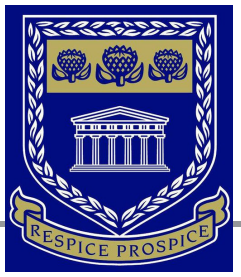
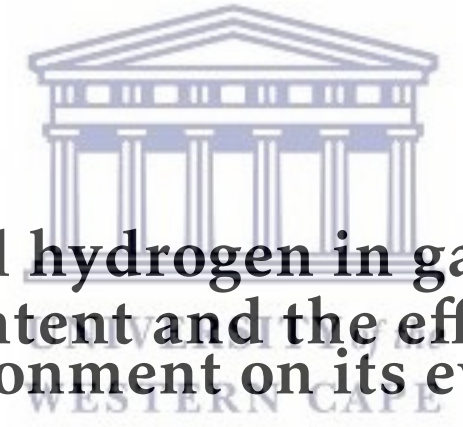
University of The Western Cape



A thesis submitted in partial fulfilment of the requirements for the degree

Doctor Philosophiae in Astronomy

Department of Physics and Astronomy



Neutral hydrogen in galaxies, its content and the effect of environment on its evolution

By:

Mika Harisetry Rafieferantsoa

Supervisor:

Prof. Romeel Davé

November 2018

Keywords

Declaration

Under the supervision of Prof. Romeel Davé, my Ph.D. journey ended with the following peer-reviewed publications:

- **Rafieferantsoa, M.**, Andrianomena, S., Davé, R., *Predicting the neutral hydrogen content of galaxies from optical data using machine learning*, 2018, MNRAS 479, 4509
- **Rafieferantsoa, M.**, Davé, R., Mufasa*: the strength and evolution of galaxy conformity in various tracers*, 2018, MNRAS 475, 955
- Davé, R., **Rafieferantsoa, M.**, Mufasa*: the assembly of the red sequence*, 2017, MNRAS 471, 1671
- Davé, R., **Rafieferantsoa, M.**, Thompson, Robert J., Hopkins, Philip F., Mufasa*: Galaxy star formation, gas, and metal properties across cosmic time*, 2017, MNRAS 467, 115

and the following recently submitted paper:

- **Rafieferantsoa, M.**, Davé, R., Mufasa*: Timescales for* HI *consumption and SFR depletion of satellite galaxies in groups*, 2018 arXiv

The publications I first-authored constitute the chapters of this thesis, with modifications to fit the required format of the Univeristy.

I declare that *Neutral hydrogen in galaxies, its content and the effect of environment on its evolution* has not been submitted to any university for a degree, and all of the work in this thesis, save for that which is properly acknowledged, is my own.

Signed: _____

Date: _____

Abstract

Using two hydrodynamic galaxy formation simulations from the Mufasa project that I helped develop, we aim to better understand the relationship between galaxy evolution and its cold gas content commonly known as the neutral hydrogen or HI. We first look at the environmental properties of the simulated galaxies and compare to those that are available observationally. As a proxy, we specifically quantify the so-called galactic *conformity*, which is the concordance between the properties of galaxies neighbouring the primaries, in chapter 2. We show that the HI, the specific star formation rate (sSFR) and the colour of galaxies show galactic conformity in qualitative agreement with previous observed data, *i.e.* the HI-rich primary galaxies are surrounded by HI-richer galaxies than the HI-poor primary galaxies, and similarly for the sSFR and the colour. We find that environment, quantified by the number of neigbouring galaxies within a fixed aperture, stellar age and molecular hydrogen (H_2) also show conformity. Galactic conformity also depends on the dark matter halo mass of the primary galaxy. The galactic conformity signal from the primaries of smaller haloes is weak but extends out to several virial radii of those structures, whereas the signal is very strong for high mass haloes but lowers quickly with distances from the primaries. We also find the galactic conformity only emerges in the later half of cosmic evolution. We next quantify the gas content and star formation depletion timescales in chapter 3. We use two carefully chosen groups of simulated galaxies and find that timescales are affected by both the mass of the virialised structure of the first infall and the galaxy stellar mass at infall: the higher the halo mass or the stellar mass the shorter the timescale. The gas or HI depletion timescale is concordant to that of the star formation quenching, indicative of direct decrease of SFR due to depletion of the extended cold gas reservoir. The neutral atomic or molecular hydrogen consumption timescale depends on the Hubble time. Galaxies tend to form stars more efficiently at lower redshift. While the halo mass of infall affects the consumption timescale of the HI, it does not correlate with the H_2. We lastly develop machine learning tools to use galaxy photometric data to predict a galaxy's HI mass in chapter 4, to allow predictions for HI from much larger optical photometric surveys. The training and testing of the algorithms are done first with the simulated data from Mufasa. We show that our model performs better than previously done with ad hoc data fitting approaches. Random Forest (RF) followed by the Deep Neural Networks (DNN) perform best among the explored machine learning techniques. Extending the trained models to observed data, namely the Arecibo Legacy Fast ALFA (ALFALFA) and REsolved Spectroscopy Of a Local VolumE (RESOLVE) survey data, we show the overall performance is slightly reduced relative to the simulated testing set owing to the small inconsistency between definition of galaxy properties between simulation and observational data, and DNN perfoms the best in this case. The application of our methods is useful for galaxy-by-galaxy predictions and anticipated to correct for incompletness in the upcoming HI deep surveys done with MeerKAT and eventually the Square Kilometre Array (SKA).

Fisaorana

Fihainana ny PhD fa tsy fandalovana. Ka raha ho tanisaina amin'ny antsipirihany daholo ny lalana nahatongavana amin'izao fahatontosana ity asa ity izao dia na boky iray ary tsy ho ampy. Saingy misy ireo tranga niavaka ary sarotra ny tsy hitsonga azy amin'ny maro.

Isaorako voalohany ny raibeko J. E. Rafieferantsoa, izay loharano nipohiran'ny saina tia karokaroka ato amiko. Na dia efa tsy miaraka amintsika intsony ary izy amin'ny fotoana anoratako ity fisaorana ity, dia mitohetra ato amiko hatrany ny fomba fihaina izay nampitainy. Isaorako manaraka izany ny ray aman-dreniko M. L. Rafieferantsoa sy G. Razafindrasoa izay loharano nipoirako ary fototry ny fitaizana izay tsy ho aiko honerana amin'izay rehetra hananako. Tsy ambakan'izany ny rahalahiko H. T. Rafieferantsoa izay niara nilalao sy nialehibe tamiko. Tsy adino i F. E. Randrianarivony. Ianareo no nanefy sy nanitsy ny fiainako ka nahatonga ahy toy izao. Manohana ahy hatrany ianareo.

As far as I can remember, I beleive R. Davé is the first person who really introduced me to the scientific method and I cannot be more grateful to have him as my supervisor. He shaped my way to approach research and I have really benefited from his highly regarded knowledge and skills in the galaxy formation and evolution theories. I started my M.Sc. project with him and our collaboration seems not to end any time soon. Thank you Romeel.

Ireo namako rehetra any Madagasikara, isaorana tamin'ny fotoana nahafinaritra.

For all my previous classmates, NASSP Honours 2012, in Cape Town that occasionally keep in touch, I am glad you were part of this journey.

Hisaorana ireo Malagasy namana izay mipetra eto Afrika Atsimo, izay nifandray matetika tamiko nandritra ny fikarakarana ity vokatra ity – Hagatiana mivady sy i Kacey, Zara mivady, Holifidy, Li sy Nandrianina ary indrindra i Fara kely – nahafinaritra eee!! Marihina fa i Hagatiana dia nandray anjara tamin'ny fanitsiana ny tsipelina sy ny haiteny anglisy. Marihina ihany koa fa i Fara kely dia nandray anjara tamin'ny fandrahoana kaly sy nifidy ny fanafodiko raha narary aho ary n.... (oups) ihany koa.

Hagatiana and Paul Udit were playing Tekken with me, I always chose Bob when I wanted to win, speeeed and w..... Paul Udit is a very good friend. We planned things, then .... women came!! Not forgetting Baby Vladmir (Emmanuel) and Mr Muscle (Sheean) whom I often play pool with. All the office mates I have had. I would have been crazier (lol) if you were not around. You became friends.

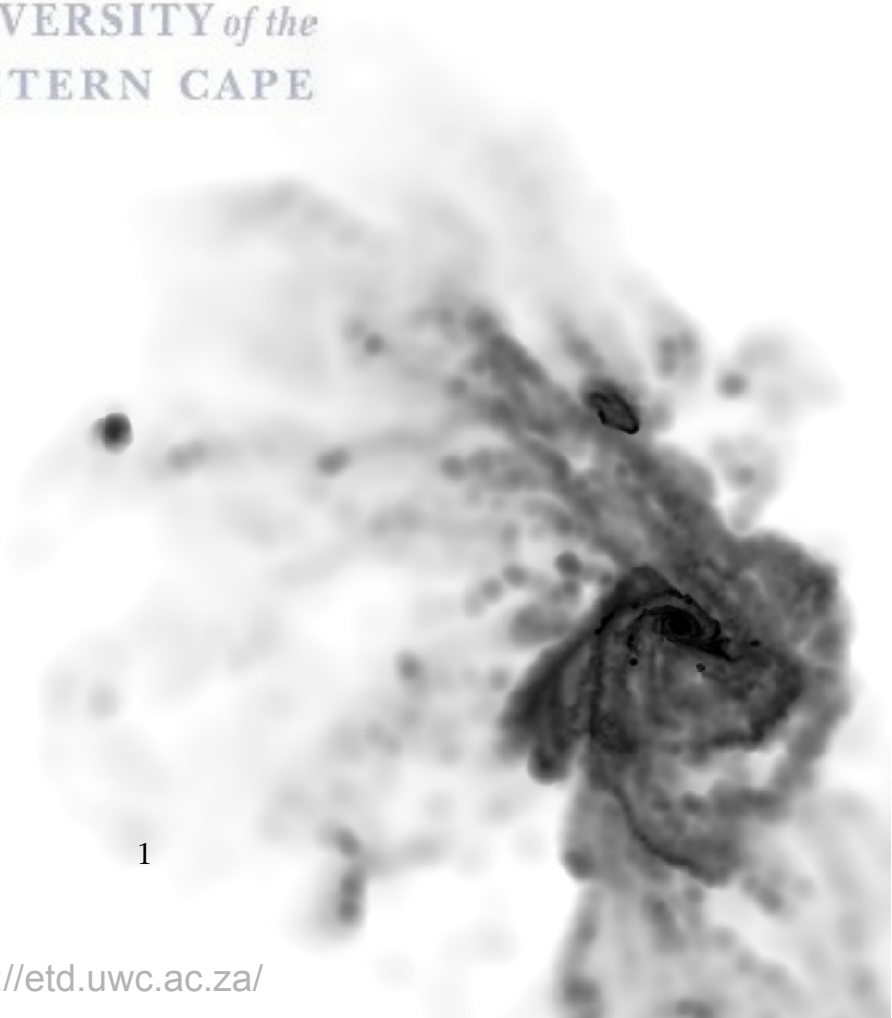
For those who whished to remain anonymous, I still dedicate this last line to remind that in some way you have contributed in the achievement of this thesis.

Contents

Ho an'i Rafieferantsoa

CHAPTER 1

Introduction



Neutral hydrogen is the first element in the periodic table of Mendeleev, often noted as HI (H-one). It contains one proton and one electron, thus its electrically neutral state. Most of the hydrogen we encounter in our everyday life is in molecular forms due to its covalent property, but far away from similar environments to our planet, it can remain in atomic or ionised forms. Neutral hydrogen was first observed in our own galaxy, the Milky Way, in 1951 by Ewen and Purcell, and since then the interest of scientists in studying HI has grown exponentially. To understand the value of HI in extra-galactic science, we first need to go back to how we currently think the Universe began and evolved.

The prevailing theory about the origin of the Universe is that it started as a near-infinitely dense singularity. The expansion during the first fraction of a second was extremely fast, reminiscent to explosion, from which the theory was named the *Big Bang* by Fred Hoyle in 1949. During the first nanoseconds of expansion, the Universe cooled sufficiently enough to form quarks and other elementary particles first and then protons and neutrons.

The highly energetic photons were constantly interacting with each other resulting in pair production and annihilation of electrons and positrons, until the Universe had expanded and cooled down enough to favor the annihilation rather than the pair production process. Interestingly, 1 electron for every $\sim 10^9$ others did not have an antiparticle. Similarly, 1 antiproton for every $\sim 10^9$ others was missing. Both were due to the CP-violation. In particle physics, CP-violation refers to the violation of the charge conjugation (C) and parity (P) symmetry. C-symmetry means that the laws of physics should remain unchanged if a charge conjugation happens. P-symmetry states that if a particle is interchanged to its antiparticle, the spatial coordinates should be reversed (mirror-symmetry).

The mass of neutrons is slightly higher than the mass of the protons, but the energy density was high enough that the proton-neutron conversion happened continually. The Universe further expanded and cooled to break the proton-neutron balance. The lower energy required to go from neutrons to protons than vice-versa resulted in fewer neutrons. The final ratio was estimated to be $\sim 87\%$ protons and $\sim 13\%$ neutrons. The decreasing energy density of the Universe ultimately allowed the protons and the neutrons to interact and the *Big Bang nucleosynthesis* (BBN; Wagoner et al., 1967) began. Protons and neutrons could merge to form deuteriums which can further interact to form helium-4

nuclei. The deuterium can stick to a neutron first to form a tritium (hydrogen-3) then to a proton to become a helium-4 (He). Another pathway was for the deuterium to stick with a proton first to form helium-3, then to a neutrino to make a He nuclei. At the completion of the first phase of the BBN, about few minutes after the Big Bang, the Universe was roughly $\sim 75\%$ hydrogen, $\sim 25\%$ helium and $\sim 0.01\%$ deuterium and tritium nuclei by mass.

From that time, the Universe was opaque because photons could not travel any further than incredibly small distances before hitting a charged particle, causing the photon to scatter.

Only about 380 000 years after the Big Bang did the Universe cool down enough that the electrons became less energetic not to escape the company of nuclei, and neutral atoms formed. This is the so called *Recombination* epoch (Peebles, 1968), when the neutral hydrogen first formed, and the Universe became transparent. The photons that exited the recombination era and managed to pass through the neutral medium of atomic hydrogens to reach us today constitute the earliest image of the Universe, known as the Cosmic Microwave Background (CMB; Penzias & Wilson, 1965). Since Recombination, the Universe remained transparent, not emitting any visible light, in an epoch known as the Cosmic Dark Ages. Only after a few hundred million years after the Big Bang were the first light-emitting objects formed.

The first luminous objects came to existence through the structure formation with the cold dark matter (CDM; Peebles, 1982; Bond et al., 1982; Blumenthal et al., 1982) paradigm. CDM collapsed under gravity to form virialized structures, often called dark matter haloes. In an oversimplified scenario, the neutral gas, particularly the HI, fell in the deeper potential of the haloes to become molecular hydrogen (H_2) followed by the formation of stars and galaxies. The dominant coolant was the fragments of molecular hydrogen because the Universe was mostly neutral (Abel et al., 2002). The first stars and galaxies produced high energy radiation that impacted the surrounding HI by exciting their electrons and eventually dissociating them from the protons, making them ionised. This phase is the Epoch of Reionisation (EoR), ending about 1 billion years after the Big Bang, and since then the Universe has only become more ionised. This was when the most interesting and complicated connection between HI and galaxy evolution started because the dense gas around galaxies was able to shield itself from radiation and remain neutral.

During the post-reionisation era, particularly for the purpose of this thesis, most of the gas component of the Universe is assumed to be optically thin, in ionisation equilibrium and embedded in an uniform background metagalactic radiation field (e.g. Haardt & Madau, 2012). The dominant gravitational field of the dark matter drags the gas which the virialized haloes would ultimately accrete and shock heat. The accretion shock is expected because the temperature of the infalling gas is largely less than the temperature of the haloes (Binney, 1977) although in smaller haloes much of the accreted gas does not shock heat and accretes in the cold mode (Kereš et al., 2005). The next section will review the different steps of star formation from the shock heated (ionised) gas to the HI to the stars.

1.1 Star formation and HI

The star formation of galaxies is fueled by molecular hydrogen. But to go from the ionised intergalactic medium to the formation of stars, the gas has to pass through its atomic neutral phase that is easier to detect observationally. The neutral state of the gas is however altered by the physical mechanisms such as cooling and heating.

1.1.1 Cooling & heating

The accreted gas is heated about the virial temperature of the halo T_v such that

$$T_v \propto \frac{M_v}{r_v} \tag{1.1}$$

with M_v and r_v the virial mass and the virial radius. Without any dissipation, the gas will reach a hydrostatic equilibrium and would stay as such indefinitely with the following radial pressure gradient

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{GM(r)}{r^2}\rho(r) \tag{1.2}$$

where P is the pressure of the gas, G the gravitational constant, $M(r)$ the mass of matter inside radius r and $\rho(r)$ the density of the gas inside r. Such an equilibrium state is only observed in X-rays. But some of it does cool and fall further to

the center of the halo. There are many astrophysical radiative cooling processes such as the primordial plasma cooling, metal-line cooling, Compton cooling and molecular cooling, but we will only consider the first two for their dominant influences. The cooling mechanisms that affect the primordial elements (H/He) are the collisional excitation and ionisation, recombination emission and free-free emission or Bremmsstrahlung. Metal-line cooling is caused by the collisional excitation of metals which revert back to their lower energy state and emit photons. Once cooled the gas can gradually fall in the deeper potential of the halo, increase its density and become neutral again (HI). Further cooling will promote the formation of H_2 and ultimately the formation of stars. The transitionary state of the gas as neutral is very crucial because HI is a very good tracer of that event, and can be observed via its radio emission.

Beside the cooling processes, the halo gas also experiences heating processes. The formation of stars produces radiation covering a wide range of energies that can heat and ionise the surrounding (falling) HI inside the halo. The cooling rate of the gas is then altered and the star formation impeded. Extra heating processes are required because without it the gas can further cool to form HI then stars and galaxies would have been more massive (e.g. White & Rees, 1978; Benson et al., 2003) than observed. This overcooling problem of the gas can be counterbalanced by the energy deposited from supernovae. The energy produced would eject and/or heat the HI (e.g. Murray et al., 2005). Another factor contributing in heating the gas is the Active Galactic Nuclei (AGN). The formation of black holes, although due to accretion of matter at the very dense region of the galaxy, outputs an enormous amount of energy that directly affects the circum-galactic medium and even reaches the intracluster medium (Fabian, 2012), thereby changing the cooling properties of the neutral hydrogen and altering the star formation and the growth of galaxies. The transitory neutral phase of the gas (HI) is therefore strongly impacted by the supernova and AGN feedback of the galaxies.

To this end, it is clear that a galaxy evolves depending on different physical mechanisms that affect the gas content of its surroundings. As a result, the global cosmic star formation rate increased to reach a maximum around 10 billion years ago (Madau & Dickinson, 2014), before decreasing until today.

Galaxy groups and clusters are the environments where the HI content of galaxies appears to transition from gas-rich to gas-poor, making them the ideal cosmic structures to focus on. But before presenting the numerical models for simulating the galaxy evolution, the next section reviews the current observational data to constrain the model and to situate the topic of this thesis.

1.2 Galaxy properties through a telescope

The Big Bang theory results in an expanding Universe, *i.e.* objects further from us recede faster than those closer. We can only observe objects or structures that either emit, absorb, or reflect electromagnetic radiation. Those photons that travel from the objects to us come in a wide range of wavelengths and brightnesses, most of which is not visible to the human eye but can be detected with telescopes. Since it takes time for the photons to reach us, it is clear that distance implies age and *vice-versa*, *i.e.* we can only see spectra from longer ago emitted by a distant object. Because of cosmic expansion, while traveling, emitted photons have their wavelength stretched to higher values until reaching us: this is called cosmological *redshift* (z). Mathematically, $z = (\lambda_{\rm obs} - \lambda_{\rm rf})/\lambda_{\rm rf}$, where $\lambda_{\rm obs}$ is the observed wavelength and $\lambda_{\rm rf}$ is the rest frame wavelength. Thus further away objects recede faster and their z are higher so we see them when they are younger, while closer objects have lower z. For instance, the recombination epoch happened at $z \simeq 1100$, the furthest galaxy ever observed at $z \simeq 11$ and the present epoch is $z = 0$. It is important to note that peculiar motions of a source can also alter its redshift.

1.2.1 Galaxy classification

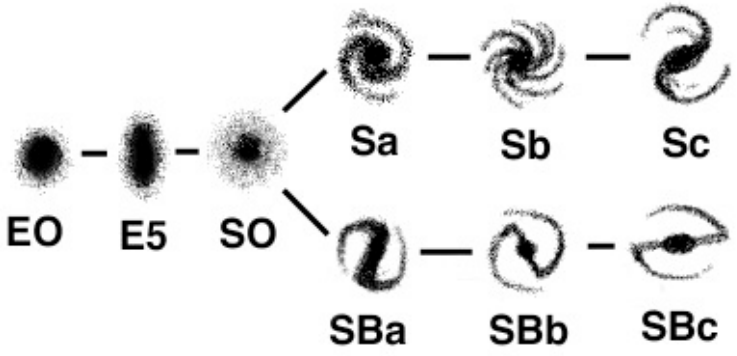


FIGURE 1.1: Tuning-fork diagram of Hubble's galaxy classification.

