encountered during close encounters of multiple halos, thereby bypassing some potentially significant interactions which enrich the interacting galaxies.

By analysing the mass-metallicity relations derived from these star formation and metallicity histories in fig. 5.14, we can see how both relations are distorted at low mass by the underpredicted metallicity histories. However, the dark simulation's shape and scatter of the relations closely resemble the network's initial TNG-Hydro predictions. The dark simulations at high mass do contain a few overpredicted metallicities for centre galaxies, though. For Uchuu galaxies with stellar masses above $10^{11.5} M_\odot$, a Spearman coefficient

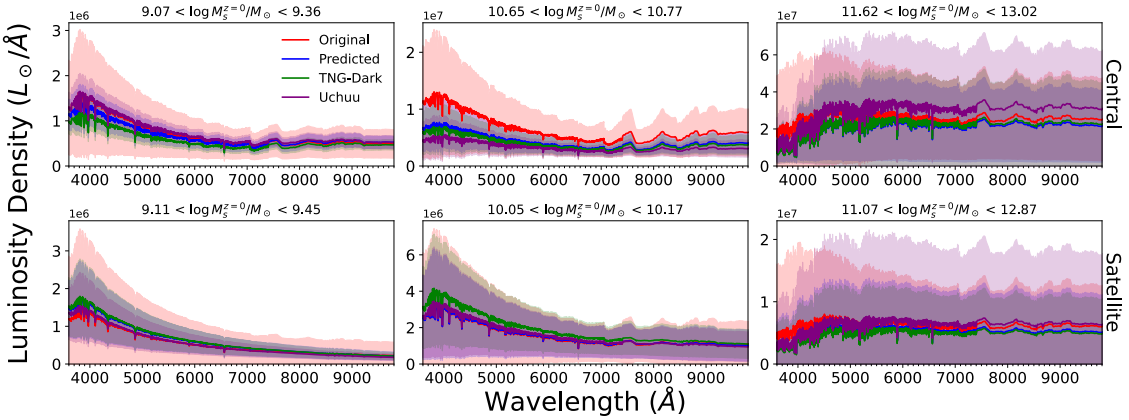


Figure 5.15: The mean and standard deviation of the evaluated spectra from four simulation datasets are presented in three bins of low, intermediate, and high stellar mass. The top row displays the results for central galaxies, while the bottom row shows those for satellite galaxies. The comparison demonstrates that the two dark simulations for satellite galaxies are well-matched to the hydrodynamic TNG simulation, as they exhibit similar means and variances in the spectra. In contrast, the agreement for central galaxies is less robust, especially for high mass galaxies, where the Uchuu spectra have a higher variance, indicating less constrained stellar mass. To clearly display the mean continuum from each simulation, emission lines are omitted from these spectra.

of 0.683 between age and metallicity demonstrates that these are the same galaxies whose mass-weighted ages are overestimated. As a result, the results have a bigger contribution from early metallicity histories and a smaller contribution from times of low star formation.

5.4.3 Spectroscopy

The spectral energy distributions of the four simulation datasets are presented in bins of stellar mass for both central and satellite galaxies in fig. 5.15. As discussed in section 4.3.1, the smaller variance of the predicted SEDs can be attributed to the absence of variability in star formation histories and implicit features such as merger-driven starbursts and quenching timescales. These factors are more prevalent in central galaxies than for satellite galaxies.

The mean amplitudes of TNG-Dark spectra exhibit subtle differences for low and intermediate mass central galaxies, being slightly smaller than average, and for satellites, being slightly larger. The increased amplitude for satellites can be attributed to higher peaks in star formation histories, as discussed in section 5.4.1. This effect impacts both networks, but it is particularly notable for satellites, which lack environmental harassment in the satellite phase (Engler et al., 2020) and do not experience mass loss due to feedback (Sorini et al., 2022). The reason for the amplitude offset in central galaxies is unclear,

but it may be due to differences in star formation histories and mass-weighted ages. The offset could result from the underpredicted star formation histories observed in Uchuu predictions.

The Uchuu spectra exhibit lower amplitudes in relation to TNG-Dark spectra, primarily due to lower star formation histories. Low mass central galaxies, on the contrary, have excess emission and are likely offset due to poorly predicted metallicity histories. However, in high mass bins, the variance in Uchuu spectra exceeds that of any TNG data, due to larger variance in mass accretion histories. The significant spread of Uchuu overdensities in high mass bins can deceive the neural network, either enhancing or suppressing star formation due to the association of overdensity with merger events or environmental quenching. While the Spearman coefficient between zero-redshift overdensity and each band magnitude is weak for the full sample, it is approximately -0.43 on average for galaxies above $10^{11.5} M_{\odot}$.

5.4.4 Photometry

Figure 5.16 displays the colour-mass diagrams obtained from the four simulation datasets, highlighting the relationship between the stellar mass and colours computed using neighboring SDSS wavebands, for both central and satellite galaxies. As established in section 4.3.4, these diagrams exhibit the bimodal colour distributions of the galaxy populations and the inclination for high-mass galaxies to be redder in colour, which is a consistent feature across all predictions of the neural network. However, this study also revealed that the unconstrained emission at UV frequencies has resulted in underpredicted colours for u and g bands, which is evident in the dark simulations.

The dark simulations show some notable differences in their predictions compared to the hydrodynamical simulation. In both dark simulations, the ages of high mass galaxies are biased towards higher values as depicted in fig. 5.11, leading to a higher fraction of red galaxies; particularly in $u - g$ and $g - r$ colours, and particularly for Uchuu galaxies with prematurely declining star formation. This overabundance of red galaxies could be due to the halo morphology in dark simulations, or the denser environments quenching interacting or infalling galaxies; as discussed in section 5.4.1.

Another feature of the dark predictions is a greater abundance of galaxies in the

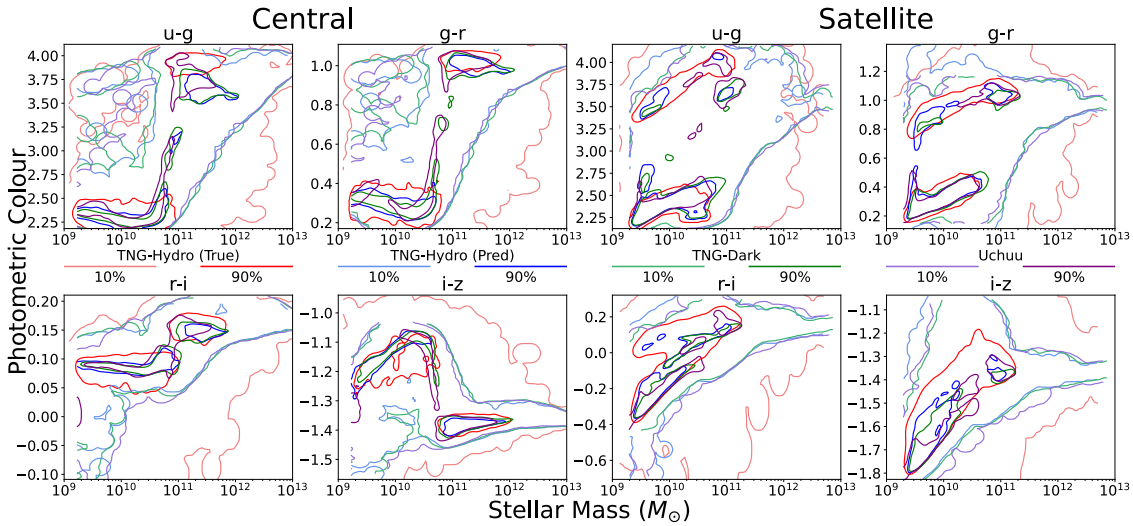


Figure 5.16: Colour-Mass diagrams from the four simulation datasets, using four colours defined as the difference between successive SDSS band magnitudes, are shown for central galaxies (left panels) and satellite galaxies (right panels). The contour lines represent the tenth percentiles (light-coloured lines) and ninetieth percentiles (dark-coloured lines) of the 2D histograms. All four datasets show the anticipated photometric colour bimodality and its relationship with mass. The dark simulations, however, show an excess of samples between the peaks of the colour distributions.