The traditional classification of galaxies was inspired by the variety of their shapes. Edwin Hubble introduced the tuning-fork diagram, in 1926, to divide the galaxy population into three categories based on the spiral arms (Figure 1.1). The galaxies are elliptical when they do not display any spiral features. They are sub-classified from E0 to E7 depending on their ellipticity with 0 being circular and 7 ellongated. Spiral galaxies are forked into two prongs: SB for barred spiral galaxies and S for purely spiral galaxies. Early and late types are also different nomenclatures of elliptical and spiral galaxies respectively. S0 labels galaxies between elliptical and spiral types. De Vaucouleurs et al. (1976) expanded on the classification of the spirals by additionally looking at galaxy rings. Morphological classifications were useful but the advent of multi-wavelength surveys, *i.e.* observation of galaxies through different filters, like the Sloan Digital Sky Survey (SDSS; York et al., 2000) or the Great Observatories Origins Deep Survey (GOODS; Giavalisco et al., 2004) opened a new path in exploring galaxy formation. In fact, the dichotomy in galaxy morphologies is also present in their color *vs* magnitude (CM) diagram. Odekon et al. (2016), for instance, classified galaxies in two ways. First by adopting the spiral critera in the Galaxy Zoo (Lintott et al., 2008) and second by fitting a double Gaussian distributions in color magnitude space.

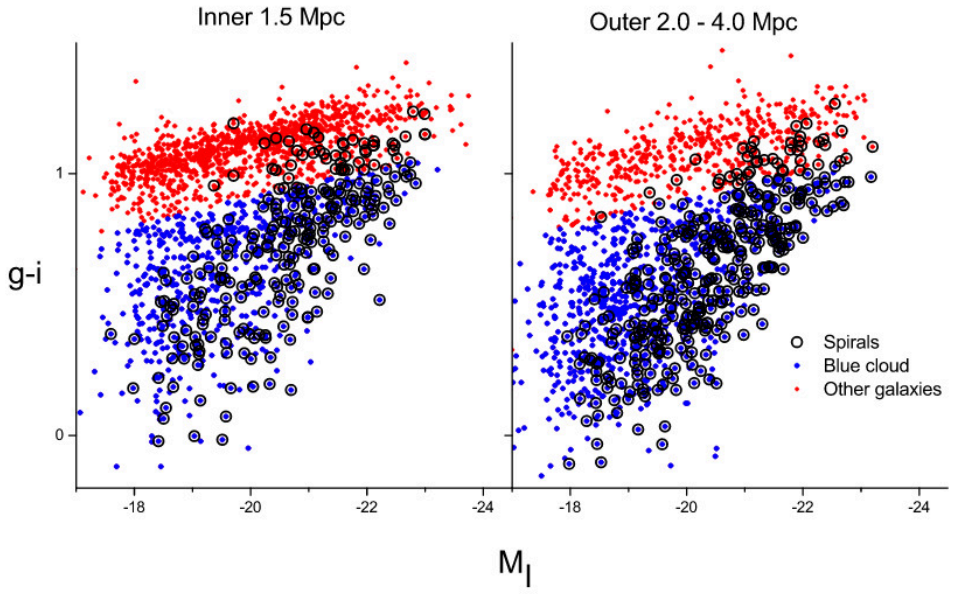


FIGURE 1.2: (Odekon et al., 2016, Fig. 2). Color-magnitude diagram of galaxies in the inner 1.5 Mpc (*Left*) of their respective groups and the control sample in the outer regions (*Right*). Spiral galaxies are mostly blue and they are more numerous further out (*right* panel *vs left* panel).

It turns out that galaxy M_i *vs* $g-i$ color separates the late types from the rest (Figure 1.2). For consistency check, Odekon et al. (2016) compared the galaxy population within 1.5 Mpc of their respective groups (Fig. 1.2, left panel) with the control sample they chose to be at 2 – 4 Mpc away from the center of mass of the galaxy groups (Fig. 1.2, right panel). It is clear, that regardless of the environmental condition, the late type galaxies (black circles) are always bluer compared to the rest of the sample. Here g and i refer to the apparent magnitudes measured in g and i filters of the Sloan Telescope (internal extinction corrected (Shao et al., 2007) and Galactic extinction corrected (Schlafly & Finkbeiner, 2011)) and M_i the absolute magnitude of the galaxy in i band. Lower value of $g-i$ means the galaxy is bluer and higher value means redder. With connection to Weinmann et al. (2006b) findings on the distinction between late-type and early-type galaxies in terms of their star formation rate, which was shown to arise from the difference in their gas contents (Morganti et al., 2006), the morphological dichotomy of galaxies (Odekon et al., 2016) is the inprint of their evolutionary phases. Furthermore, the existence of fewer passive late-type and active early-type galaxies, shown in (Weinmann et al.,

2006b), indicate a transitionary atomic neutral phase of the hydrogen affecting the star formation history of the population. To study the correlation, one needs observational information about the gas content of galaxies.

1.2.2 HI in galaxies

The HI can emit photons with a frequency of 1.420 GHz or equivalent to 21.106 cm in wavelength, hence it is often referred to as the 21 cm emission line. The neutral hydrogen atom is in its lowest energy state (very stable) when the spins of the proton and the electron are anti-parallel. The spin does not refer to a physical rotation but rather the quantum state of the particles. At the very ground state, interactions and collisions of hydrogen atoms might provide tiny amount of energy causing the electrons to spin-flip, resulting in a parallel configuration. The transition is extremely forbidden but once it happens the half life of the excited state can last millions of years before the atom returns to the ground state and emits 1 photon with a frequency of 1.420 GHz. Due to the vast amount of HI in the cosmos, we still can receive the echo of the 21 cm emission line through the telescopes in the form of flux that can be used to measure the amount of HI.

A major survey on HI with the blind – no specific target to track – extra-galactic survey Arecibo Legacy Fast ALFA (ALFALFA Haynes et al., 2018) has completed the intended two high galactic latitude regions in the northern and southern hemispheres. The survey covered 7000 deg^2 on the sky accessible by the Arecibo telescope and provide a catalog of $\sim$ 31500 HI line sources (galaxies). The REsolved Spectroscopy Of a Local VolumE (RESOLVE Kannappan et al., 2011) survey targeted $\sim$ 1500 galaxies covering ranges of star formation rate, stellar and gas masses, and located in a limited 53000 Mpc^3 volume of the sky. The presence of neutral gas in different types of galaxies at different amounts and their effects on the star formation can tell us the history of their evolution. This thesis aims to explore those effects and we will use the ALFALFA and the SDSS data to provide constraints in our simulation models.

Future surveys such as Looking At the Distant Universe with the MeerKAT Array (LADUMA; Holwerda et al., 2012) will probe 1 deg^2 of the Extended Chandra Deep Field South (ECDF-S) part of the sky for 21cm line observations. The

survey is allocated over 3400 hours of MeerKAT observing time that will detect HI signals from galaxies back to when the Universe was barely 4.5 billion years old or $z \sim 1.4$ (Baker et al., 2018). The observing techniques will be a mix of direct detections of HI emission lines for low redshift galaxies and line stacking techniques for high redshift galaxies due to decreasing sensitivity at larger distances. This will provide unprecedented amount of data to allow a detailed study of the cosmic evolution of the HI density, the evolution of HI content of galaxies as well as the effect of environment on the HI content. This thesis in part aims to provide predictions for such survey in a cosmological multi-wavelength context. Another instrument is the Hydrogen Intensity and Real-time Analysis eXperiment (HIRAX) radio telescope that will observe the 21cm signals at $z \simeq 1 - 2.5$ over very large scales to understand the properties of dark energy via the Baryon Acoustic Oscillations (BAO) (Newburgh et al., 2016). This will provide extended maps of the neutral hydrogen emissions at much higher redshift, although the sources will not be resolved. The machine learning portion of this thesis aims to connect the optically-observed galaxy population with the expected HI in order to help quantify biases in such large-scale surveys.

1.3 Hydrodynamic simulations with galaxy formation models

Data from observations are the pillar of well-contrained galaxy formation models. Once a model is well constrained by observations, we can run the simulation, at any epoch, and check the evolution of the different properties of the galaxy population.

Hydrodynamic simulation are based upon the principle of N-body simulations. A system can be thought of as an ensemble of particles. Particle in this case refers to a point mass or a volume element that can be treated individually and acted upon by relevant forces. In the context of this work, 3 types of particles are involved: **dark matter, gas** and **stars**. Dark matter and star particles are collisionless, *i.e.* they only interact with each other and with other types of particles via gravitational forces. In addition to the latter, hydrodynamic forces are present in the motion of the gas particles. *Gas volume element*, which also refers

to the gas particle, will be often mentioned in this work. The galaxy formation model includes subgrid prescriptions affecting the hydrodynamic motions of the gas particles and their processes of conversion into star particles.

I use Gizmo code (Hopkins, 2015), a descendent of Gadget-2 (Springel, 2005). The code is written in C (Kernighan, 1988) and designed to be massively parallel with the Message Passing Interface (MPI) library. It employs a TreePM algorithm (similar to Bagla & Ray, 2003) and a mesh-based Fourier technique to compute the short and long range gravity forces, respectively, that are the essential drivers of structure formation. It adopts adaptive time-steps for the motion of particles. Particles require smaller time-steps in higher density than in lower density regions. For this reason, the motion of particles in low density regions needs only updating at longer time interval. This method saves immense computational resources. All the particles are equally treated when solving for gravitational forces but the gas volume elements are additionally subjected to hydrodynamic forces.

1.3.1 Hydrodynamic forces

Given the nature of this work, to look at the properties and evolution of the gas contents of galaxies, we require a state-of-the-art hydro-solver in order to accurately model the gas evolution.

One approach to solve for hydrodynamic forces is the Lagrangian method, which follows individual gas elements and solve for the mass and momentum conservations as well as the energy balance of the Euler equations. In one such method, the volumetric distribution of a gas mass is smoothed by a user chosen kernel function, hence the name: Smoothed Particle Hydrodynamics (SPH). The method is very robust such that the equation of motion and the continuity equation can be accurately solved. Because the SPH equations are inviscid by definition, we need to add artificial viscosity to be stable. As a result, the fluid mixing stabilities (e.g. Agertz et al., 2007) and the poorly represented sub-sonic turbulence (e.g. Sijacki et al., 2012) create numerical viscosity in a inviscid situation. Solving for the SPH equations was also shown to introduce errors even for a zero-order field (Morris, 1996) and corrections of such problem contribute

into breaking the conservation of momentum and energy (Price, 2012). Therefore, the SPH technique suffers innacuracies in modeling the gas content of galaxies.

Another approach is the grid based method. The whole volume is discretised into cells and the Euler equations are solved between cells. The technique is based on the Riemann solver (Toro, 2009) that computes all conserved quantities at cell interfaces. As a result, the method handles shocks accurately. Compared to SPH, the mass is not conserved between the cells. Although numerically stable, this method suffers from higher advection errors when gas moves from one cell to another and angular momentum is not conserved. These issues can negatively impact the fluid mixing, angular momentum-supported disks and turbulence that are important processes for galaxy formation.

To alleviate those persisting issues, attempts have been made to combine the power of the two methods which have proven to be successful. One of those is the Meshless Finite Mass (MFM) scheme which I opt to use in this work. I refrain from delivering a detailed set of equations describing the MFM approach (see Hopkins, 2015, for details), I will only elaborate on its advantages. Hopkins (2015) compared the SPH methods, the grid based methods and his newly developed MFM method. Although it is a Lagrangian method, similar to the SPH method, MFM handles the fluid instabilities and mixing much more accurately without artificial diffusion terms and the need for artificial viscosity by using a Riemann solver as in an Eulerian code. Smaller Mach number problems can be accurately treated and shocks and discontinuities of fluids are well solved. A cold inviscid ideal gas element orbiting in a Keplerian potential with no self-gravity which should orbit forever in such setup achieves > 100× more orbits before showing instabilities, whereas SPH shows instabilities after just a few orbits. To achieve a moderate agreement with MFM in this context, SPH has to use > 10× neighbours which additionally favors MFM in terms of computing time.

A moving mesh without the mesh, MFM conserves angular momentum and the advection errors are substantially suppressed due to the Godunov-type (Lanson & Vila, 2008a,b) schemes. A highly dense square of fluid in a hydrostatic equilibrium moving at constant velocity should conserve its shape forever based on the Galilean invariance. MFM remains virtually the same after 10 unit time whereas with the grid codes the edges of the square starts to diffuse only at

$t = 0.2$ and the shape is completely disrupted at $t = 10$. On top of these, the meshless structures with MFM, while solving the hydrodynamic equations between particles in mesh-like fashion, maintain symmetry and do not suffer from the mesh-bending instabilities mentioned in Springel (2010). The mesh-bending requires extra adjustments to tune the irregularities which increases the computing time. The next section introduces the galaxy formation models implemented on top of the MFM solver.

1.3.2 Subgrid physics: Mufasa

To capture the cosmic evolution of galaxies, a cubical box with an edge length of $50h^{-1}$Mpc requires over 100 million gas particles with individual mass of order 10^7 $M_\odot$ ($M_\odot$: solar mass). The spatial resolution of such simulation is $\sim 0.5h^{-1}$kpc. Therefore, modeling of the star formation, galactic winds from supernovae as well as quenching prescription from AGN feedback requires the use of subgrid schemes.

The star formation prescription is typically a function of gas density. A widely used form of such a prescription is a power-law scaling between the star formation rate and the gas particle density, known as the Schmidt (1959)-law. In Mufasa, the star formation rate accounts for the molecular fraction $f_{\rm H2}$ of gas as in

$$\frac{\mathrm{d}M_*}{\mathrm{d}t} = \epsilon \frac{\rho f_{\rm H2}}{t_{\rm dyn}} \tag{1.3}$$

with ρ the gas density, $\epsilon = 0.02$ the efficiency of star formation (Kennicutt, 1998) and $t_{\rm dyn} = 1/\sqrt{G\rho}$ the local dynamical time. The $f_{\rm H2}$ is calculated from a well motivated solver for H_2 formation (see Krumholz & Gnedin, 2011) as in

$$f_{\rm H2} = 1 - 0.75\frac{s}{1+0.25s} \qquad \text{with} \qquad s = \frac{\ln\left(1+0.6\chi+0.01\chi^2\right)}{0.0396 \times Z \times \Sigma} \tag{1.4}$$

where χ is a function of metallicity Z and Σ the column density of the gas. Based on $\mathrm{d}M_*/\mathrm{d}t$, the gas volume elements are chosen stochastically to become stars. One gas volume element is converted to one star particle. In the real Universe, the star radiates energy and creates sudden changes in the hydrodynamic motions of its surrounding. Depending on its age and size, the star might become a supernovae which can cause a strong energy and momentum input

that drives materials out from the vicinity of the event. Unfortunately, modeling these phenomena, known as *stellar feedback*, for individual gas volume elements in a cosmological volume is not computationally feasible. Instead, Mufasa adopts the scaling relations obtained from very high resolution simulations of individual galaxies that self-consistently include the observed effects of stellar winds, supernova feedback and radiation from massive stars. Using the set of zoom simulations from the FIRE project (Hopkins et al., 2014), Muratov et al. (2015) found a power-law scaling relation between the stellar mass M_* of the galaxies and the gas mass outflow rate η such that

$$\eta = 3.55\left(\frac{M_*}{10^{10}\,\mathrm{M}_\odot}\right)^{-0.351} . \tag{1.5}$$

Once ejected, the gas volume elements are given initial velocities $v_w \propto v_c^{1.12}$ perpendicular to the galaxy plane, again from Muratov et al. (2015). To account for the two-phase wind that are observed, for instance in M82, Mufasa randomly heat the fluid elements depending on the remaining energy from the SN accounting for the required kinetic energy for the ejection.

While stellar feedback is promising in regulating the star formation of galaxies at lower masses ($M_* \leq 10^{10}$ M$_\odot$), it does not stop the star formation of the very massive objects as observed in massive ellipticals. Springel et al. (2005) introduced the AGN feedback model in hydrodynamic simulations, and showed that depositing only 5% of the bolometric luminosity of the black hole (BH) can remove the cold gas content of the galaxy and stop the star formation. Anglés-Alcázar et al. (2015) found similar results while they used a torque-limited BH accretion instead of Bondi accretion. Although these are promising in the future, Mufasa uses a rather empirical approach to maintain the excessive growth of galaxies at the very massive end. Gabor & Davé (2015) showed that keeping the gas in massive haloes ($M_{\rm halo} \gtrsim 10^{12}$ M$_\odot$) to the virial temperature T_{vir} can prevent any further star formation for the massive galaxies. In this scenario, the gas in the interstellar medium that is self-shielded is not heated. $T_{\rm vir} = 9.52 \times 10^7 M_h^{1/3}$ is obtained from Voit (2005) and M_h the mass of the host halo from the simulation.

Mufasa reproduces the distribution of galaxies in terms of stellar mass (SMF Davé et al., 2016) and particularly HI mass (HIMF: Rafieferantsoa & Davé, 2018) that are used to benchmark the accuracy of galaxy formation models

compared to observations. Thus it represents a plausible model to examine the physical processes driving the evolution of HI in galaxies.

1.4 Machine learning

Machine learning is a separate field on its own, but elements of it are now becoming increasingly important in astronomy applications, particularly related to simulations and big data. The rising need to provide big data for scientific analysis, especially related to cosmology, along with the limited capabilities of the current telescopes encourages me to explore the possibility of using such technique particularly for the purpose of predicting the HI content of galaxies.

The basic idea of machine learning is to start with a set of data which contains certain features, and the targets. The features are the available properties of the objects (galaxies in our case) and the targets are the properties we want to know. Very often, there is only 1 target and multiple features. The goal is to learn the relationship, often unknown, between all (or part) of the features and the target during a training phase. Once the relationship is tested to be valid, we can use the trained model on separate data. The separate data only contain the features and the target is to be predicted. If the model is well designed, the predicted targets represent a new set of data that the scientific community can explore.

With the available tools such as `scikit-learn` (Pedregosa et al., 2011) and `TensorFlow` (Abadi et al., 2015), machine learning is accessible for everyone to use. With a fair scientific understanding of the training data, predicting galaxy properties is possible with machine learning. The main advantage of the technique is the accuracy of the machine and that it utilises all the features during the training. The model will pick peculiar variabilities that humans certainly miss in a high dimensional space of features. The downside of machine learning is that the trained model remains a black box and a scientific understanding of the mapping between the features and the target is not obvious. An advantage is that the resulting model will only improve with the increasing amount of data from future surveys.

Surveys like the SDSS provide photometric properties of millions of sources (Abolfathi et al., 2018) while HI surveys such as ALFALFA (Haynes et al., 2018)

only provides thousands of HI fluxes within the SDSS sky of observation. Nonetheless the ALFALFA sample is large enough to cover a reasonable range of HI masses that can teach the machine learning models. Numerical simulations can improve the models by providing information about the galaxies without HI content. The models can provide the information about the missing data in an observed HI sample, and test whether it is simply due to the detection capability of the telescope or something else. Predicted data can also be used to adjust or improve the understanding of the consequences of the interactions between galaxies on their neutral gas content. This thesis aims to build a machine learning framework that can take observed photometric data and predict the galaxy HI contents for these purposes.

1.5 Outline of Thesis

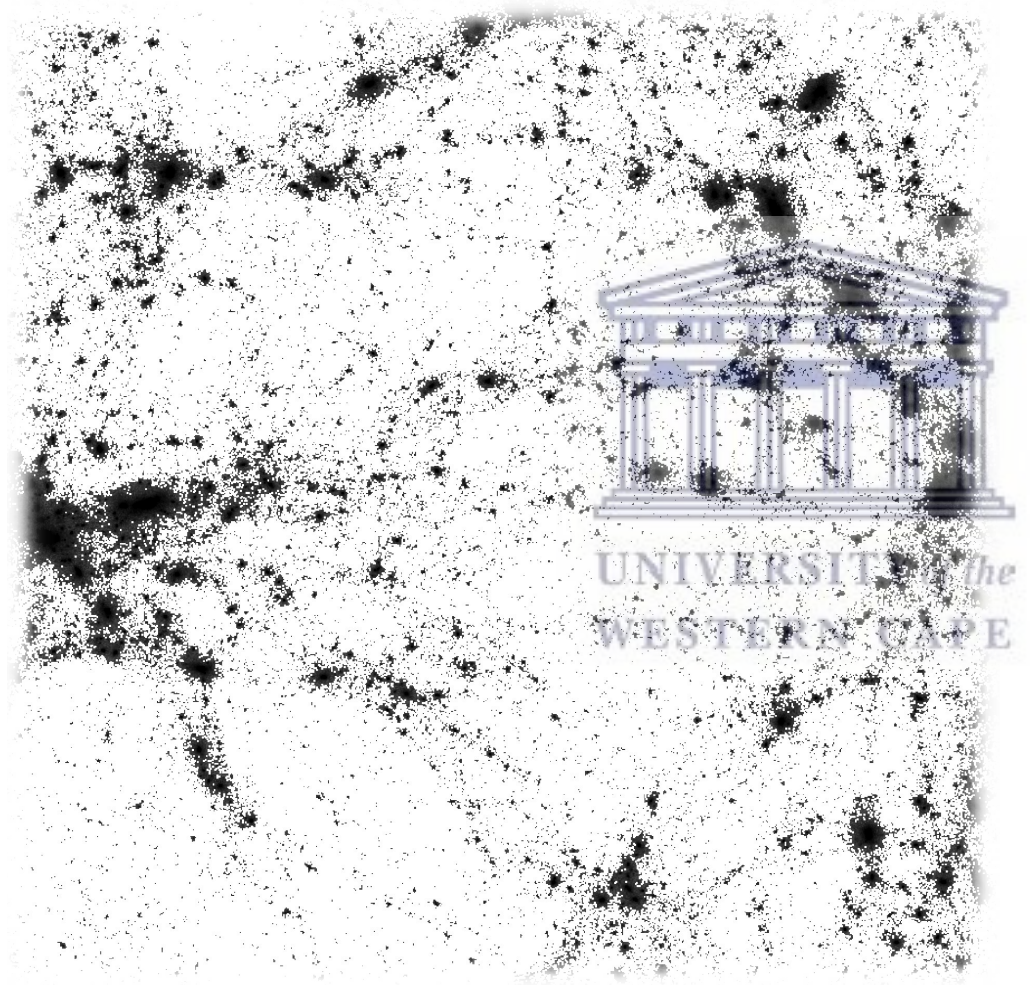
This thesis builds on Rafieferantsoa et al. (2015), which looked at the impact of environment and mergers on the HI content of galaxies. We found a halo mass (environment) dependence of the HI richness (HI mass relative to the stellar mass), the attenuation of star formation is a result of gas stripping and starvation of gas being provided to the galaxy. The e–folding time scales of such event were shown to be $\sim 1 - 3$ Gyr with large scatter. The large scatter suggested the existence of different processes affecting the star formation that we did not account for and resulted in large uncertainties. We hope to explore such processes in this work. The questions we ask are: what are the processes to go from HI to stars to galaxies and what are the different causes that might hinder or enhance those processes? How does the environment affect the HI content? If the galaxies do not contain HI, how will they evolve? How connected are the star formation and the HI content in galaxies? We can address those questions by looking at what the Universe tells us through observational data and concurrently developing galaxy formation models to interpret the observations.

Chapter 2 uses the MUFASA hydrodynamic simulation data to look at the properties of neighbouring galaxies. This will tell us the assembly history of galaxies and the effect of the environmental processes that results in the $z = 0$ distribution of galaxy properties. In Chapter 3, we follow the evolution of galaxy groups simulated at higher resolution than previous MUFASA boxes and look particularly at the HI content and the star formation histories. In Chapter 4,

we propose the use of machine learning techniques to predict HI content of galaxies in a region of the sky where only the photometric information is available. Finally, in Chapter 5, I conclude and outline some future prospects for this work.

This thesis is typset using LaTeX. The numerical programming languages used for this thesis are C (Kernighan, 1988) and Python (Rossum, 1995). Python packages extensively used are NumPy, matplotlib (Hunter, 2007) and pyGadgetReader (Thompson, 2014).

CHAPTER 2



The strength and evolution of galaxy conformity in various tracers

Abstract

We investigate galaxy conformity using the Mufasa cosmological hydrodynamical simulation. We show a bimodal distribution in galaxy colour with radius, albeit with too many low-mass quenched satellite galaxies compared to observations. Mufasa produces conformity in observed properties such as colour, sSFR, and HI content; *i.e.* neighbouring galaxies have similar properties. We see analogous trends in other properties such as in environment, stellar age, H_2 content, and metallicity. We introduce quantifying conformity using $\mathcal{S}(R)$, measuring the relative difference in upper and lower quartile properties of the neighbours. We show that low-mass and non-quenched haloes have weak conformity ($\mathcal{S}(R) \lesssim 0.5$) extending to large projected radii R in all properties, while high-mass and quenched haloes have strong conformity ($\mathcal{S}(R) \sim 1$) that diminishes rapidly with R and disappears at $R \gtrsim 1$ Mpc. $\mathcal{S}(R)$ is strongest for environment in low-mass haloes, and sSFR (or colour) in high-mass haloes, and is dominated by one-halo conformity with the exception of HI in small haloes. Metallicity shows a curious anti-conformity in massive haloes. Tracking the evolution of conformity for $z = 0$ galaxies back in time shows that conformity broadly emerges as a late-time ($z \lesssim 1$) phenomenon. However, for fixed halo mass bins, conformity is fairly constant with redshift out to $z \gtrsim 2$. These trends are consistent with the idea that strong conformity only emerges once haloes grow above Mufasa's quenching mass scale of $\sim 10^{12} M_{\odot}$. A quantitative measure of conformity in various properties, along with its evolution, thus represents a new and stringent test of the impact of quenching on environment within current galaxy formation models.

2.1 Introduction

Environment plays an important role in setting the properties of galaxies. The collapse of massive haloes and large filaments results in the shock heating of gas, which inhibits the growth of galaxies within these structures by a combination of strangulation (or starvation), ram pressure stripping, and tidal stripping. These environmental processes connect the physics of intergalactic gas and large-scale structure with the observable stellar properties of galaxies, and hence represent a key test for cosmologically-situated galaxy formation models.

A particular phenomenon that has gained attention recently is the tendency for galaxies spatially close to each other to have similar galaxy colours. This was first noted in Weinmann et al. (2006a) and given the name *galactic conformity*. They used a galaxy sample from the Sloan Digital Sky Survey (York et al., 2000, SDSS) and found that early-type central galaxies have comparatively higher fractions of early-type satellite galaxies around them, while late-type centrals tended to be surrounded by late-type satellites. Furthermore, the overall fractions of early and late types depended strongly on the mass of their respective haloes. This type of conformity was later dubbed "one-halo" conformity, since it quantifies the level of similarity within a single halo.

Conformity can arise owing purely to large-scale structure heating, because of the tendency of deep potential wells to shock-heat gas to high temperatures (e.g. White & Rees, 1978). Such shock-heated gas is expected to surround galaxies in haloes with masses above $\sim 10^{12} M_{\odot}$ (Birnboim & Dekel, 2003; Kereš et al., 2005; Gabor et al., 2010), and is likely related to the oft-mentioned bimodal distribution in galaxies properties (e.g. Kauffmann et al., 2003; Baldry et al., 2004). Such bimodality is present in both central and satellite galaxy samples. This encompassing hot halo can thus naturally give rise to a correspondence between central and satellite star formation rates and gas contents.

But such a hot halo is not stable. The hot gas in the dense central region is expected to cool quickly, leading to substantially star formation in massive galaxies, in disagreement with observations. Thus most galaxy formation models introduce some feedback mechanism (Somerville & Davé, 2015), putatively from the central active galactic nucleus (AGN), that maintains the hot hydrostatic

halo (Croton et al., 2006). This same "maintenance mode" feedback can attenuate the star formation in satellite galaxies as well (e.g. Ann et al., 2008; Gabor & Davé, 2012). Hence one-halo conformity may encode information both about large-scale structure as well as AGN feedback processes in more massive systems.

Kauffmann et al. (2013) re-analysed galactic conformity in SDSS with a more rigorous sample selection, and found that moderate-mass central galaxies show galactic conformity out to distances that are many times their virial radius, as much as 4 Mpc, which has come to be called two-halo conformity since it corresponds to properties being similar in galaxies living in different haloes. In contrast, more massive centrals only show conformity to the neighbours within about a virial radius. To explore the physical origin of conformity, Kauffmann (2015) further found that central galaxies with low star formation rate are more likely to be located in a neighbourhood with higher fraction of massive galaxies that have active AGN, suggesting that AGN feedback plays a significant role in conformity.

However, the strength or perhaps even the existence of two-halo conformity is controversial. Recent work by Tinker et al. (2018) suggested that the large spatial extent of galactic conformity found in Kauffmann et al. (2013) may be the result of misclassifications of central galaxies, and that conformity beyond at most a Mpc disappears when the central sample is more carefully selected using a group catalog rather than projected distance. This is supported by Sin et al. (2017), who further showed that the strong two-halo conformity signal found by Kauffmann et al. (2013) can be reproduced with their semi-analytic model when using Kauffmann et al.'s selection, but when true central galaxies are used, the two-halo term is very weak and essentially undetectable in the SDSS data.

Conformity analogously appears in the neutral hydrogen content of galaxies. Using 40 galaxies from the Bluedisk project, Wang et al. (2015) found that galaxies with high HI fraction live in the vicinity of other galaxies with high HI fraction. Also, conformity persists out to higher redshift. Hartley et al. (2015) studied the satellite galaxies drawn from the UKIRT Infrared Deep Sky Survey (UKIDSS Lawrence et al., 2007) and concluded that passive galaxies are more likely to be around passive galaxies with 3σ significance and that happens out to $z \gtrsim 2$. Using a set of satellite galaxies data from ZFOURGE

(Straatman et al., 2014), UDS, and UltraVISTA (McCracken et al., 2012), Kawinwanichakij et al. (2016) looked at the evolution of galactic conformity and confirmed a one-halo conformity signal with a significance of more than 3σ out to $z \sim 1.6$, and a lower but noticeable signal out to $z \sim 2.5$. Similarly, Berti et al. (2017) investigated one- and two-halo conformity with the Primus survey (conducted with IMACS; Bigelow & Dressler, 2003) data which they claimed to be uniquely suited due to large survey area of ~ 9 deg^2 and redshift precision of $\sigma_z = 0.005(1+z)$. They detect more than 2.5σ one-halo conformity out to $z \sim 1$. They also found a hint at a two-halo conformity signal from the fact that central galaxies are more likely to be quiescent when they are located in dense environment. Hatfield & Jarvis (2017) used a different method by looking at the cross-clustering of galaxies with the 2-point correlation function and claim that specific star formation rate (sSFR)-density relation, which is another way of characterising conformity, emerges at $z \sim 1$ and keeps growing until today. Hence, it appears that conformity is a real effect not only in present-day galaxy colours, but in gas content at least, as well as to higher redshifts, though the precise strength and evolution depends on the sample selection and technique used to quantify conformity.

Given the emergence of this wealth of data on galaxy conformity, there has been various attempts to explain conformity within a hierarchical structure formation paradigm. Hearin et al. (2015) used semi-analytic models on the Bolshoi simulation (Klypin et al., 2011) and found that using either $M_{\mathrm{halo},vir}$-based quenching prescription or a delayed-then-rapid quenching of the galaxies displayed zero conformity, while only their *age matching* model showed statistically significant galactic conformity. This led them to conclude that two-halo conformity is the result of the central galaxy assembly bias. Hearin et al. (2016) followed up by looking at the assembly histories of structures and found that haloes separated with more than tens of their virial radius are connected because they are situated within the same large-scale tidal environment which is the main driver of their growth. However, Zu & Mandelbaum (2018) used a colour-based halo occupancy model to argue that (weak) large-scale conformity can arise purely from the environmental dependence of the halo mass function, without requiring any assembly bias. They also confirm the result of Sin et al. (2017) and Tinker et al. (2018) that the strong two-halo conformity seen by Kauffmann et al. (2013) is primarily an artifact of mis-identified central vs. satellite galaxies.

Modern cosmological hydrodynamic simulations include all the relevant effects that are expected to give rise to galactic conformity. Such models should, in principle, implicitly include halo assembly bias, halo occupancy evolution, and any correlations between the colours of centrals and satellites arising from included feedback processes, and hence galaxy conformity should be an emergent property. Such simulations track gas and star formation properties directly as well, so can be used to study conformity in various tracers, as well as their evolution with redshift. Nonetheless, as shown in e.g. Gabor & Davé (2015), the quenching of satellites along with surrounding "backsplash" and "neighbourhood quenched" galaxies is dependent on the model for central galaxy quenching, which is at present not well constrained in galaxy formation models, and is typically included only in a prescriptive or heuristic manner. Hence it is instructive to examine conformity predictions from cosmological hydrodynamic simulations, and compare them to present and future observables as a detailed test of quenching models.

To this end, Bray et al. (2016) used the Illustris Simulation (Vogelsberger et al., 2014) to explore whether assembly bias can be an explanation for galaxy conformity. They found evidence of galactic conformity out to 10 Mpc for the smallest centrals, decreasing in strength with increasing stellar mass of the centrals. They developed a simple model based on abundance and age matching that was able to reproduce this signal, demonstrating that the galaxy colour-age relation is important for conformity.

In this paper, we use the Mufasa simulation (Davé et al., 2016) to study conformity. Our goal is to investigate conformity in a wide variety of galaxy tracers, to understand which galaxy and environmental parameters show the strongest levels of conformity, and to make predictions for conformity that can be used as a test of models. This differs in aim from previous works discussed above that have focused more on developing halo-based models for the origin of conformity, comparing to the (sparse) available data. We show that Mufasa predicts a conformity signal that is strongly dependent on halo mass, and that this signal is not limited to colour and neutral hydrogen but appears in many other galaxy properties as well. We propose a measure to quantify the strength of conformity and study its evolution in various tracers as a function of redshift. Our results indicate that galaxy conformity is a generic emergent feature of hydrodynamic galaxy formation models, and that the strength of conformity can

be a valuable test of the interplay between environment, galaxy assembly, and feedback processes particularly related to quenching.

§2.2 briefly reviews the Mufasa simulation used for this work. In §2.3, we look at the satellite galaxy properties from our simulation. §2.4 expands on the comparison of conformity between our simulated galaxy sample and the observed data. §2.5 characterises the nature of conformity comparing between various tracers, and study its evolution out to intermediate redshifts. We summarize our conclusions in §3.7.

2.2 Simulations

2.2.1 Models

For this work we employ the Mufasa simulation, which is fully described in Davé et al. (2016). Here, we briefly review the main ingredients, and expound on the key physical aspects of Mufasa that are particularly relevant for this work.

Mufasa employs the Gizmo cosmological hydrodynamic code, including a tree-particle-mesh gravity code based on Gadget (Springel, 2005), and a meshless finite mass hydrodynamic algorithm (Hopkins, 2015). For radiative processes, Mufasa utilises the Grackle 2.1 library[1] to cool the gas elements, accounting for non-equilibrium ionisation for primordial elements, as well as metal-line cooling assuming ionisation equilibrium, plus photoionisation heating computed with a spatially-uniform metagalactic flux taken from Faucher-Giguère et al. (2010). Star formation occurs in molecular gas, and only in gas elements with hydrogen number density $n_H \geq 0.13$ cm^{-3}, with the star formation rate computed following a Schmidt (1959)-law scaling, namely

$$\mathrm{SFR} = \varepsilon f_{\mathrm{H}_2} G^{-0.5} \rho_{\mathrm{gas}}^{0.5}. \tag{2.1}$$

Here ρ_{gas} is the density of the gas, $\varepsilon = 0.02$ (Kennicutt, 1998) is the star formation efficiency, G the gravitational constant, and f_{H_2} the molecular hydrogen

[1] https://grackle.readthedocs.io/en/grackle-2.1/genindex.html

fraction in the gas element computed via a subgrid prescription from Krumholz & Gnedin (2011).

We assume that star formation in the simulation produces a combination of radiation pressure and supernovae energy which manifests by kicking out its surrounding gas volume elements at a given rate η relative to the star formation rate. Each outflowing wind element is ejected away from its host galaxy and in a direction perpendicular to the ($\vec{\mathbf{v}}$, $\vec{\mathbf{a}}$) plane with a launching wind velocity v_w, where $\vec{\mathbf{v}}$ and $\vec{\mathbf{a}}$ are the velocity and the acceleration of the gas cloud prior to its launch. To choose the free parameters, we take scaling relations from Muratov et al. (2015) based on the Feedback in Realistic Environments (FIRE) simulations. In particular, we choose the mass loading factor η and the wind speed v_w as follows:

$$\eta = 3.55\left(\frac{M_*}{10^{10}\,\mathrm{M}_\odot}\right)^{-0.35}; \tag{2.2}$$

$$v_w = 2v_c\left(\frac{v_c}{200\,\mathrm{km\,s^{-1}}}\right)^{0.12}. \tag{2.3}$$

M_* and v_c are the stellar mass and circular velocity of the galaxy where the gas volume element is located. Galaxies are identified using an approximate on-the-fly friends-of-friends group finder specifically designed to be computationally fast and tuned to reproduce the same results as SKID[2] (Spline Kernel Interpolative Denmax); this is applied only to star-forming gas elements and stars. I note that the galaxy sample from skid does not contain dark matter depleted galaxy (see Figure 2.1).

Particularly relevant for this paper is that MUFASA includes an observationally motivated heuristic prescription to quench massive galaxies. Gas elements sitting in a host halo above a threshold quenching mass M_q are heated to around the virial temperature of that host. This is done except for the interstellar medium gas defined to have more than 10% neutral hydrogen fraction. The host halo is grouped on the fly with a (separate) friends-of-friends algorithm using a linking length of 0.16 times the mean inter-particle distance, including dark matter, gas, and stars. The virial temperature is taken to be $T_{vir} = 9.52\times10^7 M_h^{2/3}$ (Voit, 2005). M_q is taken to be redshift dependent, whose scaling is taken from the analytic equilibrium model of galaxy formation (Mitra et al., 2015), who obtained a best-fit scaling of $M_q = (0.96 + 0.48z)\times10^{12}\,\mathrm{M}_\odot$.

[2] http://www-hpcc.astro.washington.edu/tools/skid.html

We note that this quenching prescription is purely heuristic, and is not a physical model for AGN feedback. It is specifically designed to mimic the effects of radio mode feedback (Croton et al., 2006) from active galactic nuclei (AGN), in which jets from the central galaxies of massive haloes are observed to add enough energy into diffuse gas to counterbalance cooling (McNamara & Nulsen, 2007), without a detailed physical model for the AGN energy couples to the halo gas. Since we keep ambient gas hot all the way out to the virial radius in massive haloes, this feedback model can be regarded as a rather extreme form of maintenance mode quenching. In Davé et al. (2017a) we have showed that it results in a population of quenched central galaxies that is in reasonable agreement with key observations such as the colour-magnitude diagram of galaxies, although it appears to overproduce quenched satellite galaxies at low masses. The results on conformity here should be taken with these caveats in mind, noting that a different or more physical quenching model may yield different results.

2.2.2 Galaxy sample and operational definitions

The galaxy sample used for our analysis is obtained by simulating a cube of $50h^{-1}$Mpc on a side with 512^3 dark matter particles and 512^3 gas volume elements. The initial conditions are generated at redshift $z = 249$ using Music (Hahn & Abel, 2011) with Planck et al. (2016)-concordant cosmological parameters, namely $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, $\Omega_b = 0.048$, $H_0 = 68\ \mathrm{km\,s^{-1}\ Mpc^{-1}}$, $\sigma_8 = 0.82$ and $n_s = 0.97$. We also consider a $25h^{-1}$Mpc volume with the same number of particles and the same input physics and cosmology, having a factor of 8 better in mass resolution, in order to assess resolution convergence.

Mufasa evolves these initial conditions to $z = 0$ outputting 135 snapshots. For each snapshot, we identify galaxies, with Skid[2] (Kereš et al., 2005), as gravitationally bound collections of stars and star-forming gas. For galaxy mass resolution, we take $5.8 \times 10^8 \mathrm{M}_\odot$, where it was shown that the stellar mass functions converge with that limit (Davé et al., 2016). While this does not guarantee convergence in other properties such as colour, we will compute conformity using galaxies with $M_* > 1.8 \times 10^9 M_\odot$, three times larger than our nominal mass resolution and corresponding to approximately 100 star particles at minimum. Host haloes are identified using the friends-of-friends algorithm with a linking

length of 2% the mean inter-particle distance. The galaxies are assigned to host haloes based on their spatial location. The most (stellar) massive galaxy in a given halo is considered as the central (regardless of physical location) and the remainder are satellites.

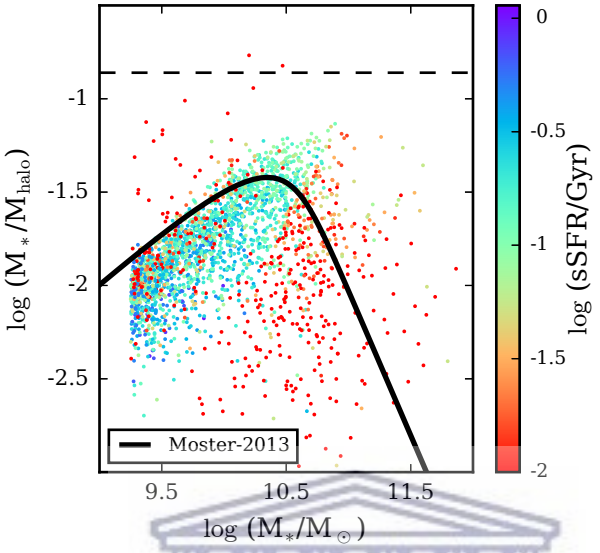


FIGURE 2.1: Stellar-to-halo mass ratio as a function of M_* for central galaxies in MUFASA. The ratio peaks around $M_* \sim 10^{10.5} M_\odot$ at around 25% of the cosmic baryon fraction (shown as the horizontal dotted line), dropping away quickly to higher and lower masses. The solid line shows the fit from abundance matching of observed galaxies by Moster et al. (2013); the predicted values are generally consistent with this.