'green valley' region, particularly in Uchuu. This may be due to variations in the halo mass accretion history and internal dynamics, leading to a broader range of calculated magnitudes in this mass range. It has been shown in fig. 4.9 that the variability in star formation history is an important factor in modelling photometry, and the star formation histories in Uchuu may have lost additional high frequency information from being based on temporal predictors that were interpolated over a sparser time domain, shown by fig. 5.2.

5.5 Future Model Amendments

The performance of the machine learning model, used to predict the star formation and stellar metallicity histories of TNG galaxies, has been positive in reproducing comparable results in pure dark matter simulations. The model has established similar quantitative relations between galaxies and halos, and has connected physical galaxy properties to observational quantities as in the hydrodynamical simulation. Nonetheless, inconsistencies exist between the predictions of the two simulations, which result from discrepancies in the growth and interactions of halos in N-body simulations in contrast to the complete cosmohydrodynamical model, and in the calculation of halo properties in the simulation data. This section identifies these crucial differences and suggests how the model may be

modified in the future to adapt to high volume N-body simulations, while still accurately characterising the connection between galaxies and halos.

5.5.1 Simulation Resolution

Our findings indicate that the Uchuu simulation, with its lower resolution compared to the TNG simulations, has negatively impacted the accuracy of predictions for low mass and slowly accreting galaxies. Notably, there are significant errors in the masses and metallicities of low mass galaxies, and especially satellite galaxies, which were already reduced in number by our quality cuts. One of the contributing factors is the power spectrum effectively being truncated by the lower resolution of Uchuu, which hinders the establishment of bound structures from small groups of simulation particles, leading to delayed growth of the lowest mass halos. As a result, low mass samples in Uchuu are not considered practical for studying the galaxy-halo connection.

The issue of computational resource limits in generating high-volume, high-resolution simulations has persisted in this field. Li et al. (2021) have introduced a machine learning model which can address this problem, by enhancement of the matter power spectrum in N-body simulations, which Ni et al. (2021) show can predictively emulate precise snapshot data from low-resolution simulation images, accurately reproducing halo substructures and correlation functions beyond the resolution limit of a coarse N-body simulation. This method could potentially enhance the Uchuu data in this study, provided that the machine learning model can construct complete merger trees accordingly. It may be necessary to develop a similar enhancement model for the merger trees themselves, given the importance of the properties of progenitor subhalos, such as the mass and metallicity of incoming gas and stars. Nevertheless, since the matter power spectra, distributions, and velocity fields can be enhanced through this approach, these properties may be deduced from the environment generated by the model.

Our findings reveal that the difference in resolution has a significant impact on halos in Uchuu that have similar masses as the lowest mass halos in TNG. Therefore, it might be beneficial in a future study to utilise the neural network on Shin-Uchuu: a smaller but higher resolution version of the Uchuu model, to recover low mass galaxies. Alternatively, the Uchuu snapshots and merger trees could be improved by leveraging the advanced resolution of Shin-Uchuu and the techniques proposed by Li et al. (2021). Since the

distribution of galaxies and the characteristics of their surroundings are influenced by the growth of large-scale structure, the latter approach could be more appropriate for producing self-consistent mock catalogues at both gigaparsec and sub-kiloparsec scales.

5.5.2 Input Parameters

Our neural network was designed to employ variables which could describe the growth, structure, and surroundings of dark matter halos, while being resistant to the influence of baryons, resolutions, and halo finders. Nonetheless, our findings have revealed that these differences have resulted in some degree of variation in the predicted galaxy statistics between pure dark matter simulations and the hydrodynamical TNG model, ranging from minor to significant. As a result, it may be appropriate to implement currently unused variables into the neural network model in the future, which could represent more appropriate indicators of the galaxy-halo relationship for pure dark matter models.

In order to eliminate inaccuracies in predictions due to the differing identification of halo mass and substructure, it may be necessary to utilise a consistent halo finder in both the hydrodynamical and N-body simulations. Although several halo finder algorithms accurately recognise halo structures, only the Rockstar algorithm accurately identifies substructures in highly dense regions such as the halo centre (Onions et al., 2012). These substructures can influence the structural variables that impact the stellar mass of both central and satellite galaxies. In future research, an "Uchuu-friendly" neural network model could be trained based on the planned TNG halo catalogue based on the Rockstar halo finder, eliminating this potential source of error.

In this study, another variable which varies between the simulations is the overdensity. In Uchuu, the overdensity is calculated using halo coordinates and masses, whereas TNG uses subhalo information. Rather than the direct use of halo tracers, and instead creating a continuous, position-dependent density field similar to Chen et al. (2020), overdensities may be smoothed with a Gaussian kernel, which can be adjusted for each simulation's resolution or density tracers. Chen et al. (2020) also utilised this method to create a position-dependent tidal field tensor, which is shown to have a strong correlation with halo assembly bias.

The behaviour of satellite galaxies is largely influenced by the environment of their

host halo, making the properties of the host halo important during the satellite phase. Inclusion of additional properties of the target halo could have improved the model, such as formation time (Artale et al., 2018), angular momentum (Bose et al., 2019), and halo-satellite separation (Engler et al., 2020; Montero-Dorta et al., 2022). The location of the satellite in relation to the host or a central halo in relation to a local cluster can provide an indication of the local environment, which could be a valuable measure of effects such as gas stripping and tidal disruption.

Since the skew parameter is not significantly different between the simulations considered in this work, and is independent of halo mass, it may be feasible to use a position or velocity-dependent environmental parameter to address the bias introduced by using over-densities to characterise local environments. However, the skew parameter still provides valuable information on halo-halo interactions over time and should be retained in the model. Nonetheless, the lack of low-mass objects in low-resolution simulations could affect the identification of major interactions based on high or low skew values, as discussed in section 5.4.2.

5.5.3 Galaxy Clustering

Further to the population statistics derived from star formation and metallicity histories, a necessary measure of the galaxy-halo connection to complement deep galaxy surveys is the clustering signal of galaxies and halos. By measuring the spatial correlation statistics of galaxies of different mass, luminosity or other observationally determined properties, the role of the galaxy-halo connection in galaxy evolution can simultaneously be measured on local and cosmological scales. Future surveys which measure these clustering statistics out to higher redshifts will investigate precisely the growth history of cosmological structure and the role of these environments in shaping the distribution of galaxies, hence it is imperative to replicate these effects on all scales to reconcile the galaxy evolution models with high fidelity surveys.

The neural network as it currently stands is likely to predict basic clustering information; as well as including environmental histories, the model includes variables which directly correlate with clustering statistics. Montero-Dorta et al. (2021) show that the specific mass accretion gradient β is a superior metric for halo and galaxy assembly bias than the traditionally used half-mass formation time, for all masses and redshifts in TNG.

Furthermore, including measures of the halo's profile and virial velocity may prove valuable to predicting secondary assembly bias (Salcedo et al., 2018; Hadzhiyska et al., 2020). However, it has been established that these input quantities are influenced by the resolution of the simulation, which in practice can affect the accuracy of the galaxy assembly bias predicted in Uchuu. Furthermore, given the sensitivity of clustering to halo concentration (Wechsler et al., 2006), a quantity characterising the profile that is fit to the halo, the use of a common halo finder algorithm may prove necessary to reproduce this precisely in future mocks; in spite of its weak influence on the halo's half-mass radius, discussed in section 5.3.2.

5.5.4 Comparison With Other Models

Aung et al. (2022) exploit the UniverseMachine (UM) model (Behroozi et al., 2019a) to calculate galaxy formation histories in the Uchuu simulation, generating statistics such as stellar mass functions and number density profiles, which Behroozi et al. (2019b) show to display reasonable consistency with observational stellar mass and luminosity functions.

UM is an empirical model which relies on MCMC optimization of star formation rates, utilising prior relations between star formation rates and quenched fractions based on the maximum circular velocity of the halo. It is able to qualitatively reproduce the environmental dependence of star formation without the explicit inclusion of the environment, suggesting that halo mass accretion is the sole factor contributing to star formation in dense environments.

Our self-consistent model has the potential to be more valuable for high fidelity mocks than empirical models like UM. It enables causal modeling of galaxy growth over time, driven by halo and environmental factors, and can be modified to predict as-yet unconsidered properties such as gas fractions and AGN growth over time. Our model may be better suited for modelling the dependence of chemical enrichment in high-fidelity mocks, as UM does not account for metallicity histories, which we have shown to be more reliant on environmental variables.

However, since quenched fractions are a direct parameter of the UM model, the colour bimodality in these mocks is likely to be more accurate, making UM more suitable for observational studies. Our machine learning model has also shown to be sensitive to certain

effects resulting from the transition to a pure dark matter simulation, such as biased growth of internal dynamics, resolution issues like delayed collapse and virialisation of haloes, and potential discrepancies in the calculation of key variables such as halo mass and radius due to differing halo finder algorithms. Therefore, a high-resolution N-body simulation with consistent halo definitions may be required to produce accurate self-consistent mocks.

5.6 Conclusion

This chapter focuses on comparing the accuracy of predictions for galaxy star formation and metallicity histories by applying the semi-recurrent neural network described in chapter 2 to pure dark matter simulations. Predictions from the hydrodynamical TNG simulation are compared with cross-matched haloes from their dark equivalents, which lack baryonic processes, to assess the impact of their absence on predictions. Secondly, the model is applied to similar halos from the Uchuu N-body simulation to examine the effects of alternative halo definitions and the lower mass resolution of the Uchuu simulation. The chapter's findings can be summarised as follows:

1. For halos of most masses and mass accretion gradients, the important input properties such as the mass accretion history of a halo are similar for both hydrodynamical and dark simulations. However, figs. 5.3 and 5.4 demonstrate that the mass accretion histories of TNG-Dark halos are exaggerated, which may be due to the lack of stellar feedback contributing to the halo morphology. This has a noticeable, similar impact on the star formation histories of low mass galaxies, discussed in section 5.4.1. Since this difference in mass accretion histories occurs at high redshift, the resulting effect on the star formation histories applies for most of the simulation's time domain.

2. Quantities pertaining to the internal structure and dynamics of the halos, specifically the half-mass radius and circular orbital velocity of the halo, are affected in a similar manner by the absence of baryons; yet they are more significantly influenced by the lower resolution of the Uchuu simulation. This delay causes the halos' germination and initial growth to occur later and at a slower pace, and causes the mass accretion and concentration of Uchuu halos to appear smaller than their TNG counterparts. For slowly growing and low-mass halos, these impacts are substantial due to the sensitivity of these variables to the simulation resolution.

3. In sections 5.4.1 and 5.4.2, we demonstrate mostly comparable neural network predictions in the dark simulations to the original hydrodynamical predictions, where the SHMR and MZR exhibit a close resemblance in shape and scatter. Nevertheless, the severely underestimated growth of low mass halos in Uchuu results in inadequate predictions of the star formation and metallicity histories of low mass galaxies. Conversely, TNG-Dark outcomes are overestimated due to an excessive amount of mass accretion during early times. In both cases, the poorest predictions are for the lowest mass galaxies in any dark matter simulation, suggesting that resolution enhancement is important for modelling the complete population of high fidelity galaxy catalogues.

4. The neural network's predictions of the number of quenched galaxies, and in some cases the rate of declining star formation, are greater than the hydrodynamical results in both dark simulations. In TNG-Dark, this is due to the difference in structural parameters, which in hydrodynamical simulations control the AGN feedback which quenches galaxies. In Uchuu, this is a consequence of a higher range of overdensities resulting from the use of Rockstar halo tracers instead of SubFind subhalos. This excess quenching corresponds to a greater abundance of photometrically red galaxies. Furthermore, in Uchuu, the abundance of red galaxies is attributed to interpolation over a coarser time domain, resulting in more significant information loss regarding time variations in their star formation history. Other than this, the spectroscopic and photometric statistics of the dark matter simulations exhibit similar physical characteristics to original results, as demonstrated in sections 5.4.3 and 5.4.4.

The ability to construct accurate summary statistics regarding the physical and observational properties of galaxies using Uchuu data is promising; it suggests that the semi-recurrent neural network design may be used as the baseline model for a gigaparsec-volume galaxy catalogue, self-consistently entailing the causal aspects of the galaxy-halo connection from high redshifts to the present day. At present, the model may be able to compete with existing and widely used empirical models of the galaxy-halo connection in N-body simulations. In future, this methodology could pave the way for a future of AI-based simulations which offer a physical understanding of galaxy evolution at both the cosmological and substructure levels, alongside spectroscopic data to complement the most ambitious galaxy surveys to date. The vastness of a complete Uchuu galaxy dataset

may be used to produce a sizeable sample of rare objects such as early massive quiescent galaxies, and discern the nature of groups of small objects such as dwarf galaxies based on their local cosmology, as well as many further insights into the logistics of galaxy evolution beyond the realm of present-day cosmic simulations, at a fraction of the computational cost.

Despite this enthralling premise for a physically motivated galaxy evolution dataset extending from sub-kiloparsec to gigaparsec scales, we have presented a series of flaws with the application of the model to pure dark matter simulation data. First of all, the properties of galaxies are sensitive to the spacial and temporal resolutions of the simulation; particularly those hosted by small, slowly developing halos. To model small-scale galaxy growth accurately, the resolution of the simulation will need to be augmented, which, thankfully, may also be achieved using machine learning. Secondly, various halo and environmental variables have proven to be affected by the halo finder algorithm or by the available data to compute the necessary quantities. It may prove decisive to the performance of a model inspired by ours to create alternative parameterisations of the halo and environment which are immune to the logistics of different N-body simulation codes, or to introduce fine-tuned modifications such as smoothed density fields to eliminate these effects. We also stress that the data presented in this chapter is not subject to the stochastic correction entailed in section 3.3, and while this has proven fruitful in amending errors in the fiducial network predictions in TNG, the predicted Fourier transforms may be subject to similar errors in Uchuu, hindering the utility of the correction in Uchuu mocks. Adjusting the model to account for methodical differences between N-body models will require a detailed comparison of halo and environmental relationships in separate pure dark matter simulations.

6

Conclusions & Outlook

In this thesis, an artificial neural network was designed to create a predictive machine learning model, capable of predicting the evolutionary histories of galaxies by encoding the galaxy-halo connection in cosmic hydrodynamical simulations. Using these results, we have analysed the historical and present-day properties of halos and cosmic environment which describe key statistics of the galaxy-halo connection, computed spectroscopic and photometric data to assess the quality of mock surveys generated by this model, and tested the model on pure dark matter simulations to assess the robustness of the neural network to changes in the simulation model. This research demonstrates the potential of a machine learning model to compute galaxy properties using high volume N-body simulation data, which could introduce a new approach to studying galaxy evolution on large and small scales.

6.1 Chapter Summaries

Chapter 2: Data, Design & Preprocessing

In this chapter, we introduce the hydrodynamical and pure dark matter simulations used in this study, describing in detail the publicly available catalogues of halo and galaxy properties which can be utilised, and our methods of calculating secondary variables such as star formation histories and environmental densities and skews. We also introduce the design of a semi-recurrent neural network, with the intent to use a combination of temporal and static input variables to predict galaxy evolution, and describe the necessary numerical and physical preprocessing of the data for use in the neural network.