Figure 2.1 shows the stellar-to-halo mass ratio in central galaxies as a function of stellar mass, colour-coded by specific SFR. This shows that quenching kicks in rather rapidly at $M_* \gtrsim 10^{10.5} M_\odot$, above which the typical halo mass lies in the regime where MUFASA's halo quenching model kicks in. There are a handful of galaxies with high $M_*/M_{\rm halo}$ ratios at lower masses, which tend to be quenched; these are typically former satellites whose orbits have taken them outside their host haloes (Gabor & Davé, 2015), and thus have had their own haloes significantly impacted by stripping. Such galaxies should also appears in observational samples, and indeed there is strong increase in the passive fraction of galaxies lying close to but still outside the virial radius of massive galaxies (Geha et al., 2012). These have only a small impact on the overall conformity statistics.

The properties of these galaxies and haloes are calculated with a modified version of CAESAR[3], which is an add-on package for the YT simulation analysis suite. The stellar mass (M_*) of a galaxy is obtained by summing the masses of all its stellar particles, and we define the galaxy's age as the time when half of these stars were formed. The galaxy star formation rate (SFR) is the summation of the instantaneous SFR of the gas elements (directly obtained from the simulation). The molecular hydrogen fraction (f_{H2}) is the total mass of H_2 from all the gas elements, which is tracked directly in the simulation, divided by its stellar mass. The atomic hydrogen (HI) content of the galaxy is the aggregate amount of all HI from the gas particles. Before summing over, the gas HI mass from the simulation is post-processed to account for the self-shielding from the metagalactic UV background radiation, by using the fitting formula for the effective optically-thin photoionization rate as a function of density taken from Rahmati & Schaye (2014, see eq. 1). The HI richness (f_{HI}, or HI fraction) is the total HI content of the galaxy divided by its stellar mass M_*. To quantify the environment, we use the projected nearest neighbour density Σ_3:

$$\Sigma_3 = \frac{3}{\pi R_3^2} \tag{2.4}$$

where R_3 is the distance of the galaxy to its 3rd closest neighbour, projected along the z-axis. The minimum mass of galaxies considered in this calculation is the stellar mass threshold mentioned below.

The colours of the galaxies are obtained using the Line Of Sight Extinction by Ray-tracing (LOSER)[4] package. Stellar spectra are interpolated from the age and the metallicity of star particles using the Flexible Stellar Population Synthesis (FSPS; Conroy & Gunn, 2010) library. The metal column density is calculated along each line of sight, converted into a dust extinction, then applied to each star particle's spectrum. The spectra of all the stars in each galaxy (from SKID) are then summed and the appropriate filter applied to get the magnitudes. See Davé et al. (2017a) for further details.

To analyse conformity, we follow the general procedure outlined in Kauffmann et al. (2013). First we subdivide our $z = 0$ central galaxies into three stellar mass

[3] https://bitbucket.org/laskalam/caesar
[4] https://bitbucket.org/romeeld/closer

bins: $\log(M_{*,\mathrm{cen}}/\,\mathrm{M}_\odot) \in \{[9.5,10],[10,10.5],[10.5,11.5]\}$. The numbers of galaxies in each bins are 747, 587, and 561, respectively. We choose $10.5-11.5$ for the largest bin (instead of $11-11.5$ that was used in Kauffmann et al. 2013) because using smaller bin results in considerable shot noise owing to small numbers of such galaxies in our simulation; we checked that the results for our larger bin are consistent with that obtained from using only $M_* > 10^{11} M_\odot$, as both predominantly live in quenched haloes.

Within each stellar mass bin, we order the central (or primary) galaxies by a given property: colour, sSFR or HI richness ($\mathrm{M_{HI}}/M_*$). We then take the objects at the lowest and the highest quartiles ($< 25\%$ and $> 75\%$), and examine the median properties of neighbours of these galaxies as a function of radius out to 4 projected Mpc. Throughout our analysis, we employ jackknife resampling among 8 simulation sub-octants to estimate our errors.

Guided by the completeness level in the SDSS-based sample of Kauffmann et al. (2013), we only consider satellite galaxies with $M_* \geq 10^{9.25} M_\odot$, comfortably above our stellar mass resolution limit. However, unlike Kauffmann et al. (2013), we use friends-of-friends identified central galaxies instead of adopting their isolation criteria. This is more closely aligned with more recent analyses that use group catalogs, which has been shown to provide a more robust measure of conformity, particularly two-halo conformity, compared to the Kauffmann et al. isolation criterion (Tinker et al., 2018).

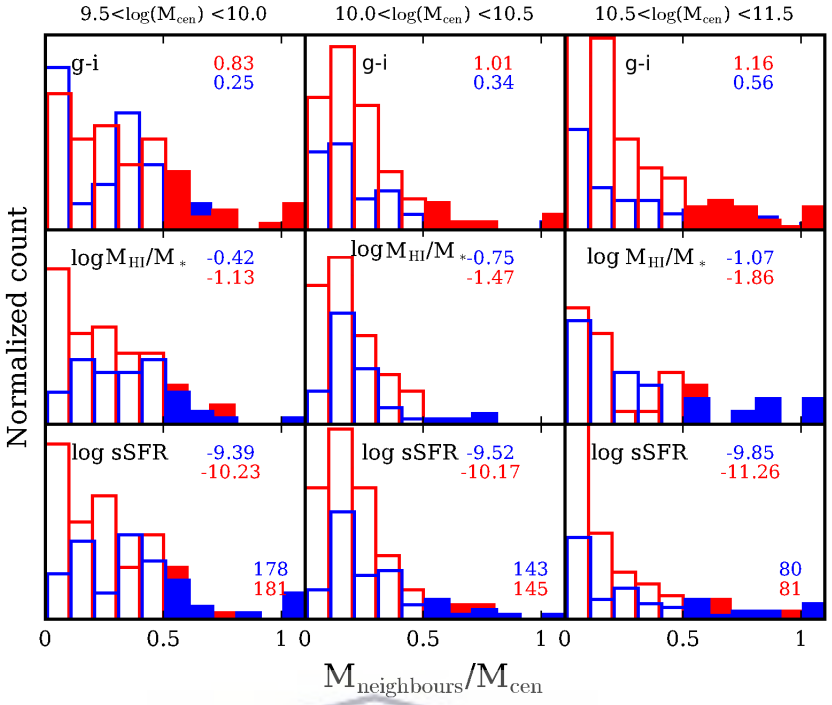


FIGURE 2.2: Stellar mass fraction of the neighbours relative to their respective central galaxies in our simulated box. We only show galaxies that are within 500 kpc projected distance and 500 $\mathrm{km\,s^{-1}}$ redshift distance from their primaries. The histograms at > 0.5 are filled to emphasize on the fractions of neighbours where their respective central galaxies would not have been classified as isolated galaxies if Kauffmann criteria were applied.

Figure 2.2 illustrates the difference in mass ratio of identified centrals to their neighbouring galaxies located within 500 projected kpc and 500 $\mathrm{km\,s^{-1}}$ redshift distance, which corresponds to the Kauffmann isolation criterion. The distribution of galaxies in the upper and lower quartiles in galaxy colour (top row), HI fraction (middle), and sSFR (bottom) are indicated in red and blue histograms, in three central mass bins (left to right columns). The numbers on the bottom right show the number of central galaxies in the respective stellar mass bin. This shows that overall a small fraction of the neighbours have $M_* > 0.5 \times M_{primary}$ (filled part of the histograms) where their respective central galaxies would not have been classified as isolated with respect to Kauffmann criteria. Given the relatively modest fraction of such neighbours ($\sim 15\%$) our conformity should be a robust prediction independent of isolation criterion, but for high-mass systems there will be systematic differences. Note that if we only used the satellite galaxies (not shown), the fraction of neighbours with $M_* > 0.5 \times M_{primary}$ drops significantly to $\sim 5\%$. We will therefore compare to

the Kauffmann et al. (2013) data for illustrative purposes only, with the caveat that the isolation criterion can play some role in the outcome.

2.3 Satellite galaxy properties in Mufasa

Conformity is a measure of the environmental impact on galaxy formation. Hence it is important to ensure that environmental processes are reasonably accurately modeled in our simulation. As a test of this, we begin by presenting an analysis of the satellite galaxy population predicted in Mufasa.

2.3.1 Satellite versus central mass functions

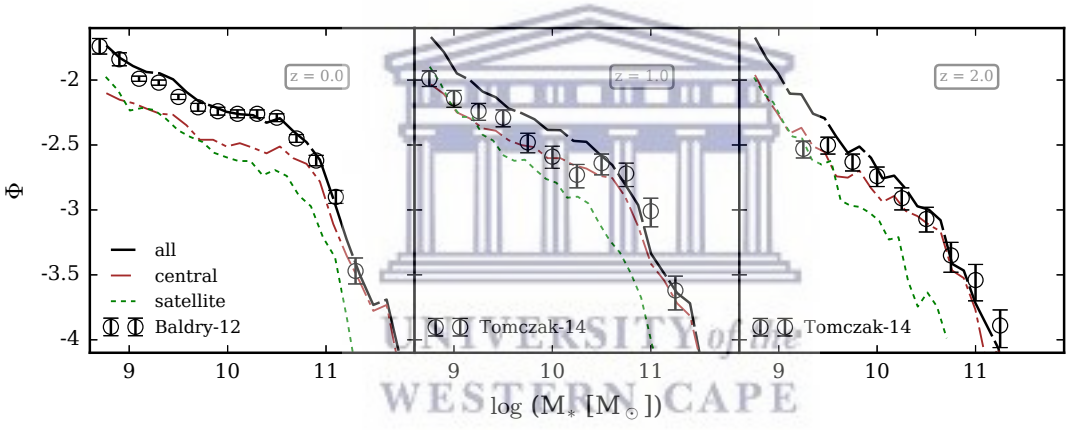


Figure 2.3: Galaxy stellar mass functions (black long-dashed lines) separated into central (brown dashed-dotted lines) and satellite (green dotted lines), at redshift $z = 0$ (left), $z = 1$ (middle) and $z = 2$ (right). At all epochs, central galaxies dominate by number at $M_* \gtrsim 10^{9-9.5} M_\odot$, and a knee begins to appear in the satellite mass function at $z \lesssim 1$. Observations of the total GSMF are shown with the black circles.

Figure 2.3 shows the GSMF of the Mufasa simulated galaxies (black long-dashed lines): separated into central (brown dashed-dotted lines) and satellite (green dotted lines) galaxies at $z = \{0, 1, 2\}$. Observational data from Baldry et al. (2012) for $z = 0$ and Tomczak et al. (2014) for $z = \{1, 2\}$ are shown with the black circles.

As discussed in Davé et al. (2016), Mufasa reproduces the observed total GSMF and its evolution out to high redshifts fairly well, albeit with a modest excess at $z \sim 1$. The massive end of the GSMF consists almost entirely of central galaxies

across all redshift explored here. The contribution from the satellites is present only at the low mass end ($M_* \lesssim 10^{9.5} M_\odot$) and it is modestly stronger at lower redshift.

Already at $z = 2$, the central GSMF starts to produce the knee while the satellite galaxies only show a hint of a knee at $z \lesssim 1$. The knee in GSMF is the result of an enhancement in star formation due to an increased contribution of wind recycling to higher masses (Oppenheimer et al., 2010), combined with the truncation of star formation in massive galaxies owing to our quenching prescription mimicking AGN feedback. At later epochs, the most massive satellites were until recently central galaxies that were massive enough to experience quenching, and hence they reflect a truncated high-mass GSMF of the centrals.

2.3.2 Star-forming versus quiescent satellite mass functions

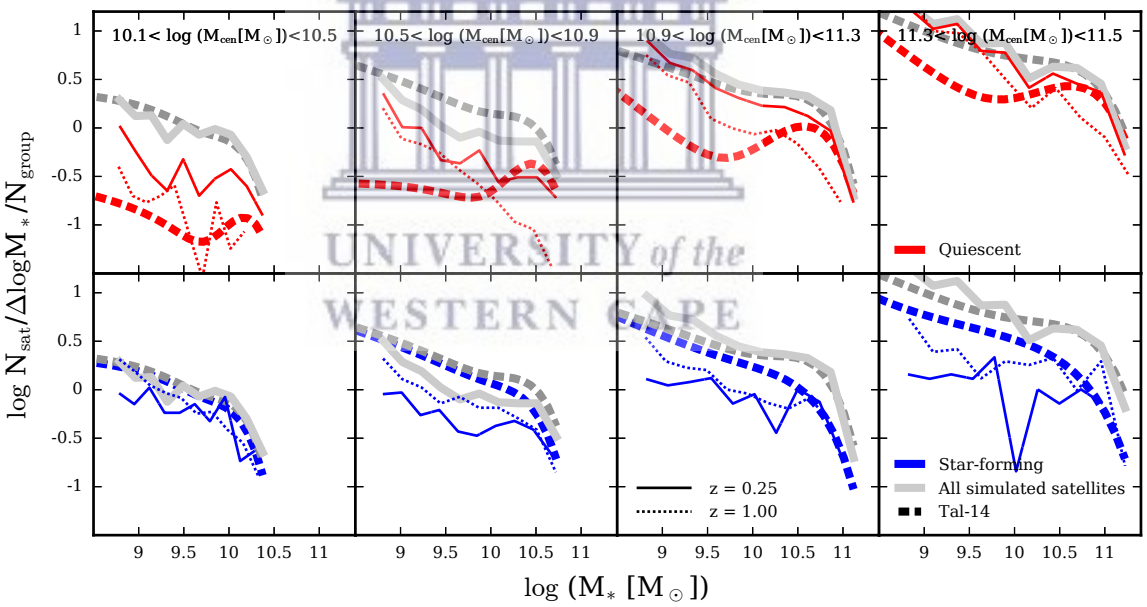


FIGURE 2.4: Satellite galaxy stellar mass functions (thick grey lines, $z = 0$); separated into star forming (lower panels, blues) and quenched (upper panels, reds) population. We show them at two different redshift $z = 0.25$ (thin solid lines), and $z = 1$ (thin dotted lines). The thick dashed lines show the double-Schechter fits from Tal et al. (2014, see Table 1): blue for star forming, red for quenched and gray for all satellites.

Figure 2.4 breaks down the GSMF in terms of satellites split into star-forming and quiescent within four central galaxy stellar mass bins, following the same

exercise as in Tal et al. (2014) using UltraVista data. Here we consider galaxies to be quiescent if $\log\left(\mathrm{sSFR/Gyr^{-1}}\right) \leq -2.2$. In accordance with Tal et al. (2014), we define Φ to be the number of galaxies per group. In other words, for the sample of all the N central galaxies within a given stellar mass bin, Figure 2.4 shows the distribution of their satellite galaxies normalized by N. In their analysis, Tal et al. (2014) defined galaxies to be central if no other more massive galaxy could be found within two projected virial radii, which is broadly consistent with our definition to be the most massive galaxies in their respective haloes (see sec. 3.2.2). They consider all non-central galaxies within two virial radii of the central galaxies to be their satellites.

The total predicted satellite GSMF, as indicated by the thick solid grey lines, are in general agreement with the Tal et al. (2014) observational data shown by the thick dashed gray lines. There is a clear trend of a higher mass function of satellites around more massive centrals. The only notable discrepancy is that the observations show up to $\sim \times 2$ more satellites at intermediate masses in the $10^{10.5} - 10^{10.9} M_{\odot}$ central stellar mass bin. At higher central mass bins, simulations tend to slightly have more satellites than the observations at the lowest mass ends. Hence overall, the total number of satellites is in broad agreement, with a hint that Mufasa underproduces the satellite population at intermediate masses ($\sim 10^{10} M_{\odot}$).

In contrast, we see more substantial discrepancies when examining the satellite GSMFs broken down by quiescent vs. star-forming. At the lowest central masses, most satellites are blue (star-forming), and Mufasa reproduces those well. However, Mufasa strongly over-predicts the (small) number of red satellites, particularly at the lowest satellite masses. At higher central masses, the observed red satellite mass function is relatively shallow (thick red dashed lines), albeit with an upturn at the lowest masses, whereas Mufasa predicts a steeper red satellite GSMF. Particularly, for the most massive centrals, Mufasa predicts that red satellite galaxies dominate at all masses, while in Tal et al. (2014) they only dominate at $M_* \gtrsim 10^{10.5} M_{\odot}$.

The overall redshift evolution predicted in Mufasa is qualitatively consistent with that seen by Tal et al. (2014), where they used three redshift bins such as $(0.2, 0.5)$, $(0.5, 0.85)$ and $(0.85, 1.20)$. For illustration, we only show the simulated sample at $z = 0.25$ (thin solid lines) and $z = 1$ (dotted lines) which span

the redshift ranges used in their sample. The star-forming satellite mass functions show very little evolution with time, with a hint of being slightly higher at higher redshift. They dominate the mass function at low masses, and are more prevalent at earlier epochs. Meanwhile, there is a mild increase with time for the quenched satellite population, but the trend with mass is much more significant; there is a much larger number of red satellites per group around massive centrals, and at high masses they dominate over the star-forming satellites. These general trends are qualitatively in agreement with the observations. However, as noted before, Mufasa strongly over-produces the number of low-mass red satellites, at essentially all central masses. Moreover, Mufasa tends to grow massive red satellites more rapidly than the low-mass red satellites, which is opposite to what is seen in the data in which the most massive satellites are already in place fairly early on.

Several avenues could lead to these discrepancies. For instance, Mufasa identifies central galaxies via a 3-D friends-of-friends scheme, while in Tal et al. (2014), centrals are identified as having no other more massive system within two projected virial radii, which can blend systems in projection. In particular, it could be that the "bump" of massive red satellites identified by Tal et al. (2014) are actually nearby massive central galaxies. Furthermore, Tal et al. (2014) must do substantial background subtraction to count satellites, which can be uncertain particularly when low numbers of satellites are present. We could in principle mimic the first selection effect, but the second effect would require detailed modeling of the background population, which is beyond the scope of this work. In general, without a more careful mimicking of the observational selection effects, the significance of these discrepancies is not completely clear.

If we take the discrepancies at face value, it suggests that Mufasa over-quenches satellites in haloes of all masses, particularly low masses satellites. This is consistent with the findings in Davé et al. (2017a), but here we see this across a range of mass bins and epochs. The discrepancy for high-mass centrals may owe to the extreme nature of our quenching mechanism, where we keep gas hot all the way out to the virial radius. For low-mass centrals, however, the quenching mechanism is not obviously relevant, except if such low-mass centrals happen to be in the vicinity of high-mass centrals ("neighbourhood quenched";

Gabor & Davé, 2015). Hence the discrepancies may reflect details of hydrodynamic processes, or else some of the systematic effects mentioned in the previous paragraph. For now, we will consider the broad agreement as sufficient for examining galaxy conformity in Mufasa, with the caveat that Mufasa does not well reproduce the low-mass satellite population when broken up by colour.

2.3.3 Halo-centric satellite colours

Conformity was first specified as a commonality in colours between nearby centrals and satellites. Hence an important aspect to investigate in our models is the colours of satellite galaxies as a function of radius. In this section we examine the halo-centric colours of satellite galaxies.

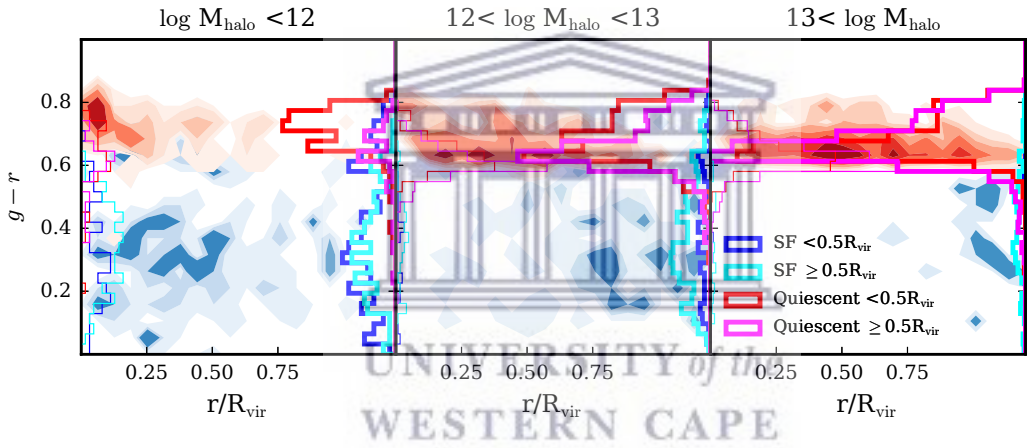


Figure 2.5: $g-r$ colour of neighbouring galaxies depending on their distances to the centre of their corresponding haloes ($z = 0$). The red (blue) shaded area shows the distribution of the quiescent (star forming) satellites in our simulated sample. The histograms along the right y-axes show the distribution of the satellites depending on their specific star formation rate and their distance to the halo centers. We also show the colour distributions of our smaller box (higher resolution, but poorer statisctics) on the left y-axes.

Figure 2.5 shows $g-r$ contour plots of all satellites in Mufasa, as a function of radius scaled by the virial radius R_{vir}. The red shaded area shows the quiescent satellites, while the star-forming satellites are depicted by the blue shaded area; we remind the reader that these are divided at sSFR=$10^{-2.2}$ Gyr^{-1}. The right y-axis histograms show the colour distributions of the satellites: blues (reds) for star forming (quiescent) satellites at $r < 0.5R_{vir}$, and cyan (magenta) for star forming (quiescent) satellites at $r \geq 0.5R_{vir}$. We show for three different

halo mass bins. As a test of numerical convergence, we also show the colour distributions of our $25h^{-1}$Mpc box in thin histograms on the left y-axis.