We compute historical halo properties along the main progenitor branch of each sample's merger tree, and environmental properties using the GriSPy periodic nearest-neighbour search algorithm developed by Chalela et al. (2021). Our key baryonic targets, the star formation and metallicity histories of galaxies, are computed using the mass and metallicity weighted stellar age spectra of all stellar particles bound to the halo. We apply a multiplicative correction as a function of halo mass and time to lower resolution data, and apply a quantile transformation to most data in the model, including a method of normalisation accounting for the properties of temporal quantities.

We show in this chapter that the data adjustments which we implement result in concurrent galaxy properties between the TNG100 and TNG300 data suites, which due to differences in resolution, have different star formation efficiencies (Pillepich et al., 2017a). We also show that our choice of quantile transformation to a Gaussian distribution is a practical means to handle sparsely distributed data such as halo mass accretion rates, while offering two methods of normalising temporal data for different calculations by the network. Our neural network design is capable of implementing multiple static quantities and several temporal variables without exceedingly many degrees of freedom, which we show in the following chapter to be superior to a more conventional design.

Chapter 3: Neural Network Predictions

In this chapter, we evaluate the quality of predictions of the neural network in relation to the original TNG simulation data. We show here that accurate statistics such as the

stellar-halo mass relation and stellar mass-metallicity relation can be derived from the predicted star formation and metallicity histories, further to the development of these historical properties at different times being accurately predicted as a function of halo mass. This indicates that galaxy evolution can be predicted from historical halo data using our artificial neural network.

However, the neural network is less capable of predicting variability in these star formation histories, which have negative consequences on the calculated scatter in stellar mass and metallicity. We use an identical neural network to predict the absolute Fourier transforms of these star formation and metallicity histories, and use these predictions to compute a stochastic correction, to improve the predicted star formation and metallicity histories. This method has made notable improvements to the key galaxy-halo relations and to the predicted geometry of the star formation histories, yet it does not fully reproduce the galaxies' formation histories due to the difficulty in modelling rare, high mass star formation events.

We assess the importance of input variables in neural networks for replicating galaxy-halo statistics by replacing groups of variables with random noise, identifying their impact on the model's predictions. We find that variables related to halo mass and substructure predict the scatter of the stellar-halo mass relation, while environmental factors correlate with the mass-metallicity relation. Previous studies focusing on a single redshift struggled to measure these effects, highlighting the value of historical input data. While the bulk of these statistics can be inferred with relatively few parameters, we show that specific parameters dominate the influence over other physically related variables; which implies that the model can identify significant parameters in the galaxy-halo connection.

Chapter 4: Observables

In this chapter, we apply the Flexible Stellar Population Synthesis code developed by Conroy et al. (2009); Conroy & Gunn (2010) to calculate spectroscopic and photometric data from the predictions of the neural network detailed in the previous chapter. Additionally, we evaluate the quality of improvement made to observational galaxy statistics by the aforementioned stochastic correction to our predicted star formation and metallicity histories. These results show that observational statistics such as the photometric colour bimodality of the galaxy population can be predicted indirectly by the neural network.

The shortcomings of the predicted galaxy formation histories, such as the lack of high frequency features, have influenced the computed observational statistics, resulting in less variance in luminosities and colours in most mass bins. The stochastic correction has made considerable improvements, producing more accurate spectral amplitudes in bins of stellar mass, in accordance with improved scatter in stellar mass and metallicity. However, just as the correction has limited ability to predict missing high mass features, the true diversity of galaxy spectra is not fully recovered. While improvements of spectral amplitude and emission line luminosity are significant, photometry is not largely affected by the correction.

Chapter 5: Dark Simulations

Finally, we apply the neural network to the TNG-Dark and Uchuu dark matter simulations, where the similarity of predictions and statistics between these two simulations and the hydrodynamical simulation data is used to gauge the suitability of the model for use in high volume N-body simulations; in this case, the Uchuu simulation. By comparing previous results with those from the TNG-Dark simulation, we measure the effects that the lack of baryonic physics in the simulation has on the input and output variables to the model. By comparing all of the above with predictions based on Uchuu data, we investigate the effects of lower simulation resolution and alternative halo structure properties on the predictions of the neural network.

Important input properties such as mass accretion history are similar in both dark and hydrodynamical simulations. However, the mass accretion histories in TNG-Dark halos are exaggerated due to the absence of stellar feedback. Quantities related to the internal structure and dynamics of halos, like the half-mass radius and circular orbital velocity, are also affected by the absence of baryons and lower resolution in the Uchuu simulation. This delay causes slower growth and smaller mass accretion and concentration in Uchuu halos compared with TNG halos. This has a noticeable impact on the predicted star formation histories of low mass, slowly accreting galaxies, suggesting the importance of resolution enhancement for accurate modeling of the complete population. Observationally, differences in formation histories in dark simulations such as excess quenching results in a greater abundance of red galaxies.

We argue that the model and the N-body simulation data can be augmented in a number of ways to minimise the discrepancies we have shown between the data produced by the neural network in hydrodynamical and dark simulations. First of all, a future rendition of the neural network may impose alternative quantities, such as a smoothed density field or distance from a host halo as measures of cosmic environment which do not depend on the resolution of the simulation or the halo properties acquired by the algorithm identifying halos and constructing merger trees. Secondly, as the simulation resolution is an important factor in shaping the model's predictions, the simulation data may be enhanced using machine learning techniques such as those employed by Li et al. (2021).

6.2 Research Outlook

The work presented in this thesis has shown that a predictive machine learning model designed to emulate galaxy formation histories based on halo catalogues and merger trees is capable of returning physically congrous and meaningful results in N-body simulations, in spite of methodical differences between the training and testing data. While certain corrections will be necessary to emulate reliable data on all mass scales, this work represents the foundation of a method of efficiently producing intricate and comprehensive galaxy catalogues, entailing the physics of the galaxy-halo connection across cosmic time, and supporting galaxy surveys intending to test theories of galaxy evolution on such scales in the real universe. Here we contemplate potential applications of such a design in future research in galaxy evolution and cosmology.

Exploring The Galaxy-Halo Connection

While the predictions of the model may have some limitations in the level of predictable detail, this presents an opportunity to study the contributions of evolutionary processes. The stochastic correction method outlined in this thesis may be developed to implement the correlation between phases of star formation and metallicity histories, providing some insight into the nature of delays between star formation events and metal synthesis. These fluctuations may be correlated with input features once accurate observables are made. Having predicted the power spectrum of star formation and metallicity histories using the neural network design, a feature importance test could inform the halo and environment

data which influences galaxy evolution on different timescales. Iyer et al. (2020) show how such power spectra are related to baryonic processes such as AGN feedback, galactic winds and mergers, which could be explained in part by the galaxy-halo connection.

Despite showing that the temporal galaxy-halo connection can be learned via artificial intelligence, the relationship remains complex and non-intuitive. However, by application of the model to a gigaparsec-scale N-body simulation such as Uchuu, a vast dataset will be generated which can be used to gain an explanation of the exact role of halo and environmental properties in shaping galaxies over time. (Wadekar et al., 2020) use symbolic regression techniques to derive analytical expressions for HI mass and assembly bias as a function of environment and halo substructure per TNG snapshot, which can be reconciled with empirical HOD models. As the relationships between galaxy and halo properties become more convoluted, however, larger datasets will be required to generalise this to mutliple regimes of galaxy-halo coevolution. Alternatively, Shapley explainability modelling (Lundberg & Lee, 2017) can evaluate the driving parameters of the GHC across time; for individual samples and for classifications of galaxies, such as galaxies of a specified mass or morphology.

A catalogue of synthetic galaxies in a high fidelity dark matter simulation may be very practical for testing the role of dark matter in cosmological star formation, such as explaining the trends of star formation rates with halo mass and age in SDSS (Scholz-Díaz et al., 2022a,b). Furthermore, a similar model may be designed to predict the evolution of interesting galaxy properties, such as the kiloparsec-scale kinematics of star formation gas, believed to play a key role in seeding rapid star formation (Förster Schreiber & Wuyts, 2020). However, cosmological simulations tend to produce distinct galaxy statistics due to differences in the subgrid models governing gas thermodynamics and star formation, and different choices of cosmological parameters. Thus far, this neural network has been based on a single simulation, yet TNG can be less similar to observations than other simulations, in aspects such as green valley quenching (Angthopo et al., 2020), cold gas fractions (Davé et al., 2020) and cosmic star formation rate density (Yates et al., 2021). It may prove insightful to train the neural network on multiple simulations to optimise these subgrid and cosmic parameters to conform to these observations, thereby illustrating the importance of baryonic processes in hydrodynamical models; and perhaps produce a suite of large simulations with varying parameters similar to the CAMELS suite (Villaescusa-

Navarro et al., 2021), which may inform the role of these processes on scales beyond the 25 Mpc/h limit of CAMELS.

High Redshift Galaxies

In addition to the ongoing high volume surveys such as that conducted by the DESI Collaboration (2016), the advent of deep surveys such as those conducted by JWST (Dunlop et al., 2021; Malkan et al., 2021) has opened the possibility to survey galaxies at increasingly high redshifts. In particular, JWST has the sensitivity necessary to capture resolved spectroscopy of galaxies forming in the epoch of reionisation: a critical epoch which marks the most recent major phase change in the universe's history. Star-forming galaxies have been identified in abundance at redshifts as high as $z \sim 8$, dominating the ionisation of the IGM, yet it is unclear how these galaxies evolved, how abundant they were at reionisation redshifts, and how they contributed to the rapid ionisation of the IGM (Robertson, 2022).

Modelling the evolution of galaxies in the reionisation epoch has been achieved with a specialised set of cosmological simulations, yet this has also required a tradeoff between small, resolved simulations such as SPHINX (Rosdahl et al., 2018) and large, coarse simulations such as BlueTides (Feng et al., 2015; Wilkins et al., 2017); or the use of zoom simulations such as FLARES (Lovell et al., 2020) which by construction under-resolve the surrounding large scale structure, as discussed in section 1.2.1. While these simulations reproduce accurate galaxy statistics such as UV luminosity functions and provide practical insights into the correlations between galaxy properties (e.g. stellar mass and metallicity), little is known about the detailed mechanics of early galaxy evolution.

Post-reionisation, there are some fascinating and uncommon entities that emerge. Among them are massive quiescent galaxies, which, based on accepted theories of galaxy evolution and observations since cosmic noon, become increasingly scarce as we look to higher redshifts (Brennan et al., 2015); and galaxy clusters, which are uncommon due to heirachical structure growth, yet serve as valuable indicators of spatial structure during earlier epochs and foster a distinctive environment that shapes the evolution of galaxies (Kravtsov & Borgani, 2012).

The growing frequency of the detection of these objects in infrared surveys highlights the necessity for an advanced model of massive galaxy evolution which accurately captures

their abundance throughout cosmic history. Moreover, these early quenched galaxies possess intriguing characteristics, such as their unusually compact nature given their mass (van Dokkum et al., 2008; Glazebrook et al., 2017). However, their rarity in modern deep surveys and the difficulty of identifying their emission lines pose challenges when attempting to measure these properties through observations. JWST is anticipated to observe the evolution of such galaxies beyond $z \sim 4$, a crucial period when substantial star formation activity in massive galaxies is believed to take place (Forrest et al., 2020).

With a machine learning model which can predict the evolution of early galaxies in relation to their halo mass accretion and large scale environment, the growth of galaxies in the early universe may be computed using a combination of high redshift simulations. By application of this model to a large N-body simulation, the evolutionary properties of massive galaxies and clusters and candidates for massive compact galaxies can be investigated. The complete Uchuu simulation, for example, contains $\sim 10^5$ merger trees of $\sim 10^{12} M_{\odot}$ up to $z \sim 4$, providing a great amount of diversity of accretion and interaction histories of these progenitors. With the forthcoming data from JWST, more of these rare and distant galaxies are expected to be identified, with enough spectroscopic and photometric data to infer their star formation histories, verifying the accuracy of the evolutionary model. Upon concluding this PhD, it is the evolution of early, massive galaxies in JWST which I will investigate as a PDRA at the Swinburne Centre for Astrophysics and Supercomputing in Melbourne, Australia.

Bibliography

Abadi, M. et al. 2016, arXiv e-Print:1603.04467

Agarwal, S., Davé, R., & Bassett, B. A. 2018, MNRAS, 478, 3410

Aggarwal, C. 2018, Neural Networks And Deep Learning (Cham, Switzerland: Springer-Verlag Publications)

Alpaydin, E. 2020, Introduction To Machine Learning, 4$^{\text{th}}$ edn. (Cambridge, MA, USA: MIT Press)

Anbajagane, D., Evrard, A. E., & Farahi, A. 2021, MNRAS, 509, 3441

Angthopo, J., Negri, A., Ferreras, I., de la Rosa, I. G., Dalla Vecchia, C., & Pillepich, A. 2020, MNRAS, 502, 3685

Angulo, R. E., Baugh, C. M., Frenk, C. S., & Lacey, C. G. 2014, MNRAS, 442, 3256

Artale, M. C., Zehavi, I., Contreras, S., & Norberg, P. 2018, MNRAS, 480, 3978

Aung, H. et al. 2022, MNRAS, 519, 1648

Baldry, I. K., Glazebrook, K., Brinkmann, J., Ivezic, Ž., Lupton, R. H., Nichol, R. C., & Szalay, A. S. 2004, ApJ, 600, 681

Bartelmann, M. 2010, Rev. Mod. Phys., 82, 331

Baugh, C. M., Lacey, C. G., Gonzalez-Perez, V., & Manzoni, G. 2021, MNRAS, 510, 1880

Behera, J., Chittenden, H. G., & Tojeiro, R. in prep.

Behroozi, P., Wechsler, R. H., Hearin, A. P., & Conroy, C. 2019a, MNRAS, 488, 3143

——. 2019b, MNRAS, 488, 3143

Behroozi, P. S., Wechsler, R. H., & Wu, H.-Y. 2012a, ApJ, 762, 109

Behroozi, P. S., Wechsler, R. H., Wu, H.-Y., Busha, M. T., Klypin, A. A., & Primack, J. R. 2012b, ApJ, 763, 18

Bennett, C. L. et al. 2013, ApJSS, 208, 20

Berlind, A. A., & Weinberg, D. H. 2002, ApJ, 575, 587

Berta, S., Lutz, D., Genzel, R., Förster-Schreiber, N. M., & Tacconi, L. J. 2016, A&A, 587, A73

Blecha, L. et al. 2015, MNRAS, 456, 961

Bluck, A. et al. 2020, MNRAS, 499, 230

Bluck, A. F. L., Piotrowska, J. M., & Maiolino, R. 2023, ApJ, 944, 108

Blumenthal, G. R., Faber, S. M., Primack, J. R., & Rees, M. J. 1984, Nature, 311, 517

Boecker, A., Neumayer, N., Pillepich, A., Frankel, N., Ramesh, R., Leaman, R., & Hernquist, L. 2022, MNRAS, 519, 5202

Boggess, N. W. et al. 1992, ApJ, 397, 420

Bond, J. R., Kofman, L., & Pogosyan, D. 1996, Nature, 380, 603

Bose, S., Eisenstein, D. J., Hernquist, L., Pillepich, A., Nelson, D., Marinacci, F., Springel, V., & Vogelsberger, M. 2019, MNRAS, 490, 5693

Boylan-Kolchin, M., Springel, V., White, S. D. M., Jenkins, A., & Lemson, G. 2009, MNRAS, 398, 1150

Brennan, R. et al. 2015, MNRAS, 451, 2933

Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000

Byler, N., Dalcanton, J. J., Conroy, C., & Johnson, B. D. 2017, ApJ, 840, 44

Castignani, G. et al. 2022, ApJSS, 259, 43

Castro, T., Borgani, S., Dolag, K., Marra, V., Quartin, M., Saro, A., & Sefusatti, E. 2020, MNRAS, 500, 2316

Chabrier, G. 2003, PASP, 115, 763

Chalela, M., Sillero, E., Pereyra, L., Alejandro García, M., Cabral, J. B., Lares, M., & Merchán, M. 2021, Astronomy & Computing, 34, 100443

Chaves-Montero, J., Angulo, R. E., Schaye, J., Schaller, M., Crain, R. A., Furlong, M., & Theuns, T. 2016, MNRAS, 460, 3100

Chaves-Montero, J., & Hearin, A. 2021, MNRAS, 506, 2373

Chen, Y., Mo, H. J., Li, C., Wang, H., Yang, X., Zhang, Y., & Wang, K. 2020, ApJ, 899, 81

Child, H. L., Habib, S., Heitmann, K., Frontiere, N., Finkel, H., Pope, A., & Morozov, V. 2018, ApJ, 859, 55

Chittenden, H. G., & Tojeiro, R. 2022, MNRAS, 518, 5670

Chittenden, H. G., Tojeiro, R., & Kraljic, K. in prep.