We can see a strong bimodality distribution in terms of colour that extends out to the virial radius. This bimodality exists for all halo masses, although the relative number of red and blue satellites changes substantially with halo mass. There are a few dusty star-forming satellites that have red colours lying underneath the red contour, but these are small fraction of the total. The strong bimodality in colour was first noted observationally by Kauffmann et al. (2003); Balogh et al. (2004), and is interpreted to indicate that satellite galaxies quench fairly rapidly once the quenching process begins (e.g. Wetzel et al., 2015). Mufasa broadly reproduces these observed trends. The thin histograms from the $25h^{-1}$Mpc volume along the left y-axis are qualitatively similar, although the bimodality is somewhat weaker maybe because the smaller volume does not produce as many massive quenched systems and the model performs differently at that resolution. Despite the periodic boundary approach used in our simulations, smaller boxes always produce smaller cosmic structures than bigger boxes. Therefore, the formation of quenched systems, with our halo-mass based quenching prescription, is affected by the size of the box.

Examining the trends with halo size, we clearly see a decrease of star forming population towards larger halo masses. For less massive haloes, we see that quiescent satellites are mostly located at the core (inner $\sim 10-15\%$) of the groups while the star forming satellites are almost evenly distributed. For intermediate-mass haloes, the star forming satellites inside $< 0.5R_{vir}$ rarify and the quiescent galaxies extend out to $\geq 0.5R_{vir}$. For the most massive haloes, only the satellite galaxies at the very edge of the haloes are forming stars. These trends are again qualitatively consistent with observations, as more massive central galaxies tend to be quenched and have more environmentally-quenched satellites (e.g. Peng et al., 2012).

Overall, these results taken together generally indicate that the satellite population in Mufasa qualitatively reproduces observations, including the satellite population evolution split by red versus blue galaxies. The most notable discrepancy is that Mufasa predicts an excess of low-mass red satellites in massive haloes. Broadly, this is expected to strengthen the conformity of red centrals with red neighbours in massive haloes, hence the predicted conformity is likely to be overestimated in this regime. Modulo this caveat, Mufasa provides

a fairly viable platform to study the radial distribution of galaxy properties around central galaxies.

2.4 Conformity in $_{s}$SFR, HI richness and colour of the galaxies.

Traditionally, galaxy conformity is known as the tendency for central galaxies and their satellites to have similar colours (Weinmann et al., 2006a). However, conformity could in principle be associated with any galaxy property. For instance, conformity has recently been quantified in neutral hydrogen (Kauffmann et al., 2013), with central and satellite galaxies found to be similar in their HI richness. One could equivalently define conformity between different galaxy properties. For example, one could quantify by how much blue central galaxies have higher HI satellites, or quantify how older central galaxies have lower sSFR satellites. We will call this *cross-conformity*, differentiated from *auto-conformity* (or just, conformity).

In this section, we examine galaxy conformity in MUFASA in terms of colour, HI content and sSFR. We also make predictions for cross-conformity among these properties. Here we will take bins of central galaxy stellar mass in order to be able to compare to observations particularly of Kauffmann et al. (2013), but we forego a detailed matching of selection for particular samples that can be critical for proper quantitative interpretation (as discussed in §3.1). Instead, we focus on the nature and strength of conformity as predicted in MUFASA.

Using the approach of Kauffmann modulo our definition of central galaxies, the closest neighbours are satellites, representing one-halo conformity, and those farther out are other haloes' central and satellite galaxies representing two-halo conformity. Hence for each neighbour property, we generate a plot of three mass bins, which we show in columns, and three central galaxy properties, which we show as rows. We will consider conformity in the neighbour properties of colour, sSFR and HI richness; thus we have three such plots.

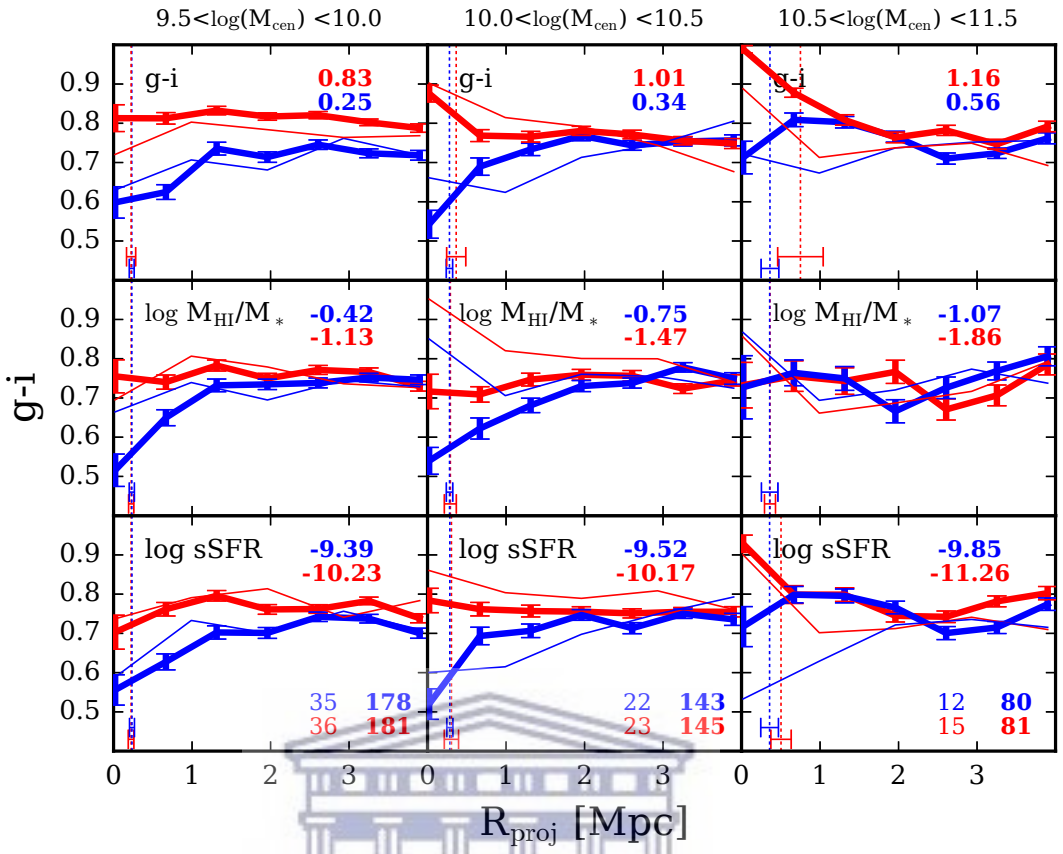


FIGURE 2.6: Median *g-i* colours of galaxies as a function of projected distances around central galaxies divided in three stellar mass bins (columns). Within each stellar mass bins, the red/blue curves denote the median colours of galaxies around the upper/lower quartiles of central galaxies in g-i (top panels), HI richness (middle panels), and sSFR (bottom panels). The median values of each quartiles of the central galaxies are shown on top of each panel. The number of central galaxies in each stellar mass bins are shown on the bottom right of the bottom panels: the lower (upper) quartile is in red (blue) (normal font: m25n512, bold font: m50n512). The thick lines show our fiducial box (m50n512) and the thin lines our smaller (higher resolution) box (m25n512). The errorbars are estimated with jackknife resampling. The vertical dashed lines show the median values of the $R_{\rm vir}$ of the hosts of the central galaxies (colour coded), with the errorbars at the bottom left quantifying the standard deviation. The top row represents the conformity signal in *g-i*, while the bottom two rows represent the cross-conformity of centrals' HI richness and sSFR with satellites' *g-i*.

Figure 2.6 shows the first of these plots, depicting the $g-i$ colour of the neighbouring galaxies as a function of projected distance to their respective centrals $R_{\rm proj}$. The error bars are from jackknife resampling. The top row represents traditional conformity: the tendency for neighbouring galaxies to share the colour of their central galaxy. The median colour of central galaxies in the top and bottom quartiles of colour are shown as the values in the upper right,

for each mass bin. The values at the bottom right of the bottom panels show the number of central galaxies for each stellar mass bins (color coded, where the normal (bold) fonts are for m25n512 (m50n512)). The thick lines show our fiducial box (m50n512), but we also show our smaller (higher resolution) box (m25n512) with the thin lines for comparison. The vertical dotted lines (colour coded) show the median values of the virial radii of the central galaxies with 1σ uncertainties shown towards the bottom of each panel. We show the plot out to $R_{\rm proj} = 4$ Mpc, beyond which point we will later show the conformity signal essentially disappears in all tracers.

In the lowest mass bin, conformity is evident at all scales; the central galaxies have a median $g-i = 0.83$ magnitudes, and the neighbours tend to have a similar colour all the way out to 4 Mpc. The colour of the bluest quartile of centrals is $g-i = 0.25$, but here the neighbours are generally redder than the central, with $g-i \approx 0.6-0.7$; nonetheless they are still clearly bluer than the neighbours of red centrals.

The intermediate and high-mass bins show conformity as well, but restricted only to within $R_{\rm proj} \lesssim 1-2$ Mpc, with a trend for less extended conformity in larger galaxies. The conformity is also relatively weak, as the colour difference in the neighbours is $\lesssim 0.2$ magnitudes, while the colour difference between the centrals in the highest and lowest quartiles is much larger, $\sim 0.5-0.7$.

Thus Mufasa clearly shows evidence for colour-colour conformity. While it weakens with projected distance, it still extends well beyond $R_{\rm vir}$, and thus we predict two-halo conformity to exist albeit with a strength that rapidly diminishes with radius. We will quantify the strength of conformity in §2.5. It could be that the two-halo conformity reflects only backsplash satellites whose orbit takes them beyond the nominal virial radius, or neighbourhood-quenched galaxies (Gabor & Davé, 2015); we will investigate the nature of these galaxies in detail in future work.

The middle row shows the level of colour conformity when selecting central galaxies by HI fraction. Cross-conformity between HI richness and colour is equivalently evident as for colour-colour (auto-)conformity. Here, at low masses the cross-conformity only extends to ~ 2 Mpc, which is similar to what is seen at intermediate masses.

Finally, in the bottom panel, we show the cross-conformity between sSFR and colour. For an extinction-free stellar population, these two would essentially be identical. We have included extinction in computing our colours, so the similarity in trends with the colour-colour conformity is not completely trivial, but nonetheless the trends do look quite similar. As such, herein we will primarily consider sSFR as a proxy for colour.

An interesting note is that for low and intermediate mass galaxies, the bluest, most HI-rich, or most star-forming galaxies all have close neighbours that are significantly bluer than faraway neighbours. However, in the most massive bin, this turns around, and the satellite colours become redder moving in from $R_{\rm proj} \lesssim 1$ Mpc. The trend in colour is also present but somewhat weaker than in sSFR, since it is partially offset by decreasing extinction from the lower gas content at small radii. This shows the strong effect that our halo-based quenching model has on truncating the gas content and star formation in satellite galaxies.

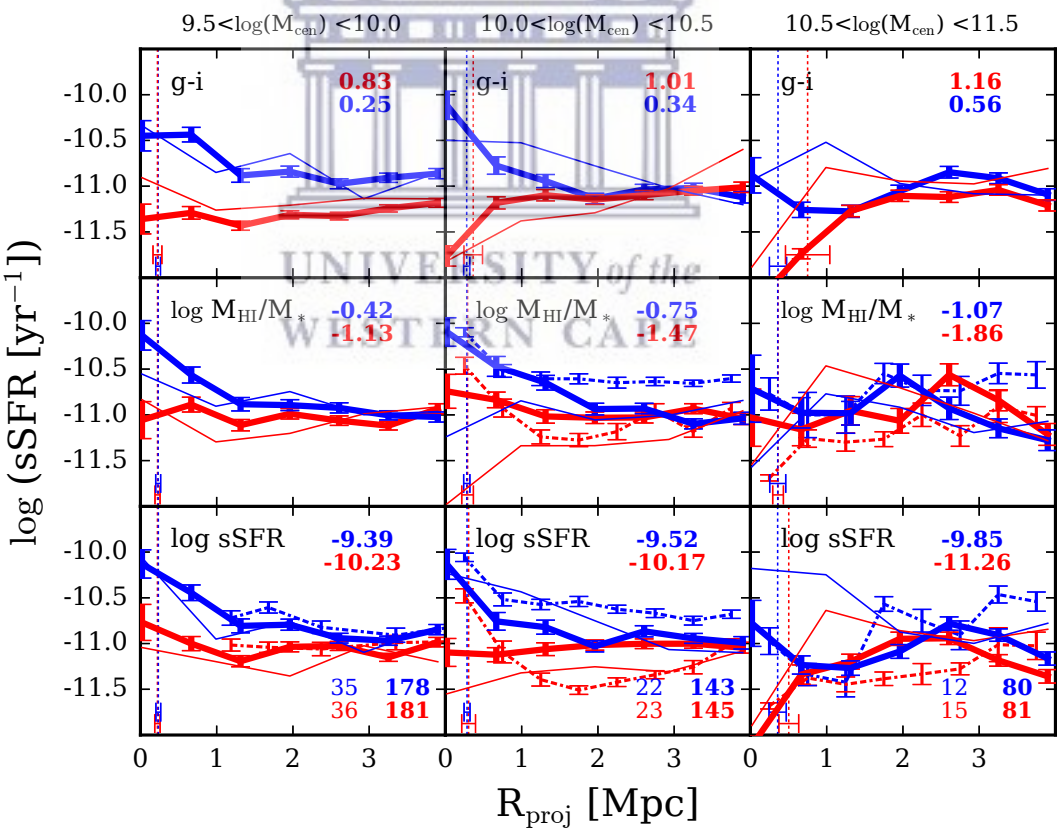


FIGURE 2.7: Similar to 2.6, except showing the sSFR of the neighbours. Dashed lines (colour coded) show SDSS observations from Kauffmann et al. (2013) and Kauffmann (2015) for reference, with the caveat that we have not mimicked their selection in detail.

In Figure 2.7, we show conformity in the sSFR's of neighbouring galaxies. The trends are quantitatively similar to that seen in the colour conformity case, except flipped in sign since higher colours correspond to lower sSFR. Again, conformity is seen out to ~ 2 Mpc in lower-mass galaxies, and is strongly radial dependent. Here, we can see that the sSFR conformity clearly increases at low separations for HI and sSFR, for low and intermediate mass galaxies.

Conformity has been measured in SDSS with respect to sSFR (Kauffmann et al., 2013) and HI (Kauffmann, 2015). These data are shown as the dashed lines in the lower two rows. The median values of the simulated galaxy properties are in reasonable agreement with the observed trends. At low and intermediate central masses, the amplitude of the separation between the top and bottom quartiles are in agreement with data at $R_{\rm proj} \lesssim 1$ Mpc, and also show an increase of neighbours' sSFR to small separations. For the largest masses, the upturn at small radii in neighbour colour, HI content, and sSFR is not as strong, and this again is in broad agreement with observations. We emphasise that our isolation criterion based on using only central galaxies in friends-of-friends haloes is not the same as that used in (Kauffmann et al., 2013). As we showed in Figure 2.2 there are only mild differences between these different isolation criteria and strongest only for the most massive centrals, and thus it is instructive to compare to the (Kauffmann et al., 2013) data even if the criteria are not identical. Meanwhile, our results provide predictions that are more comparable to recent group catalog-based analyses such as that of Tinker et al. (2018).

In detail, at intermediate masses the observed neighbours of lowest-quartile centrals become substantially higher (bluer) at small separations, while Mufasa does not predict this trend (though the agreement is good for highest-quartile centrals). One explanation for this discrepancy could be that mergers increase sSFR at low separations, but this enhancement is not properly reflected in the simulations, perhaps owing to resolution. At larger radii, the data continues to show strong conformity that is not seen in Mufasa, but these data are somewhat in doubt owing to interloper contamination as discussed in Sin et al. (2017); Tinker et al. (2018). Our results are broadly in agreement with the semi-analytic or abundance matching models presented in those works, for which conformity disappears at distances not far beyond the virial radius in more massive galaxy samples.

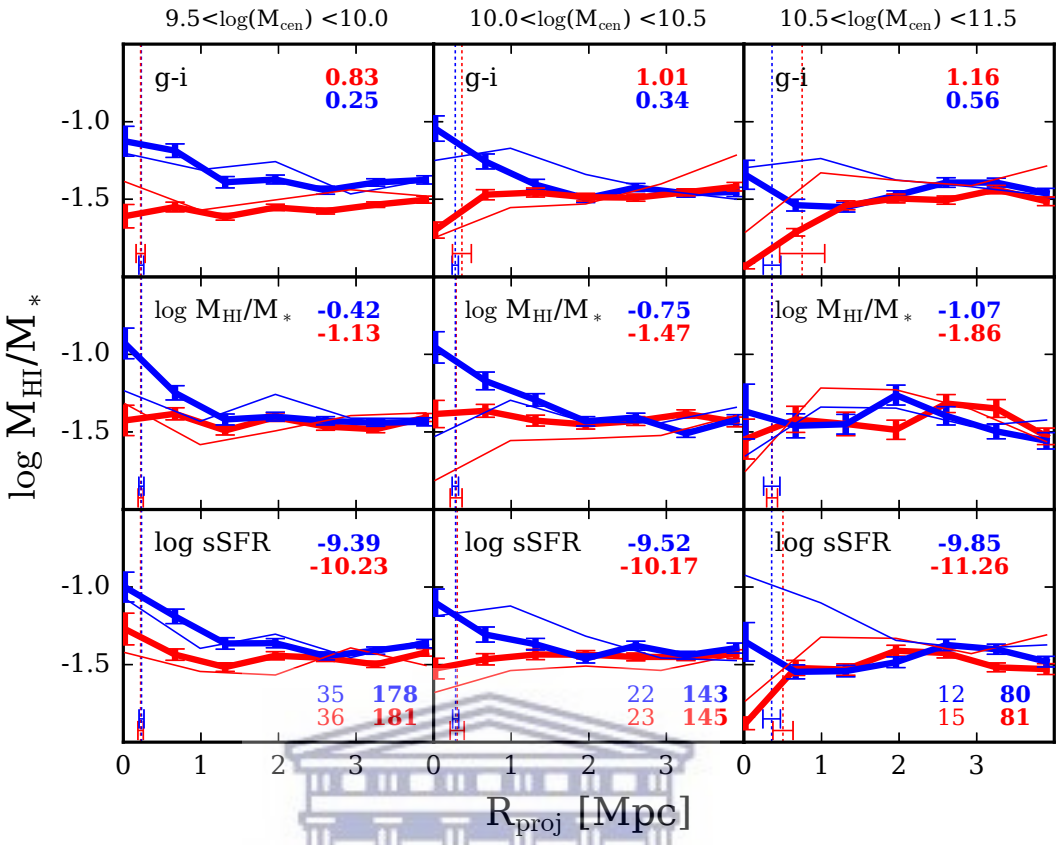


FIGURE 2.8: Similar to 2.6, except showing the HI richness of the neighbours.

For completeness, we consider in Figure 2.8 the conformity trend in the HI richness of neighbours, for central galaxies split by colour (top row), HI richness (middle), and sSFR (bottom). The trends generally mimic those in the other cases, in that conformity is only detectable out to $R_{\rm proj} \lesssim 2$ Mpc, and is only strong at $R_{\rm proj} \lesssim 1$ Mpc. The neighbours' HI richness, like colour and sSFR, becomes higher (bluer) towards small separations for low-mass centrals, but become lower (redder) for high-mass centrals.

Comparing the $50h^{-1}$Mpc (thick) and $25h^{-1}$Mpc (thin) line predictions, we see that in general the conformity signature is qualitatively similar at both resolutions, though they are not within each others' formal error bars. As shown in Davé et al. (2017b), the two volumes are not completely converged in terms of many of their stellar and gas properties. Ironically, the convergence in conformity strength appears to be best-converged for the smallest centrals; at higher masses, the trends are less well mimicked in the two volumes, though overall the trends are similar. On top of the rather small sample from m25n512, we

also speculate that the discrepancy might also be from the model behaving differently at different resolutions. We note that the error bars for the $25h^{-1}$Mpc results (not shown for clarity) are generally larger than for the $50h^{-1}$Mpc.

Overall, Mufasa displays fairly strong galaxy conformity within $\lesssim 1$ Mpc projected radius, and weak conformity in most cases out to ~ 2 Mpc, which disappears at high masses. The conformity signature is present and similar in all permutations of central vs. neighbour galaxy properties considered here, namely colour, HI richness, and sSFR. The trend to small radius at large central masses shows the impact of halo quenching, but for lower central masses we see bluer, more gas rich, and higher sSFR neighbours towards small radii. Comparing to observations of Kauffmann et al. (2013); Kauffmann (2015), the trends generally agree at projected radii less than about 1 Mpc. Mufasa does not produce as strong conformity at larger scales as inferred by those data, but at large separations these data may be impacted by interloper contamination. Observational comparisons are sensitive to details of selection effects etc. particularly at large radii, so we consider the outcome of conformity in Mufasa and the broad agreement at $R_{\rm proj} \lesssim 1$ Mpc to be encouraging, and leave a more detailed data comparison for future work.