Choi, J., Dotter, A., Conroy, C., Cantiello, M., Paxton, B., & Johnson, B. D. 2016, ApJ, 823, 102

Chruslinska, M., & Nelemans, G. 2019, MNRAS, 488, 5300

Chua, K. T. E., Pillepich, A., Vogelsberger, M., & Hernquist, L. 2019, MNRAS, 484, 476

Chua, K. T. E., Vogelsberger, M., Pillepich, A., & Hernquist, L. 2022, MNRAS, 515, 2681

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. 2015, arXiv e-Print:1511.07289

Conroy, C. 2013, ARA&A, 51, 393

Conroy, C., Gunn, J., & White, M. 2009, ApJ, 699, 486

Conroy, C., & Gunn, J. E. 2010, ApJ, 712, 833

Correa, C. A., & Schaye, J. 2020, MNRAS, 499, 3578

Cui, W., Davé, R., Peacock, J. A., Anglés-Alcázar, D., & Yang, X. 2021, Nat. Astr., 5, 1069

Curtis-Lake, E. et al. 2012, MNRAS, 429, 302

Davé, R., Anglés-Alcázar, D., Narayanan, D., Li, Q., Rafieferantsoa, M. H., & Appleby, S. 2019, MNRAS, 486, 2827

Davies, J. J., Crain, R. A., Oppenheimer, B. D., & Schaye, J. 2019, MNRAS, 491, 4462

Davies, J. J., Pontzen, A., & Crain, R. A. 2022, MNRAS, 515, 1430

Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371

Davé, R., Crain, R. A., Stevens, A. R. H., Narayanan, D., Saintonge, A., Catinella, B., & Cortese, L. 2020, MNRAS, 497, 146

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. 2020, Mathematics For Machine Learning (Cambridge, UK: Cambridge University Press)

DESI Collaboration. 2016, arXiv e-Print: 1611.00036

Dhoke, P., & Paranjape, A. 2021, MNRAS, 508, 852

Dodelson, S. 2003, Modern Cosmology (Amsterdam, The Netherlands: Academic Press)

Donnan, C. T., Tojeiro, R., & Kraljic, K. 2022, Nat. Astr., 6, 599

Donnari, M. et al. 2020, MNRAS, 500, 4004

Donnari, M., Pillepich, A., Nelson, D., Marinacci, F., Vogelsberger, M., & Hernquist, L. 2021, MNRAS, 506, 4760

Dooley, G. A., Griffen, B. F., Zukin, P., Ji, A. P., Vogelsberger, M., Hernquist, L. E., & Frebel, A. 2014, ApJ, 786, 50

Duckworth, C., Starkenburg, T. K., Genel, S., Davis, T. A., Habouzit, M., Kraljic, K., & Tojeiro, R. 2020, MNRAS, 495, 4542

Duckworth, C., Tojeiro, R., & Kraljic, K. 2019, MNRAS, 492, 1869

Dunlop, J. S. et al. 2021, PRIMER: Public Release IMaging for Extragalactic Research, JWST Proposal. Cycle 1, ID 1837

Einasto, J. 1965, Trudy Astrofizicheskogo Instituta Alma-Ata, 5, 87

Engler, C. et al. 2020, MNRAS, 500, 3957

Euclid Collaboration. 2013, LRR, 16

Falcón-Barroso, J., Sánchez-Blázquez, P., Vazdekis, A., Ricciardelli, E., Cardiel, N., Cenarro, A. J., Gorgas, J., & Peletier, R. F. 2011, A&A, 532, A95

Feng, Y., Di-Matteo, T., Croft, R. A., Bird, S., Battaglia, N., & Wilkins, S. 2015, MNRAS, 455, 2778

Ferrero, I. et al. 2021, A&A, 656, A106

Forrest, B. et al. 2020, ApJ, 903, 47

Fraser, T. S., Tojeiro, R., & Chittenden, H. G. 2022, MNRAS

Frenk, C., & White, S. 2012, Annalen der Physik, 524, 507

Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., & Schneider, D. P. 1996, AJ, 111, 1748

Fukugita, M., & Peebles, P. J. E. 2004, ApJ, 616, 643

Förster Schreiber, N. M., & Wuyts, S. 2020, ARA&A, 58, 661

Galárraga-Espinosa, D., Garaldi, E., & Kauffmann, G. 2023, A&A, 671, A160

Gallazzi, A., & Bell, E. F. 2009, ApJs, 185, 253

Gallazzi, A., Charlot, S., Brinchmann, J., White, S. D. M., & Tremonti, C. A. 2005, MNRAS, 362, 41

Galárraga-Espinosa, D., Garaldi, E., & Kauffmann, G. 2023, A&A, 671, A160

García, R., & Rozo, E. 2019, MNRAS, 489, 4170

Genel, S. 2016, ApJ, 822, 107

Géron, A. 2020, Hands-On Machine Learning With Scikit-Learn, Keras, And TensorFlow: Concepts, Tools, And Techniques To Build Intelligent Systems, 2nd edn. (Sebastopol, Canada: O'Reilly Media Inc.)

Glazebrook, K. et al. 2017, Nature, 544, 71

Glorot, X., Bordes, A., & Bengio, Y. 2011, in Proceedings of Machine Learning Research, Vol. 15, Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, ed. G. Gordon, D. Dunson, & M. Dudík (Fort Lauderdale, FL, USA: PMLR), 315

Goodfellow, I. J., Bengio, Y., & Courville, A. 2016, Deep Learning (Cambridge, MA, USA: MIT Press)

Guo, Q., White, S., Li, C., & Boylan-Kolchin, M. 2010, MNRAS, 404, 1111

Guth, A. H. 1981, Phys. Rev. D, 23, 347

Habouzit, M. et al. 2022, MNRAS, 511, 3751

Hadzhiyska, B., Bose, S., Eisenstein, D., & Hernquist, L. 2020, MNRAS, 501, 1603

Hadzhiyska, B., Liu, S., Somerville, R. S., Gabrielpillai, A., Bose, S., Eisenstein, D., & Hernquist, L. 2021, MNRAS, 508, 698

Haggar, R., Pearce, F. R., Gray, M. E., Knebe, A., & Yepes, G. 2021, MNRAS, 502, 1191

Hani, M. H., Gosain, H., Ellison, S. L., Patton, D. R., & Torrey, P. 2020, MNRAS, 493, 3716

Haslbauer, M., Dabringhausen, J., Kroupa, P., Javanmardi, B., & Banik, I. 2019, A&A, 626, A47

Hearin, A. P., Zentner, A. R., van den Bosch, F. C., Campbell, D., & Tollerud, E. 2016, MNRAS, 460, 2552

Hellwing, W. A., Cautun, M., van de Weygaert, R., & Jones, B. T. 2021, PhysRevD, 103, 063517

Henriques, B. M. B., Yates, R. M., Fu, J., Guo, Q., Kauffmann, G., Srisawat, C., Thomas, P. A., & White, S. D. M. 2020, MNRAS, 491, 5795

Hoffmann, K. et al. 2014, MNRAS, 442, 1197

Hong, S., Zou, Z., Luo, A.-L., Kong, X., Yang, W., & Chen, Y. 2022, MNRAS, 518, 5049

Huchra, J. P., & Geller, M. J. 1982, ApJ, 257, 423

Ishiyama, T., Fukushige, T., & Makino, J. 2009, PASJ, 61, 1319

Ishiyama, T., Nitadori, K., & Makino, J. 2012

Ishiyama, T. et al. 2021, MNRAS, 506, 4210

Iyer, K. G. et al. 2020, MNRAS, 498, 430

Jenkins, A., Frenk, C. S., White, S. D. M., Colberg, J. M., Cole, S., Evrard, A. E., Couchman, H. M. P., & Yoshida, N. 2001, MNRAS, 321, 372

Jiang, F., & van den Bosch, F. C. 2014, MNRAS, 440, 193

Jiang, L., Helly, J. C., Cole, S., & Frenk, C. S. 2014, MNRAS, 440, 2115

Johnson, B. et al. 2013a, Flexible Stellar Population Synthesis (FSPS) For Python

Johnson, B. D. et al. 2013b, ApJ, 772, 8

Johnston, H., Westbeek, D. S., Weide, S., Chisari, N. E., Dubois, Y., Devriendt, J., & Pichon, C. 2023, MNRAS, 520, 1541

Kannan, R., Smith, A., Garaldi, E., Shen, X., Vogelsberger, M., Pakmor, R., Springel, V., & Hernquist, L. 2022, MNRAS, 514, 3857

Kauffmann, O. B. et al. 2020, A&A, 640, A67

Kelley, L. Z., Blecha, L., & Hernquist, L. 2016, MNRAS, 464, 3131

Kennicutt, R. C., & Evans, N. J. 2012, ARA&A, 50, 531

Kewley, L. J., Nicholls, D. C., & Sutherland, R. S. 2019, ARA&A, 57, 511

Knebe, A., Kravtsov, A. V., Gottlober, S., & Klypin, A. A. 2000, MNRAS, 317, 630

Kobulnicky, H. A., & Phillips, A. 2003, ApJ, 599, 1031

Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., Klypin, A. A., Gottlober, S., Allgood, B., & Primack, J. R. 2004, ApJ, 609, 35

Kravtsov, A. V., & Borgani, S. 2012, ARA&A, 50, 353

Kravtsov, A. V., & Klypin, A. A. 1999, ApJ, 520, 437

Lacey, C. G. et al. 2016, MNRAS, 462, 3854

Lagache, G., Cousin, M., & Chatzikos, M. 2018, A&A, 609, A130

Lambas, D. G., Alonso, S., Mesa, V., & O'Mill, A. L. 2012, A&A, 539, A45

L'Huillier, B., Park, C., & Kim, J. 2015, MNRAS, 451, 527

Li, Y., Ni, Y., Croft, R. A. C., Matteo, T. D., Bird, S., & Feng, Y. 2021, PNAS, 118

Lia, C., Portinari, L., & Carraro, G. 2002, MNRAS, 330, 821

Liao, L.-W., & Cooper, A. P. 2022, MNRAS, 518, 3999

Lipton, Z. C., Berkowitz, J., & Elkan, C. 2015, arXiv e-Print 1506.00019

Lovell, C. C., Vijayan, A. P., Thomas, P. A., Wilkins, S. M., Barnes, D. J., Irodotou, D., & Roper, W. 2020, MNRAS, 500, 2127

Lovell, C. C., Wilkins, S. M., Thomas, P. A., Schaller, M., Baugh, C. M., Fabbian, G., & Bahé, Y. 2021, MNRAS, 509, 5046

Lu, L., Shin, Y., Su, Y., & Karniadakis, G. 2020, CiCP, 28, 1671

Lundberg, S., & Lee, S.-I. 2017, NIPS

Madden, S. C. et al. 2020, A&A, 643, A141

Malkan, M. A. et al. 2021, PASSAGE-Parallel Application of Slitless Spectroscopy to Analyze Galaxy Evolution, JWST Proposal. Cycle 1, ID 1571

Mansfield, P., & Avestruz, C. 2021, MNRAS, 500, 3309

Marinacci, F. et al. 2018, MNRAS, 480, 5113

Maseda, M. V. et al. 2014, ApJ, 791, 17

McAlpine, S., Harrison, C. M., Rosario, D. J., Alexander, D. M., Ellison, S. L., Johansson, P. H., & Patton, D. R. 2020, MNRAS, 494, 5713

McGibbon, R., & Khochfar, S. 2022, MNRAS, 513, 5423

Montenegro-Taborda, D., Rodriguez-Gomez, V., Pillepich, A., Avila-Reese, V., Sales, L. V., Rodríguez-Puebla, A., & Hernquist, L. 2023, MNRAS, 521, 800

Montero-Dorta, A. D., Chaves-Montero, J., Artale, M. C., & Favole, G. 2021, MNRAS, 508, 940

Montero-Dorta, A. D., Rodriguez, F., Artale, M. C., Smith, R., & Chaves-Montero, J. 2022, arXiv e-Print: 2212.12090

More, S. et al. 2023, arXiv e-Print: 2304.00703

More, S., van den Bosch, F. C., Cacciato, M., Mo, H. J., Yang, X., & Li, R. 2009, MNRAS, 392, 801

Mouhcine, M., Lewis, I., Jones, B., Lamareille, F., Maddox, S. J., & Contini, T. 2005, MNRAS, 362, 1143

Nadler, E. O. et al. 2023, ApJ, 945, 159

Nagao, T., Maiolino, R., Marconi, A., & Matsuhara, H. 2011, A&A, 526, A149

Naiman, J. P. et al. 2018, MNRAS, 477, 1206

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563

Navarro, J. F. et al. 2004, MNRAS, 349, 1039

Nelson, D. et al. 2015, Astronomy and Computing, 13, 12

——. 2019a, MNRAS, 490, 3234

——. 2017, MNRAS, 475, 624

——. 2019b, CompAC, 6, 2

Ni, Y., Li, Y., Lachance, P., Croft, R. A. C., Matteo, T. D., Bird, S., & Feng, Y. 2021, MNRAS, 507, 1021

Onions, J. et al. 2012, MNRAS, 423, 1200

Ownsworth, J. R., Conselice, C. J., Mortlock, A., Hartley, W. G., Almaini, O., Duncan, K., & Mundy, C. J. 2014, MNRAS, 445, 2198

Pan, Y. et al. 2023, MNRAS, 519, 4499

Papovich, C. et al. 2018, ApJ, 854, 30

Paranjape, A., Kovač, K., Hartley, W. G., & Pahwa, I. 2015, MNRAS, 454, 3030

Pasquali, A., Gallazzi, A., Fontanot, F., Van Den Bosch, F. C., De Lucia, G., Mo, H. J., & Yang, X. 2010, MNRAS, 407, 937

Pedregosa, F. et al. 2011, Journal of Machine Learning Research, 12, 2825

Peebles, P. J. E. 1982, ApJ, 263, L1

Peebles, P. J. E., & Ratra, B. 2003, Rev. Mod. Phys., 75, 559

Peng, Y.-J., & Maiolino, R. 2013, MNRAS, 438, 262

Perlmutter, S. 2000, International Journal of Modern Physics A - IJMPA, 15, 715

Pfeffer, J., Cavanagh, M. K., Bekki, K., Couch, W. J., Drinkwater, M. J., Forbes, D. A., & Koribalski, B. S. 2023, MNRAS, 518, 5260