2.5 The Nature of Conformity

In the previous section we showed that galactic conformity is present in Mufasa in specific star formation rate, HI richness, and colour of galaxies at similar levels, with conformity being stronger at small separations and extending farther out in lower-mass central galaxies. Physically, one might regard conformity as a reflection of quenching processes associated with hot massive haloes (e.g. Peng et al., 2012; Gabor & Davé, 2015). Mufasa assumes a quenching halo mass scale of $\sim 10^{12} M_{\odot}$, and hence the effects of conformity might be expected to be stronger in haloes above this mass scale, since the gas surrounding satellites has now been forcibly heated to the virial temperature. However, starvation and stripping processes can happen in lower-mass haloes through tidal interactions and harrassment. Hence it is an interesting question to quantify how conformity changes with halo mass, particularly across our quenching mass threshold.

To this end, in this section we subdivide our sample with respect to halo mass above and below our nominal quenching scale, as opposed to subdividing in stellar mass as in the previous section in order to more closely compare with data. Furthermore, we consider conformity in a wider range of properties beyond only what has been observed, to a more exhaustive set of galaxy properties. This will help us identify which properties display the strongest conformity, and hence in some sense drive galactic conformity.

2.5.1 Conformity in non-quenched haloes

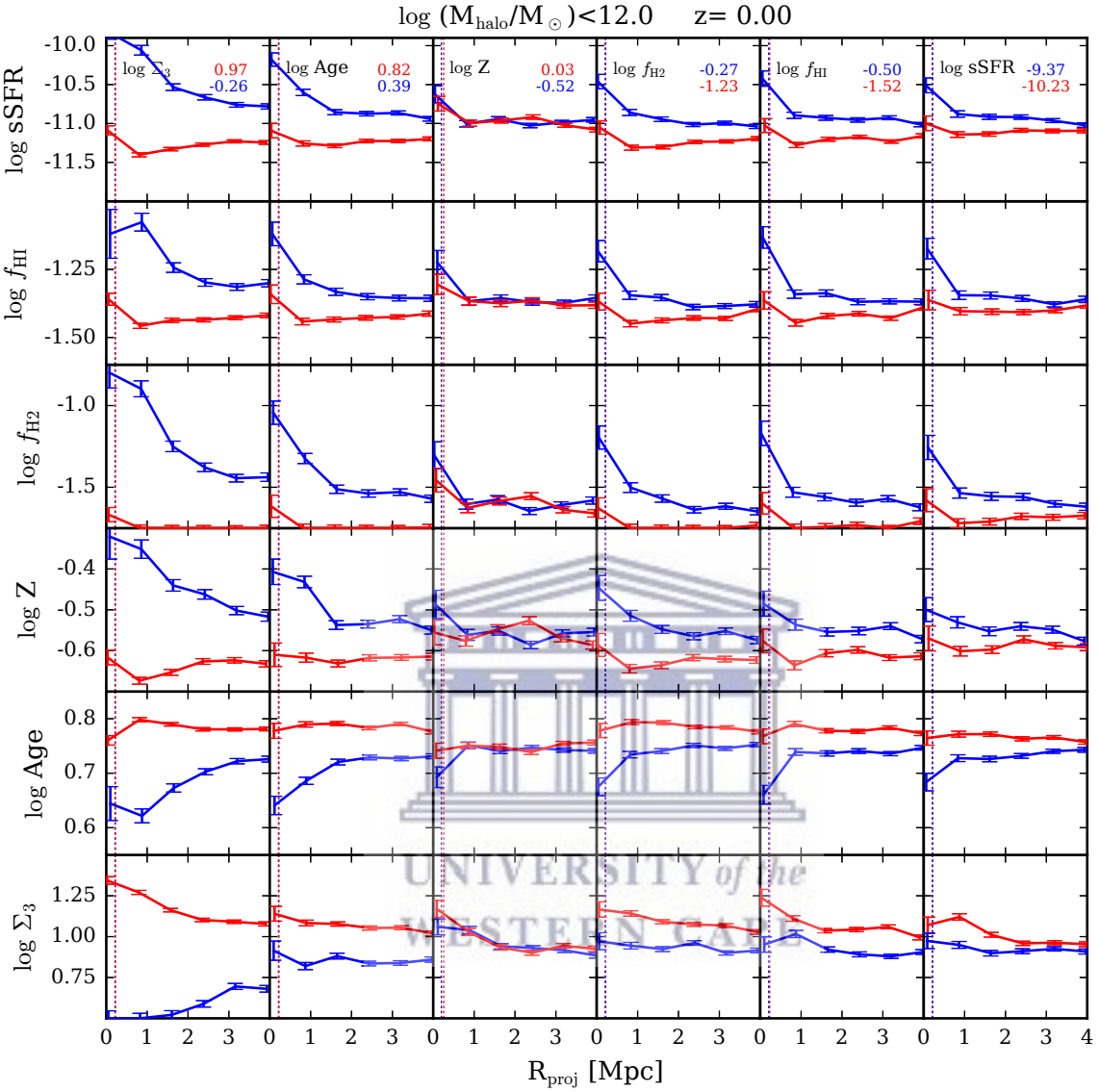


FIGURE 2.9: Neighbour galaxy median properties versus projected distance to their central galaxies living in haloes with mass $M_{halo,cen} < 10^{12} M_{\odot}$, binned by central galaxies in highest and lowest quartile for each given property. The specific property shown in each column is indicated in the upper left of the top row, while the median quantities for the central galaxies are indicated on the upper right (color coded with the lines). The dashed vertical lines are the median values of the R_{vir} of the central galaxies in the highest and lowest quartiles.

Figure 2.9 shows, from top to bottom panels, the sSFR (yr^{-1}), f_{HI} (HI richness), f_{H_2} (molecular hydrogen fraction), Z (gas phase metallicity), Age (median value of the stellar ages, in Gyr) and Σ_3 (third projected nearest neighbour density) of central galaxies in haloes of $\log(M_{halo}/M_{\odot}) < 12$, ordered by each of those

properties (columns). The trends for sSFR and $f_{\rm HI}$ were shown in Figure 2.7 and Figure 2.8 binned by central stellar mass, but here we show them binned by halo mass. The diagonal panels (from lower left to upper right) show the (auto-)conformity in each quantity, while the off-diagonals show the cross-conformity with the rows representing the neighbour properties and the column representing the central properties.

Before we delve into a detailed analysis, it's important to keep in mind the nature of galaxies in this halo mass bin. Note that the median sSFR of the reddest quartile of central galaxies is log (sSFR/Gyr^{-1})=−1.23, which is not a quenched galaxy by our (or almost any reasonable) definition. This is not surprising, since in Mufasa, we do not apply our quenching prescription in this range of halo masses. Nonetheless, it is worth bearing in mind that here we are mostly examining trends among star-forming galaxies, subdivided into redder vs. bluer, more gas-rich versus less, etc. Since there exist trends with galaxy mass in these quantities (e.g. Davé et al., 2017b), binning in these quantities implicitly includes some trend with galaxy mass. Ideally, one would remove this effect by sub-binning in narrow bins of galaxy (or halo) mass, but given our statistics, this is not feasible.

Let us consider the auto-conformities first, along the diagonal panels. For all quantities except Z, the conformity is well-defined out to remarkably large scales, exceeding the 4 Mpc (projected) limit that we consider here, with a strength that diminishes relatively slowly. We also saw this behavior in some but not all cases around the low-mass centrals shown in Figures 2.6-2.8. This implies that neighbours know about their large-scale environment out to quite large distances. This likely does not reflect halo (AGN) quenching, ram pressure stripping, or other processes traditionally associated with massive haloes, since it is not confined near the haloes themselves, nor are the selected massive haloes. Instead, it represents conformity driven by the growth of large-scale structure by the tendency for massive galaxies (with e.g. redder colours and lower gas content) to live around other massive galaxies.

The most striking trend is seen in the first column of Figure 2.9. Here, we see that the difference between the satellite properties is always greatest when the centrals are subdivided by our environment measure Σ_3. This is true not only for the Σ_3 auto-conformity but for every cross-conformity measure as well. From this, we conclude that the most important driver of conformity is the

nearby galaxy density, and clearly identifies galactic conformity in low-mass haloes as environmentally-driven. This is perhaps unsurprising, since conformity is inherently itself a tracer, *i.e.* if galaxy properties vary smoothly with environment, then having many galaxies of a similar type close to each other will give rise to strong signals in both Σ_3 and galaxy conformity.

The next strongest level of difference between satellite quartiles is provided by separating centrals in stellar age (second column from left). This is qualitatively consistent with the findings of Hearin et al. (2016), who argued that assembly bias is crucial for driving two-halo conformity. Although we have restricted this sample to $M_{\rm halo} < 10^{12}$ $M_\odot$, it is possible that the environmental dependence of the halo mass function within this mass range may drive conformity, with the age aspect being a consequence rather than a driver (Zu & Mandelbaum, 2018); this would be consistent with our results that the environment shows the strongest trend in conformity. To test this, we would need to construct an age-matched sample within each environment, but unfortunately given our limited simulation volume, the results are too noisy to extract clear trends.

We next examine the gas and star formation rate conformity properties. Interestingly, the conformity signature is at least as strong when centrals are subdivided by f_{H2}, as compared to either sSFR or f_{HI}. This is notable since current observations have mainly probed the latter two (since they are the most observationally accessible given current data), but Mufasa predicts that these conformity signals are actually similar if not weaker compared to that seen in molecular gas fractions.

Finally, the metallicity has a very minor conformity signal, only at $R_{\rm proj} < 1$ Mpc. The most curious aspect is that qualitative trend in the metallicity of neighbours relative to the sSFR. Looking at the top row of the metallicity column, then at low $R_{\rm proj}$, one sees that low-Z centrals tend to have neighbours with slightly higher sSFR. This is consistent with the fundamental metallicity relation (e.g. Mannucci et al., 2010; Lara-López et al., 2010), where at a given mass, galaxies with high sSFR have low-Z and vice versa; this trend is reproduced in Mufasa (Davé et al., 2017b), and apparently extends to satellites as well. In contrast, if you look at the sSFR (6th) column, fourth panel down, it curiously shows that centrals with high sSFR are surrounded by neighbours with *high* metallicity – in other words, the trend is flipped, and metallicity actually shows anti-conformity! The differences are subtle but robust, which we will

quantify later on. This may arise as an age effect – one can see in the Age (2nd) column, fourth panel down, that galaxies with young ages tend to be much more metal-rich, because they have had longer time to form stars and hence enrich themselves.

In general, we note that the overall amplitude of conformity looks most similar in the columns of our plot, rather than in the rows. This means that the conformity is generally most driven by the *central* galaxy property being examined, but relatively independent of the neighbour property being examined. In contrast, the trends with radius are most similar when examining a particular neighbour galaxy property.

The physical interpretation of one-halo conformity in non-quenched haloes likely traces back to halo assembly bias. This is because, in this regime, there is no explicit physics that will turn nearby galaxies red, as there is in the quenched haloes case where there is strong local quenching feedback. Hence the similarity of galaxy colours likely arises owing to the tendency for haloes living in denser regions to have formed earlier and be surrounded by more gravitationally shock-heated gas, which results in an overall reduction of the accretion rates onto all galaxies living in such extended structures. This is essentially a form of assembly bias, in which galaxies at a given halo mass experience different growth histories depending on their large-scale assembly history. Such an interpretation is consistent with that of Tinker et al. (2018) from SDSS and Bray et al. (2016) from examining the origin of conformity in the Illustris simulation.

In summary, $M_{\rm halo} < 10^{12}$ $M_\odot$ haloes show noticeable conformity across almost all galaxy properties, with the strongest absolute differentiations among neighbours occurring in environment (Σ_3). Specific SFR, $f_{\rm HI}$ and Age have actually rather weak conformity signal relative to that, and somehow comparable to molecular fraction $f_{\rm H2}$. There is little conformity in galaxy metallicities, and in fact shows opposite trends depending on whether one considers cross-conformity versus the central's metallicity or the neighbour's metallicity. Some of these trends may arise owing to mass trends among central galaxies when binned into upper and lower quartiles. We will discuss conformity more quantitatively in §2.5.

2.5.2 Conformity in quenched haloes

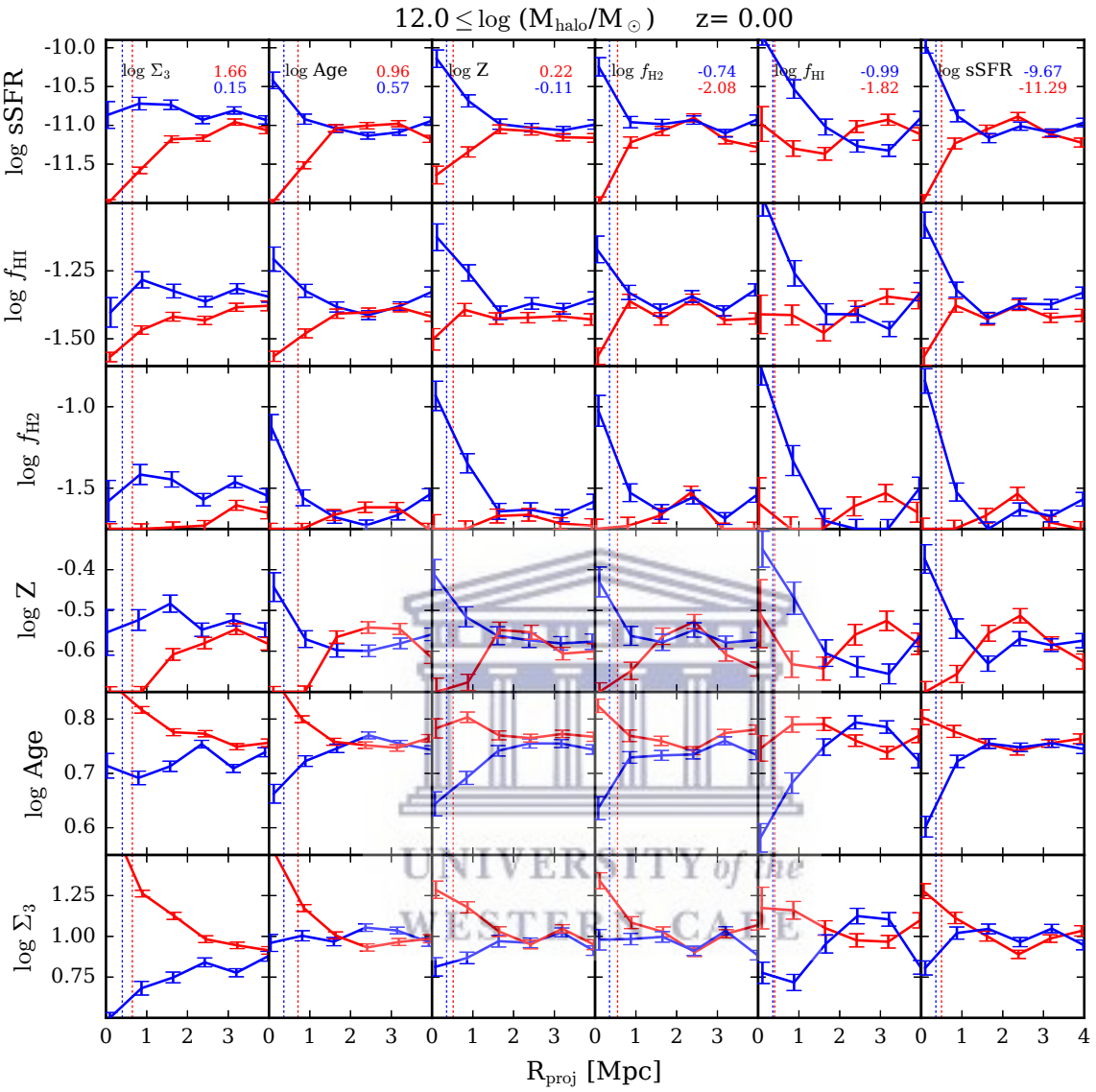


FIGURE 2.10: Similar to Figure 2.9 except for centrals living in $M_{halo,cen} \geq 10^{12}\ M_{\odot}$.

We now conduct a similar investigation, but for centrals within quenched haloes having $M_{halo} \geq 10^{12} M_{\odot}$. Figure 2.10 shows the analogous plot to Figure 2.9 for these quenched haloes. Note that the typical sSFR of upper-quartile centrals is only a factor of 2 lower than in the low-mass halo case, but for lower-quartile centrals the sSFR is more than an order of magnitude lower, showing the onset of a considerable quenched population in massive haloes. Similarly, the magnitude of the difference is substantially larger between the highest and lowest quartile in gas content as well.

The trends at high halo masses are clearly different than for the lower mass haloes examined in the previous section. Most strikingly, conformity in just about every property is now restricted only to $R_{proj} \lesssim\sim 1-1.5$ Mpc, with the exception of Σ_3. Within that range, the radial trend is much stronger than in the low-mass halo case, as quenching clearly plays a role in impacting the satellite galaxy population. The one-halo conformity is obviously well stronger than the two-halo conformity, showing that satellites in particular are very strongly impacted by environmental-specific processes owing to quenching.

Looking at individual properties, we see once again that environment (Σ_3) shows the strongest absolute levels of conformity, and its auto-conformity extends out to ~ 4 Mpc. The cross-conformities tend to nearly vanish beyond $\gtrsim 2$ Mpc. But in contrast to the low-mass halo case, the strengths and radial trends of the conformities in other properties are remarkably similar, and even the metallicity shows strong conformity.

These trends show qualitatively the behaviour one would expect, with centrals that are high-sSFR, gas-rich, young, and low-density having like neighbours. The exception again is metallicity, where low-Z centrals tend to have high-Z satellites and neighbours, and vice versa. The anti-conformity of metallicity is an interesting testable prediction.

Recall that in §2.3.2 we highlighted an issue with these simulations in that they overproduce the number of low-mass quenched satellites. This is expected to increase the 1-halo conformity term for the reddest (and analogously least star-forming and gas-rich) quartile, since quenched central galaxies are also red. The magnitude of this effect is difficult to quantify without a new feedback model that is able to mitigate this discrepancy. We are currently working on such a model, but do not have final results at this time. For now, we note that the quenched halo predictions may change depending on the new input physics required to fix this discrepancy.

In summary, both low-mass and high-mass (quenched) haloes show conformity in virtually all quantities, but the trends are qualitatively different. High-mass haloes show a relatively confined (spatially) extent of conformity, with a very strong radial trend, and similar conformity strength in all properties except for Σ_3 which is the strongest. In the next section we discuss our approach to

quantifying these trends, in order to more carefully inter-compare and study their evolution with redshift.

2.5.3 Quantifying conformity

Previous works have generally focused on quantifying conformity by measuring its detectable radial extent, within a given tracer. The next logical step would be to quantify the conformity strength in a manner such that we can intercompare the strengths amoung various tracers. In the previous sections, we have focused on comparing the *absolute* strength of conformity within a given neighbour property, binned by central property. However, it is not obvious how to compare this strength between different properties, since effectively this absolute conformity signal (say, the difference between the red and blue curves) has units associated with that property. Thus we need a new a measure that also enables cross-comparisons between various neighbour properties, in order to more robustly determine which neighbour property shows the strongest conformity.

The fundamental idea of conformity is to quantify how well the neighbours follow the trends of their central. In this sense, a good measure of conformity to inter-compare properties would be to measure the difference between the neighbour properties, relative to how much difference there is in the central galaxy properties. This may be regarded as a *relative* conformity, as opposed to the absolute conformity that we have investigated in the previous sections. As such, we define (relative) conformity strength $\mathcal{S}$ as follows:

$$\mathcal{S} = \frac{1}{N_{\rm bin}} \sum_{i=0}^{N_{\rm bin}-1} \frac{Q1_i - Q4_i}{Q1_{\rm cen} - Q4_{\rm cen}} \tag{2.5}$$

where $Q1_{\rm cen}$ ($Q4_{\rm cen}$) is the property of the central galaxies in the 1^{st} (4^{th}) quartile, $Q1_i$ ($Q4_i$) is the (logarithm of the) property of the galaxies neighbouring the 1^{st} (4^{th}) quartile in the radial bin i. In this way, we normalize the difference in neighbour properties by the difference in the central property. Generally, we will consider $\mathcal{S}(R < 2$ Mpc), *i.e.* summing all bins that are within 2 Mpc of the central galaxy. We will also consider $\mathcal{S}(R)$, *i.e.* taking a single bin at each radius.

Broadly, $\mathcal{S} = 1$ means that the neighbours show exactly the same difference between the top and bottom quartiles as the centrals; this is full conformity. $\mathcal{S} = 0$ means no conformity. We will characterise weak conformity as $\mathcal{S} < 0.5$, and strong above this. It is possible to get $\mathcal{S} > 1$ which indicates very strong conformity, with neighbours showing a greater difference in a property than even their centrals, or even $\mathcal{S} < 0$ which is anti-conformity. Hence this quantity provides an intuitive measurement of conformity that is independent of the given property. We note that this is not a new definition of conformity. We use similar definition of conformity from Kauffmann et al. (2013) but only extend it to be presentable in one number.