Pillepich, A. et al. 2017a, MNRAS, 475, 648

——. 2019, MNRAS, 490, 3196

——. 2017b, MNRAS, 473, 4077

Planck Collaboration. 2016, A&A, 594, A13

Poole, G. B., Mutch, S. J., Croton, D. J., & Wyithe, S. 2017, MNRAS, 472, 3659

Poudel, A., Heinämäki, P., Tempel, E., Einasto, M., Lietzen, H., & Nurmi, P. 2017, A&A, 597, A86

Prada, F., Klypin, A. A., Cuesta, A. J., Betancort-Rijo, J. E., & Primack, J. 2012, MNRAS, 423, 3018

Press, W. H., & Davis, M. 1982, ApJ, 259, 449

Rau, S., Vegetti, S., & White, S. D. M. 2013, MNRAS, 430, 2232

Riggs, S. D., Loveday, J., Thomas, P. A., Pillepich, A., Nelson, D., & Holwerda, B. W. 2022, MNRAS, 514, 4676

Robaina, A. R., Bell, E. F., van der Wel, A., Somerville, R. S., Skelton, R. E., McIntosh, D. H., Meisenheimer, K., & Wolf, C. 2010, ApJ, 719, 844

Robertson, B. E. 2022, ARA&A, 60, 121

Roca-Fàbrega, S. et al. 2021, ApJ, 917, 64, 2106.09738

Rodriguez-Gomez, V. et al. 2015, MNRAS, 449, 49

Rosas-Guevara, Y. et al. 2019, MNRAS

Rosdahl, J. et al. 2018, MNRAS, 479, 994

Rubin, V. C., & Ford, W. K. J. 1970, ApJ, 159, 379

Rubin, V. C., Ford, W. K. J., & Thonnard, N. 1980, ApJ, 238, 471

Saghiha, H. 2017, PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn

Saha, K., & Naab, T. 2013, MNRAS, 434, 1287

Salcedo, A. N., Maller, A. H., Berlind, A. A., Sinha, M., McBride, C. K., Behroozi, P. S., Wechsler, R. H., & Weinberg, D. H. 2018, MNRAS, 475, 4411

Sánchez Almeida, J., Terlevich, R., Terlevich, E., Cid Fernandes, R., & Morales-Luis, A. B. 2012, ApJ, 756, 163

Schaye, J. et al. 2014, MNRAS, 446, 521

Schneider, P. 2015, Extragalactic Astronomy And Cosmology, 2nd edn. (Heidelberg, Germany: Springer-Verlag Publications)

Scholz-Díaz, L., Martín-Navarro, I., & Falcón-Barroso, J. 2022a, MNRAS, 511, 4900

——. 2022b, MNRAS, 518, 6325

Scoville, N. et al. 2013, ApJSS, 206, 3

SDSS Collaboration. 2002, AJ, 123, 485

Shi, J. et al. 2020, ApJ, 893, 139

Simha, V., Weinberg, D. H., Davé, R., Fardal, M., Katz, N., & Oppenheimer, B. D. 2012, MNRAS, 423, 3458

Simpson, C. M., Grand, R., Gómez, F., Marinacci, F., Pakmor, R., Springel, V., Campbell, D., & Frenk, C. 2018, MNRAS, 478, 548

Slone, O., Jiang, F., Lisanti, M., & Kaplinghat, M. 2021, arXiv e-Print:2108.03243

Smoot, G. F. 1999, in Conference on 3K cosmology (ASCE)

Snyder, G. F., Peña, T., Yung, L. Y. A., Rose, C., Kartaltepe, J., & Ferguson, H. 2022, MNRAS, 518, 6318

Somerville, R. S., & Davé, R. 2015, ARA&A, 53, 51

Sorini, D., Davé, R., Cui, W., & Appleby, S. 2022, MNRAS, 516, 883

Sousbie, T. 2011, MNRAS, 414, 350

Springel, V. 2010, MNRAS, 401, 791

Springel, V. et al. 2017, MNRAS, 475, 676

Springel, V., White, S. D. M., Tormen, G., & Kauffmann, G. 2001, MNRAS, 328, 726

Stark, D. P., Ellis, R. S., Chiu, K., Ouchi, M., & Bunker, A. 2010, MNRAS, 408, 1628

Suzuki, T. L. et al. 2016, MNRAS, 462, 181

Tantalo, R., & Chiosi, C. 2004, MNRAS, 353, 917

Tinsley, B. M. 1980, Fund. Cosmic Phys., 5, 287, 2203.02041

Tojeiro, R. et al. 2017, MNRAS, 470, 3720

Torrey, P. et al. 2019, MNRAS, 484, 5587

Trevisan, M., Mamon, G. A., Thuan, T. X., Ferrari, F., Pilyugin, L. S., & Ranjan, A. 2021, MNRAS, 502, 4815

Vale, A., & Ostriker, J. P. 2004, MNRAS, 353, 189

van Dokkum, P. G. et al. 2008, ApJL, 677, L5

Van Loon, M. L., Mitchell, P. D., & Schaye, J. 2021, MNRAS, 504, 4817

Veena, P. G., Cautun, M., van de Weygaert, R., Tempel, E., Jones, B. J. T., Rieder, S., & Frenk, C. S. 2018, MNRAS, 481, 414

Villaescusa-Navarro, F. et al. 2021, ApJ, 915, 71

Viroli, C., & McLachlan, G. J. 2017, arXiv e-Print 1711.06929

Vogelsberger, M., Marinacci, F., Torrey, P., & Puchwein, E. 2020, Nat. Rev. Phys., 2, 42

Wadekar, D., Villaescusa-Navarro, F., Ho, S., & Perreault-Levasseur, L. 2020, arXiv e-Print:2012.00111

Wang, W. et al. 2017, MNRAS, 469, 4063

Wechsler, R. H., & Tinker, J. L. 2018, ARA&A, 56, 435

Wechsler, R. H., Zentner, A. R., Bullock, J. S., Kravtsov, A. V., & Allgood, B. 2006, ApJ, 652, 71

Weinberg, S. 2008, Cosmology (Oxford, UK: Oxford University Press)

Weinberger, R., Springel, V., & Pakmor, R. 2020, ApJSS, 248, 32

Welker, C., Devriendt, J., Dubois, Y., Pichon, C., & Peirani, S. 2014, MNRAS Letters, 445, L46

Wetzel, A. et al. 2022, arXiv e-Print 2202.06969

White, M. 2001, A&A, 367, 27

White, S. D. M., & Frenk, C. S. 1991, ApJ, 379, 52

White, S. D. M., & Rees, M. J. 1978, MNRAS, 183, 341

Whitler, L., Stark, D. P., Endsley, R., Leja, J., Charlot, S., & Chevallard, J. 2023, MNRAS, 519, 5859

Wilkins, S. M., Feng, Y., Matteo, T. D., Croft, R., Lovell, C. C., & Waters, D. 2017, MNRAS, 469, 2517

Yang, X., Mo, H. J., & van den Bosch, F. C. 2003, MNRAS, 339, 1057

Yates, R. M., Péroux, C., & Nelson, D. 2021, MNRAS, 508, 3535

Yuan, S., Hadzhiyska, B., & Abel, T. 2023, MNRAS, 520, 6283

Yusofi, E., Khanpour, M., Khanpour, B., Ramzanpour, M. A., & Mohsenzadeh, M. 2022, MNRAS Letters, 511, L82

Zavala, J., & Frenk, C. S. 2019, Galaxies, 7, 81

Zehavi, I., Kerby, S. E., Contreras, S., Jiménez, E., Padilla, N., & Baugh, C. M. 2019, ApJ, 887, 17

Zhang, Y., de Souza, R. S., & Chen, Y.-C. 2022, MNRAS, 517, 1197

Zhao, D., Du, M., Ho, L. C., Debattista, V. P., & Shi, J. 2020, ApJ, 904, 170

Zhao, D. H., Jing, Y. P., Mo, H. J., & Börner, G. 2009, ApJ, 707, 354

Zheng, Z. et al. 2005, ApJ, 633, 791

Zwicky, F. 1933, Helvetica Physica Acta, 6, 110