2.5.4 Conformity as a function of mass and radius

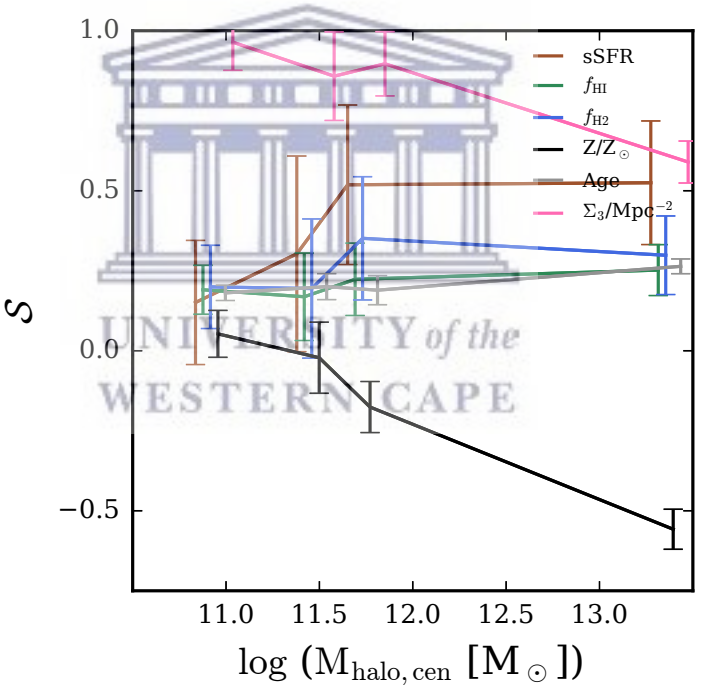


FIGURE 2.11: Halo mass dependence of the auto-conformity strength in MUFASA at redshift $z = 0$. Halos have been binned so that each bin contains approximately the same number. 1σ error bars are derived by jackknife resampling. Environment shows the strongest conformity at low masses, but gets weaker with halo mass. sSFR shows the strongest conformity among the remaining quantities, and increases with halo mass from $M_{\rm halo} \sim 10^{12} M_\odot$. Metallicity displays anti-conformity.

In Figure 2.11, we shows $\mathcal{S}(R < 2$ Mpc) at $z = 0$ for central galaxies binned by their halo masses: we chose each bins to contain the same number of central galaxies. Σ_3 signal shows the strongest conformity signal, quantifying the trend in the previous sections that conformity is most strongly related to environment. The Σ_3 conformity is however anti-correlated with halo mass, so environment is the primary driver at low and intermediate halo masses, but at the highest halo masses it is comparable to the other quantities. While we discussed previously that this perhaps not surprising since conformity corresponds to the similarity in properties of nearby galaxies, while Σ_3 is itself a measure of how many nearby galaxies there are, it is not a trivial result, since it indicates that galaxies that lie particularly close to each other in dense regions have similar properties. Moreover, the trend with mass is interesting, and is driven by the difference between the central galaxy properties, *i.e.* the denominator of $\mathcal{S}$, increasing. Hence, effectively, large-scale structure is the primary driver of conformity when quenching processes are not present, but once they are, then it is no longer so dominant.

Besides Σ_3, sSFR clearly shows the strongest conformity at most halo masses. Recall that in the previous section we found that Age and $f_{\rm H2}$ were stronger in an absolute sense. However, when computing $\mathcal{S}$, we normalize this to the difference in the central galaxy property, and in this case we discover that sSFR conformity is stronger than that of Age, $f_{\rm H2}$, or $f_{\rm HI}$. Thus it appears that, beyond the obvious dependence of conformity on environment, sSFR shows the strongest levels of similarity between the neighbours and the central galaxies. Though we do not show it here, we expect galaxy colour would show a very similar trend to sSFR, as we found in §2.4.

As discussed previously, metallicity shows anti-conformity at most halo masses, increasingly so to larger halo masses. Recall this is the SFR-weighted gas-phase metallicity, not the stellar metallicity; one might expect a different trend for stellar metallicity, as central galaxies within large haloes tend to have metal-rich stars but their star-forming gas may be dominated by recent infall that is relatively metal-poor.

Observational analyses from, e.g., Tinker et al. (2018); Sin et al. (2017) suggest that galactic conformity is only prevalent in the relatively nearby environment, within $\lesssim 1$ Mpc. It can be seen in Figure 2.10 that, particularly in massive haloes, that were generally the targets of these studies, the conformity signal

predicted in Mufasa drops quickly with radius. However, this is less true in low-mass (non-quenched) haloes (Figure 2.9), which shows overall weaker conformity but much less radial dependence.

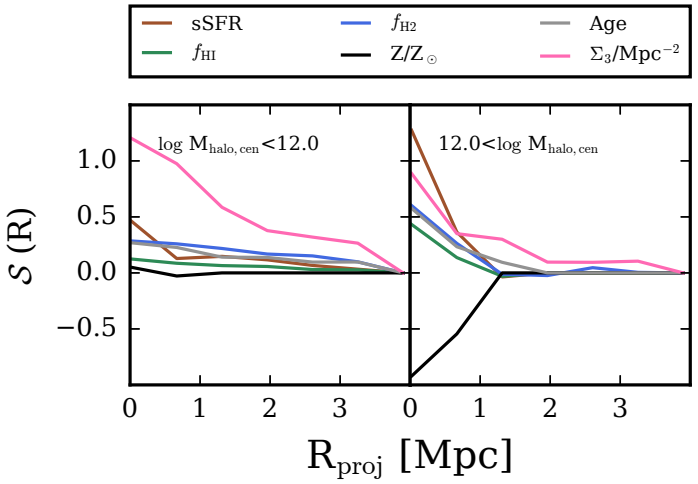


Figure 2.12: Radial dependence of conformity strength $\mathcal{S}(R)$. Here $\mathcal{S}$ is computed within each radial bin. Low-mass haloes show weak conformity with little radial trend, except for Σ_3. Massive haloes show stronger conformity for neighbours close to the central, but little conformity at $R_{proj} > 1$ Mpc.

To quantify this, we show in Figure 2.12 the conformity signal as a function of projected radius R_{proj}, for non-quenched (left panel) and quenched (right) haloes. The qualitative trends evident in Figures 2.9 and 2.10 are evident here. The conformity is strongest in the environmental measure Σ_3 at all radii, except close in for massive haloes. Conformity is relatively weak with modest radial dependence in the low-mass haloes, while it is strong at $R < 1$ Mpc in the massive haloes but mostly nonexistent beyond this (except in Σ_3). In fact, in the innermost bin, the sSFR shows stronger conformity that even Σ_3, exceeding unity. It is thus perhaps not surprising that conformity was first noticed for colours of satellites around sizeable galaxies (Weinmann et al., 2006a) – Mufasa predicts the conformity signal is quite strong under these conditions.

In summary, conformity is strongest in environment, confirming the environmental nature of galaxy conformity. Beyond this, the next strongest conformity is in sSFR, which is mildly stronger than gas fraction and Age. Metallicity presents an odd anti-conformity particularly in quenched haloes, which may be a signature of recent accretion of metal-poor star-forming gas into massive

central galaxies. Massive haloes show a strong conformity signal in most quantities only at $\lesssim 1$ Mpc, while less massive haloes show a weaker conformity with a weaker gradient.

2.5.5 One-halo vs. two-halo conformity

A particularly interesting quantity to examine is the relative contribution of one-halo versus two-halo conformity to the total conformity signal, as there has been substantial debate in the literature regarding the strength, origin, and even existence of two-halo conformity. Here we explicitly examine this for each halo, computing only the conformity associated with galaxies that are satellites within the haloes of the chosen central, versus those outside the halo up to 2 Mpc. Note that this is not a strict radial cut, since FOF haloes typically have non-spherical shapes.

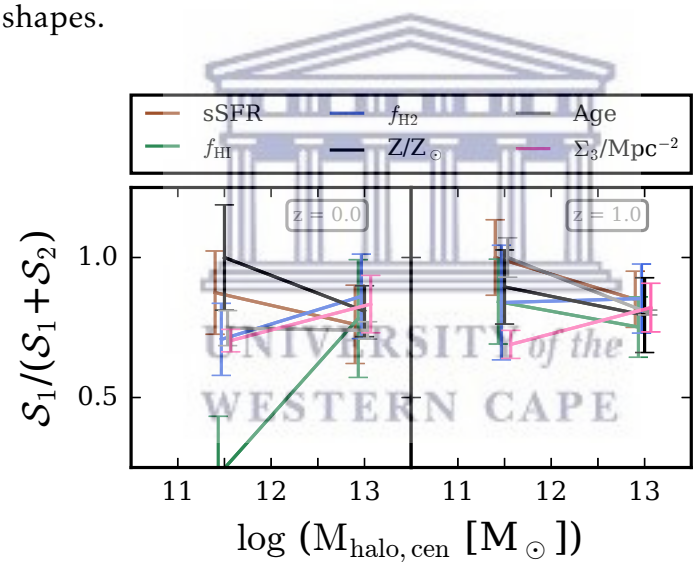


FIGURE 2.13: Halo mass evolution of the relative contribution of the one-halo conformity. $\mathcal{S}_1$ ($\mathcal{S}_2$) is one(two)-halo conformity. One-halo conformity generally dominates the signal for all quantities except HI in low-mass haloes.

Figure 2.13 shows the fractional contribution of one-halo conformity ($\mathcal{S}_1$) versus two-halo ($\mathcal{S}_2$), specifically $\mathcal{S}_1/((\mathcal{S}_1+\mathcal{S}_2)$, as a function of halo mass. Owing to small number statistics for the one-halo term particularly at low halo masses, we separate our central galaxy sample into only two bins of halo mass, $M_{halo} < 10^{12}$ $M_\odot$ and $M_{halo} \geq 10^{12}$ $M_\odot$. The left panel shows the $z = 0$ results, while the right shows $z = 1$.

It is clear that one-halo conformity dominates the overall strength for almost all quantities for both low and high mass haloes, at both $z = 0$ and $z = 1$. The only deviant case is the HI fraction in low mass haloes, which likely owes to the physical effect that satellite galaxies around star-forming centrals can have their HI stripped relatively easily, so that such centrals actually end up having different HI fraction relative to their satellites but more similar to distant galaxies. Typically, one-halo conformity is $\sim 3-6\times$ stronger than two-halo conformity, according to our $\mathcal{S}$ measure. We note that while the strength of one-halo conformity dominates in most circumstances, the strength of two-halo conformity at small radii depends on halo mass, and this may be partly the cause of divergent results in the literature regarding two-halo conformity.

2.5.6 Evolution of conformity

Conformity has been observed to exist out to $z \gtrsim 1$ and beyond, although the strength of the evolution is difficult to quantify at this time. In this section we examine the evolution of conformity strength $\mathcal{S}$ predicted by MUFASA.

There are two ways to track the conformity back in time for a galaxy population. One way is to find the conformity strength for the most massive progenitors of the $z = 0$ central galaxies; we will call this $\mathcal{S}_{\rm track}$, since we are tracking individual galaxies. Another way is to consider the centrals in the same halo mass bins as at $z = 0$, separated at $M_{\rm halo} = 10^{12} M_\odot$; we will call this $\mathcal{S}_{\rm ev}$, which shows the overall evolution of conformity. In both cases we take the absolute value $|\mathcal{S}(< 2\ {\rm Mpc})|$ for plotting purposes, and note that only the metallicity shows negative values (anti-conformity).

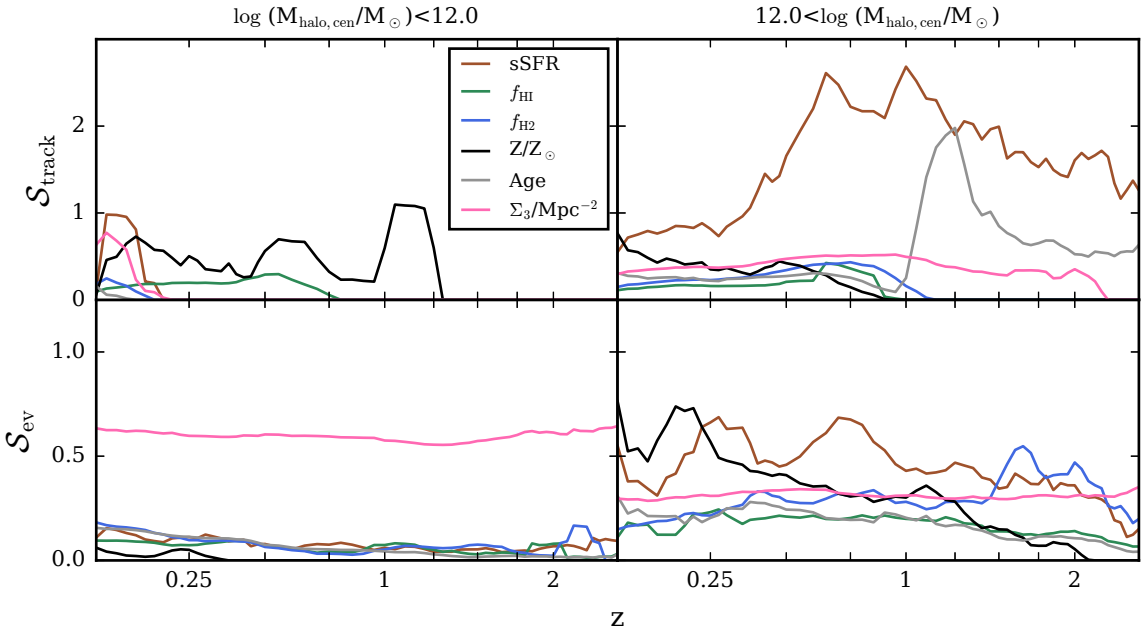


FIGURE 2.14: Evolution of the strength of galactic conformity. The halo mass range of the central galaxies are given on top of the columns. Upper panels show the evolution the conformity strengths tracking back the $z = 0$ progenitors in time. Lower panels show the evolution of conformithy within the listed halo mass bins at each redshift independently. Conformity is a late-time emergent phenomenon more associated with high-mass haloes, suggesting that it is dependent on quenching physics.

Figure 2.14 shows the evolution of galactic auto-conformity strength, (*i.e.* quantifying the diagonal panels of Figures 2.9 and 2.10), for central galaxies binned into two halo mass ranges: $\log(M_{\rm halo}/M_\odot) < 12$ and $12 < \log(M_{\rm halo}/M_\odot)$. For clarity of the plot, we set the strength to 0 when $\mathcal{S}$ crosses 0 and stay on the other side of the 0-line for two successive snapshots. In addition, we smoothed the curves to their moving averages over $\Delta z = 0.1$ for $z < 0.5$, and 0.2 elsewhere, to average over the fluctuations among individual redshift snapshots.

The left panels show $\mathcal{S}_{\rm track}$ (top) and $\mathcal{S}_{\rm ev}$ (bottom) for low mass (unquenched) haloes. For these, conformities are generally weak at all redshifts. Tracking our specific central galaxies back from $z = 0$, $f_{\rm HI}$ and Z emerge earlier than the others at $z \sim 1$, while the others only appear at $z < 0.25$. Because these are small haloes, we are limited in how far back we can track these galaxies before they lack resolution to follow. However, it is evident that we can track sufficient numbers till at least $z \sim 1$, so it is interesting that the low-mass conformity we see at $z = 0$ in many quantities is actually quite a late-time phenomena for these particular galaxies. This suggests that conformity in e.g. Σ_3 or sSFR requires

halo masses that approach $\sim 10^{12} M_{\odot}$, and at significantly smaller halo masses there is no conformity except the metallicity anti-conformity. With larger statistical samples from upcoming larger-volume simulations, we will be able to test this idea more finely.

Examining $\mathcal{S}_{\rm ev}$ for low-mass haloes, we see that conformity is very uniform and weak at all redshifts, with the exception of Σ_3 as noted before. Again, this suggests that there is mass threshold for the emergence of conformity that is close to $\sim 10^{12} M_{\odot}$. There is a hint that conformity strength increases with time for a fixed halo mass sample, but given the small statistics at $z \gtrsim 2$ it is difficult to draw firm conclusions.

Turning to the high mass haloes (right panels), we see much stronger levels of conformity in some quantities. In particular, in $\mathcal{S}_{\rm track}$ (top right) sSFR conformity is extremely strong out to $z \gtrsim 2$. Hence for massive haloes, even tracking them back in time shows that classic (colour-based) conformity emerges quite early on. We note that this trend is driven by a few massive haloes that have a large number of satellites, since conformity in massive haloes is generally restricted to relatively small radii and hence are dominated in statistics by satellites. This also causes some odd behaviour such as the Age-$\mathcal{S}_{\rm track}$ which spikes up around $z \sim 1$; we hesitate to over-interpret this behavior without more statistics. It is also the case, as at lower masses, that gas conformity is a relatively late-time phenomenon, appearing only at $z \lesssim 1$.

For $\mathcal{S}_{\rm ev}$ in high-mass haloes, again we see that the conformity strength is fairly constant with redshift. The metallicity (anti-)conformity shows the most significant increase with time. This is consistent with the idea that conformity in most quantities is primarily driven by environment (which correlates strongly with halo mass), except for metallicity where the anti-conformity is driven by a different effect namely the late infall of low-metallicity gas into massive galaxies.

Another way to quantify conformity is to ask, how far from the central galaxy should we expect conformity to be evident? To quantify this, we measure the distance from the central galaxies where the two-halo conformity still exists (R_2), relative to the mean $R_{\rm vir}$ of the central galaxy subsample. R_2 is measured as the largest projected distance from the central galaxies where the highest and lowest quartile samples are first within 1σ of each other (e.g. where the

blue and red lines are within each other's error bars in Figure 2.9). We note that R_2 is thus mildly sensitive to the level of the error bars, and hence the sample size.

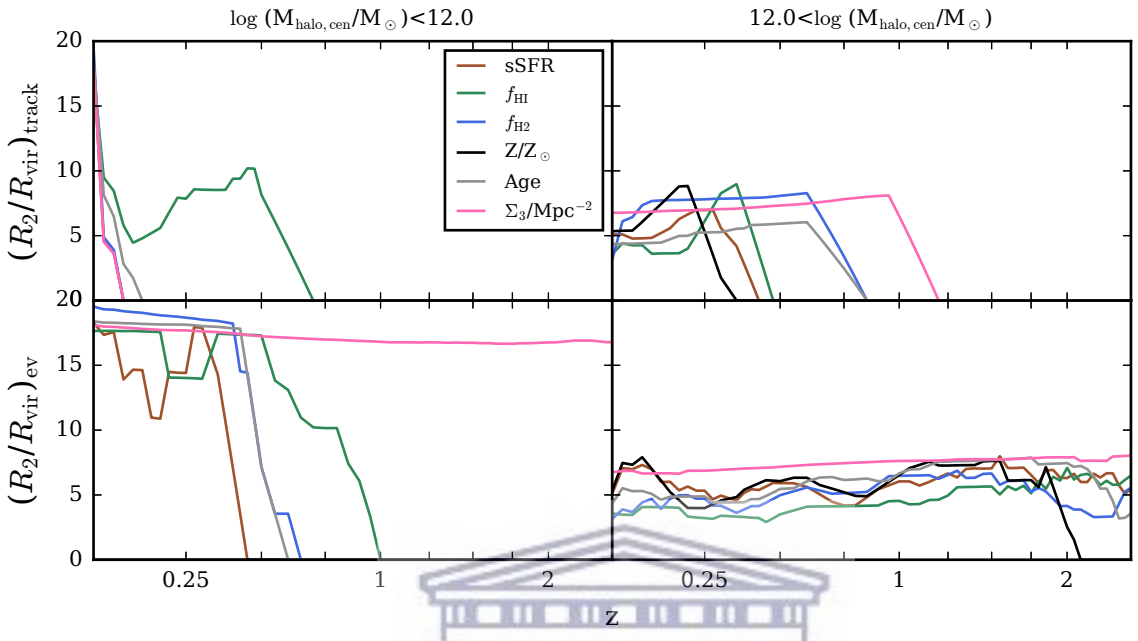


FIGURE 2.15: Extent of the two-halo conformity. R_2 is the distance from the primary galaxies to the point where the two-halo conformity still exists, *i.e.* blue and red lines in Figure 2.9 are within each other's error bars. At low masses, conformity emerges at late epochs and quickly to large numbers of virial radii. At high masses, once it appears, conformity is evident out to a fairly constant $\sim 5R_{vir}$.

In Figure 2.15 we show R_2 as a function of redshift for low (left panels) and high (right) halo mass bins and tracked either using progenitors (top panels) or at a fixed halo mass bin (bottom), analogous to Figure 2.14.

At $z = 0$ for small haloes, we see as in Figure 2.9 that the conformity extends out to many virial radii; we truncate this at 2 Mpc, corresponding to ≈ 18 median virial radii in that halo mass bin. However, tracking these galaxies back in time shows that the conformity radius disappears extremely rapidly, with the exception of HI fraction for which it extends to $z \sim 2$. Even using a fixed halo mass bin, R_2 starts dropping beyond $z \gtrsim 0.25$, and disappear by $z \sim 1$.

For massive haloes, R_2 extends typically out to $\sim 5R_{vir}$, as long as it is present. Tracking galaxies back, we see that when conformity emerges, it emerges quickly to this radius. Of course, this exact radius as well as the exact redshift depend on the sample size, but the trend of rapid emergence is robust. When looking

within a fixed halo mass bin, we see the radius to which noticeable conformity extends is remarkably constant out to $z \sim 2$ in terms of typical virial radii.

In summary, conformity in low-mass haloes appears to be a late-time effect, likely arising from haloes as they approach $\sim 10^{12} M_{\odot}$. Tracking massive haloes, conformity is evident and strong in sSFR at all redshifts, but in the gas properties and stellar age it is relatively weak and emerges at $z \lesssim 1$. When looking at $> 10^{12} M_{\odot}$ haloes at each redshift, conformity is fairly constant in terms of both strength as well as number of radii to which it extends. Hence we predict that conformity should be evident in reasonably massive galaxies at approximately the same level at higher redshifts as at $z = 0$. These trends are qualitatively consistent with observations showing that conformity continues to be evident out to $z \sim 2$.

2.6 Summary & Conclusion

We have examined galaxy conformity as an emergent property of a cosmological hydrodynamic simulation of galaxy formation. In particular, we employ the Mufasa simulation that has been shown to reasonably reproduce a range of observed galaxy physical properties over much of cosmic time (Davé et al., 2016, 2017b). The approach for quenching massive galaxies in this simulation is purely heuristic, utilising a slowly-evolving threshold halo mass above which diffuse halo gas is prevented from cooling, which well reproduces the observed red/blue bimodality in the galaxy population (Davé et al., 2017a) but does not invoke a specific physical model driving the quenching. Galaxy conformity provides a relatively unexplored statistic with which to quantify how such quenching impacts the properties of galaxies in various environments. To this end, we examine conformity in galaxy properties that have been previously looked at in the literature, namely colour, HI fraction and specific SFR, along with a range of other galaxy properties such as molecular content, stellar age, gas-phase metallicity, and environmental density.

Our main findings are as follows:

- Mufasa yield approximately the observed fraction of total satellite galaxies as a function of central galaxy mass. However, it overproduces the

low-mass quenched satellites, most noticeably around high mass central galaxies when the quenched satellite dominate the count (Figure 2.4).

- Satellite galaxy colours are generally bimodal for all halo masses and radii, with quenched satellites at small radii and star-forming ones farther out. The transition radius between these varies with central mass: $< 10^{12} M_{\odot}$ haloes have quenched satellites dominating only within the inner 10-15% of $R_{\rm vir}$, while $> 10^{13} M_{\odot}$ haloes have star-forming satellites primarily in the outer 10-15% (Figure 2.5).

- Very broadly, galaxy conformity is evident in Mufasa at all stellar masses examined ($M* > 10^{9.25} M_{\odot}$), in essentially all quantities examined. It typically extends significantly beyond the virial radius, though the strength diminishes with radius (Figures 2.6,2.7, 2.8,).

- Focusing on conformity in $g-i$ colour, specific star formation rate (sSFR) and HI , we find that conformity at low (central) galaxy mass is relatively weak, compared to high mass galaxy, but extends to quite large radii, while conformity at high galaxy mass declines more quickly with radius. The cross-conformity among these three quantities show similar trends to the (auto-)conformity (Figures 2.6, 2.7, 2.8,).

- We subdivide our galaxies into "unquenched" ($M_h < 10^{12} M_{\odot}$) and "quenched" ($M_h > 10^{12} M_{\odot}$) haloes, which provides a more direct view of the impact of quenching on neighbouring galaxies. Low-mass haloes show relatively weak conformity compared to high-mass haloes, extending to large radius, likely arising from assembly bias by which galaxies residing in more dense environments experience earlier growth and more suppression of accretion today. Meanwhile, high-mass haloes show conformity rapidly declining with radius and typically disappearing at projected radii above ~ 1 Mpc. This strong qualitative difference demonstrates that the presence of a sustained hot halo in our model has a major impact in driving strong galaxy conformity in our simulation (Figures 2.9, 2.10).

- Qualitatively, the strength of the conformity or cross-conformity signal is most directly correlated with the central galaxy property being examined, while the radial trend of the (cross-)conformity signal is dependent on the neighbour property. Environment (as measured by the density to the third nearest projected neighbour) and stellar age appear to have the

largest absolute (cross-)conformity signals, while metallicity shows a very small signal. The conformity strength in sSFR, $f_{\rm HI}$, and $f_{\rm H2}$ are comparable (Figures 2.9, 2.10).

- We introduce a measure $\mathcal{S}(R)$ (where R is the projected radius) to quantify the conformity strength, defined as the difference in the neighbour galaxies' properties in the first and fourth quartiles relative to that of the central galaxies. By normalizing to the central galaxies, this constructs a dimensionless "relative" conformity strength that can be used to intercompare different properties. $\mathcal{S}(R)$ is unity if the neighbours show as much difference as centrals, while it is zero if they show no difference, and can be negative if the satellites' difference is in the opposite sense to the centrals'.

- $\mathcal{S}(R < 2\text{Mpc})$ is strongest for environment, but this strength declines with halo mass. Low-mass ($M_h \sim 10^{11} M_\odot$) haloes show very weak conformity ($\mathcal{S} \sim 0.1 - 0.2$) in all quantities, but for sSFR and $f_{\rm H2}$ this rises to moderate strengths ($\mathcal{S} \sim 0.3 - 0.5$) by $M_h \sim 10^{12} M_\odot$. Age and $f_{\rm HI}$ always show relatively weak conformity. Metallicity, interestingly, shows increasing anti-conformity ($\mathcal{S} < 0$) with halo mass, possibly owing to the increasing number of satellites coupled with the reduced gas-phase metallicity in centrals from small amounts of fresh accretion (Figure 2.11).

- Consistent with qualitative impressions, $\mathcal{S}(R)$ for low-mass haloes is weak but present at all R in most quantities, while in massive haloes it is strong at small radii but declines rapidly and vanishes at $R \gtrsim 1$ Mpc. At the smallest radius, *i.e.* for satellite galaxies, $\mathcal{S}(R)$ is strongest for sSFR (or equivalently galaxy colour), and even in low-mass haloes this declines relatively quickly with R (Figure 2.12).

- For quantities that show significant conformity, $\mathcal{S}(R)$ is dominated by one-halo conformity over two-halo. An exception to this is for $f_{\rm HI}$ in small haloes, where small satellites of gas-rich centrals can get their neutral gas stripped, resulting in more similarity in this quantity with neighbours farther out (Figure 2.13).

- Tracking conformity back in time for the $z = 0$ galaxy population, we find that conformity is a late-time phenomenon for low-mass haloes for all properties except metallicity. Massive haloes develop conformity earlier,

and the sSFR conformity is generally always the strongest, with $\mathcal{S}$ significantly exceeding unity at intermediate redshifts. These trends are consistent with the idea that significant conformity only occurs once halo start to be quenched (Figure 2.14).

- Tracking conformity within fixed halo bins at all redshifts, we find that conformity strength is fairly constant, with low-mass haloes always showing very weak conformity and high-mass haloes showing stronger conformity (Figure 2.15).

- Our conformity results are generally qualitatively but not quantitatively consistent between our fiducial simulation and our $25h^{-1}$Mpc volume with 8× better resolution, in part due to the different performance of the model at different resolutions. In fact, at our fiducial resolution, our feedback model appears to overproduce low-mass quenched satellites (Figure 2.4), particularly in quenched haloes. It is likely that part of the strong one-halo conformity signal in our massive haloes arises from such overquenched satellites.

Overall, galaxy conformity appears to be a generic prediction of models that quench massive galaxies approximately in accord with observations. However, the strength, extent, and dependence on specific property are all likely to depend significantly on the precise physical model driving the quenching. In our case, we have tested a heuristic but observationally-consistent scenario where the hot gas in massive haloes is kept hot, which yields concordant central galaxy properties, and showed that it has a substantial impact on galaxy conformity developing in massive haloes. The different behaviours of our models at different resolutions might be due to our ad-hoc quenching prescription or the stellar feedback scaling relation. The smaller box with higher resolution produce less quenched structures that highly contribute in the change of the thermal properties of the gas and consequently alter the star formation history of the galaxies. The decoupling and coupling of the ejected gas volume elements with the hydrynamical forces could also result in different behaviour of the surrounding medium. The effects of resolutions on those prescriptions is less explored but we keep them for the future when appropriate data for comparisons are available. This chapter presents a first step towards quantifying conformity in a way that will allow observations to be compared to models more stringently,

and thereby potentially provide a new way to constrain AGN feedback in modern galaxy formation models. With such environmental effects on galaxy evolution and properties, the next chapter will focus on the HI content and SFR of galaxies in groups, particularly the removal and depletion timescales.

CHAPTER 3

MUFASA: Timescales for HI consumption and SFR depletion of satellite galaxies in groups

Abstract

We investigate the connection between the HI content, SFR and environment of galaxies using a hydrodynamic simulation that incorporates scaling relations for galactic wind and a heuristic halo mass-based quenching prescription to regulate the simulated cosmic star formation history and reproduce the galaxy stellar mass distribution of the observed Universe. We focused on 2 zoomed-in groups of galaxies, selected based on their masses and their hierarchical formation histories. The delay time τ_d, time during which the satellite galaxies behave similarly to central galaxies, ranges from $\sim 1-3$ Gyr at $z = 0$ depending on the halo mass they first fall into: the higher the halo mass the shorter the τ_d. At $z \sim 1$ we find $\sim 0.3 \lesssim \tau_d \lesssim 2$ Gyr. Lower stellar mass galaxies at infall time have higher τ_d. The timescales are identical for HI depletion and SF quenching. The fading time τ_f, time during which the HI content and the SFR exponentially drop, ranges between ~ 150 Myr at $z \sim 0$ and ~ 80 Myr at $z \sim 1$. Halo mass of the galaxy first infall does not show any effect on τ_f. For a given distance from the main halo of interest, higher redshift galaxies have higher cold gas content but are as efficient as lower redshift galaxies with lower gas content to form stars. The redshift trend remains for the highest halo mass at first infall but vanishes at lower masses. Given the same amount of HI, galaxy can form stars more effectively if they live in more massive structures. Difference in halo mass of infall does not affect the conversion of H_2 into stars. Given that galaxies mostly live in groups and clusters, our finding is useful prediction for the upcoming 21cm signal surveys such as LADUMA on MeerKAT and further surveys on SKA.

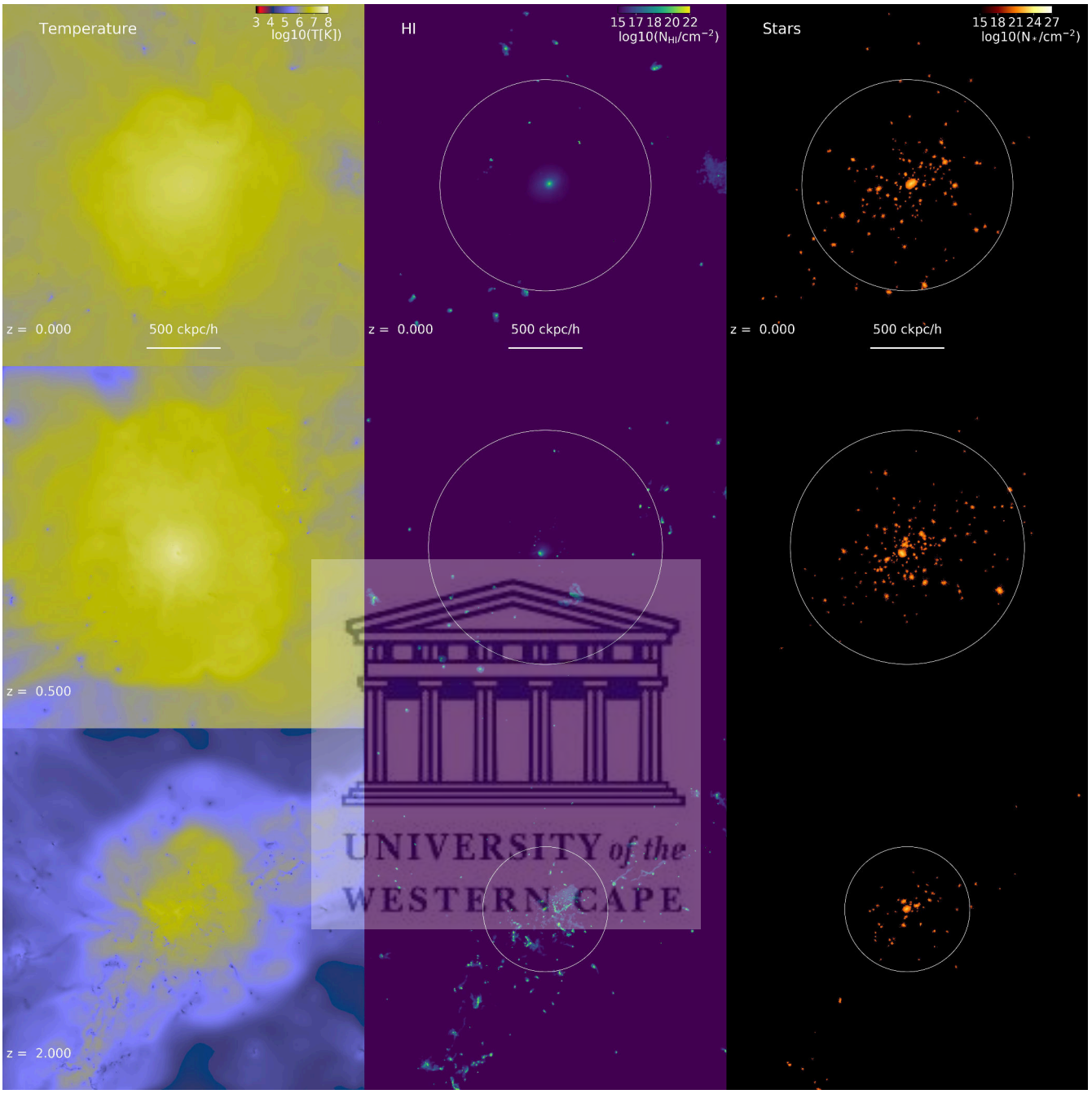


FIGURE 3.1: Temperature (*left*), HI (*middle*) and stellar (*right*) distributions of the zoomed-in halo (Halo 10) at different redshift.

3.1 Introduction

Gas inflow is the main fuel for *in-situ* star formation in galaxies. The current understanding of galaxy evolution stipulates that star formation-driven feedback mechanisms eject some of the infalling gas (Somerville & Davé, 2015),

preventing it from directly cooling and forming stars. In addition to that, the immediate (e.g. Weinmann et al., 2006b; Lacerna et al., 2018) or extended (e.g. Kauffmann et al., 2013; Rafieferantsoa & Davé, 2018) environment of galaxies can also affect their growth. It has long been argued theoretically that the cold gas content of galaxies is ram pressure-stripped when galaxies move at high velocities through a dense medium within a massive host halo (Gunn & Gott, 1972). Furthermore, the extended diffuse gas around the galaxies can be removed by the hot gaseous environment of dense structures such as galaxy groups or clusters (Larson et al., 1980). Relatively recent observations favour such claims (e.g. van Gorkom et al., 2003; van Gorkom, 2004).

Star formation of galaxies is then destined to end for galaxies moving within such environments, but depending on the circumstances, it happens on different timescales. The enormous amount of data provided by the Sloan Digital Sky Survey (SDSS) has permitted many studies of the effects of galaxy environments on the star formation quenching timescales. Kauffmann et al. (2004) used the SDSS data to analyse the correlation between star formation history and different observed time indicators such as the 4000$\AA$ break strength and the Balmer-absorption index Hδ_A: they found no dependence with environment, indicating a star formation timescale not less than 1 Gyr. Peng et al. (2010) used SDSS with zCOSMOS (Lilly et al., 2007) to disentangle two different quenching mechanisms. First, the *environment quenching* –star formation quenching related to the location of the galaxies – is less of a function of redshift out to $z \sim 1$, which they argued to be the consequences of the formation of large-scale structures leading to the end of star formation of more than half of the satellite galaxies. Second, the *mass quenching* – star formation quenching due to the size of the galaxies – they found to vary on a shorter timescale and that is proportional to the star formation rate of the galaxies. Hirschmann et al. (2014) estimated a timescale of ~ 5 Gyr for the local Universe low mass satellite galaxies seen within SDSS. Recently, Fossati et al. (2017) used a different set of data to look at the environmental processes affecting the galaxy evolution. With galaxies from the five CANDELS/3D-*HST* fields (Grogin et al., 2011; Koekemoer et al., 2011; Brammer et al., 2012), they found a quenching timescale of 2 – 5 Gyr. The quenching processes were shown to be less dependent of the host mass but quantitatively correlate with galaxy stellar masses and redshifts such that smaller galaxies at lower redshift have longer quenching time. They also argued that the galaxies used in their analysis stop forming stars due to

fuel (gas) depletion. In addition, their finding corroborates the *delayed-then-rapid* quenching model (Wetzel et al., 2013): they found a longer delay phase where the satellite galaxies indistinguisably behave like central galaxies and a short phase (~ 0.5 Gyr) where the star formation rate drops below the threshold limit.

Much work has also been done on satellite quenching from the theoretical side. McGee et al. (2009) studied accretion history with a semi-analytic galaxy sample covering a wide range of environment, *i.e.* from groups to clusters. Their sample showed that cluster galaxies typically originate from smaller group galaxies, and as a consequence the environmental effect on the galaxies lasted > 2 Gyr. Their finding suggests that before $z = 1.5$, galaxies should be mildly or not experiencing any environment processes. Simha et al. (2009) used an SPH cosmological simulation and found a decreasing but continuous gas accretion of the satellite galaxies. They showed that the gas depletion happens within $\lesssim 1$ Gyr timescale.

De Lucia et al. (2012) used a semi-analytic galaxy formation model based on the Millennium Simulation to conclude that the majority of small galaxies are pre-processed satellite, *i.e.* they were previously located in smaller groups prior to their current one. With such an origin, their model predicted a relatively long timescale for galaxies to lose their gas and halt their star formation (~ 6 Gyr). Wetzel et al. (2015) predicted in his letter the rapid environmental quenching timescale. They used 48 high-resolution dark matter simulation from Exploiting the Local Volume in Simulations (ELVIS) and found a shorter quenching time of $\lesssim 2$ Gyr for less massive ($\sim 10^8$ $M_\odot$) satellite galaxies (with an additional $1 - 2$ Gyr when including the pre-processing event). The quenching timescale positively correlates with the stellar mass of the satellite galaxies up to $\sim 10^9$ $M_\odot$ before decreasing towards more massive objects.

In our previous work, Rafieferantsoa et al. (2015) used a cosmological hydrodynamical simulated galaxy sample and found a halo mass dependent timescale where galaxies in more massive structures lose their gas faster than those living in less massive hosts: ~ 1 and ~ 2.5 Gyr e-folding times respectively. The range of predicted quenching times suggest that the timescales depend on a variety of factors, including physical ones such as halo mass and redshift, and perhaps numerical ones such as resolution and hydrodynamic methodology (Agertz et al., 2007). Therefore it is undeniable that we need to further quantify the timescale

of galaxy quenching, using cosmologically-situated simulations spanning a range of masses but with sufficiently high resolution that adequately models the stripping processes.

In this chapter we use high resolution simulated galaxies from a cosmological zoom simulation of two $\sim 10^{13} M_{\odot}$ galaxy groups using the Gizmo code that adequately handles mixing instabilities (Hopkins, 2015). In particular we study satellites within our groups to directly quantify the duration from infall until when the galaxies lose their gas and stop forming stars. We find a sudden transition for our galaxies from being gas rich to gas poor, along with a similar scenario in terms of star formation. Our quenching timescale follows the *delayed-then-rapid* scheme, and we quantify both the delay time and the fading time (the time to fully quench after quenching begins) in our simulations. The delay time drops rapidly with the infall mass of the satellite, and also decreases at higher redshifts. In all cases the fading time is much smaller than the delay time, and it is somewhat longer for more massive infalling objects.

§3.2 briefly reviews the Mufasa simulation used for this work, with the description of the method used to get our galaxy sample as well as the progenitors tree. In §3.3, we look at the distribution of galaxy properties used in this work. In §3.4, we quantify the depletion timescale of HI and SFR while §3.5 and §3.6 show the relation between gas and SFR, and gas and halo mass respectively. We summarize our findings in §3.7.

3.2 Simulations

3.2.1 Models

To analyse the evolution of galaxy HI content, we use the Mufasa galaxy formation models fully described in Davé et al. (2016). The same model was used in Rafieferantsoa & Davé (2018) with minor differences. This section briefly describes the main prescriptions in Mufasa necessary for galaxy formation as well as the changes we made to fit the context of this work.

First is the star formation rate (SFR) prescription that is based on the molecular content of the gas particles. It follows a power law scaling, namely Schmidt (1959)-law that reads

$$\mathrm{SFR} = \epsilon \frac{\rho f_{\mathrm{H2}}}{t_{\mathrm{d}}} \tag{3.1}$$

where $\epsilon = 0.02$ is the star formation efficiency (Kennicutt, 1998), f_{H2} the molecular hydrogen fraction in the gas volume element and t_{d} the dynamical time of the gas with density ρ. The star formation prescription is only applied to gas particles above a number density threshold of $\mathrm{n_{thresh}} = 0.13\ \mathrm{cm^{-3}}$. Such high density is only reached when the gas particles undergo radiative cooling. Mufasa uses the Grackle 2.1 library[1] for such implementation, that accounts for the H/He-elements and metal-line cooling. During the simulation, a uniform background metagalactic radiation field is assumed following Faucher-Giguère et al. (2010).

The galactic wind produced by the supernovae energy as well as the radiation pressure due to the evolution of the star population is also accounted in our model. Instead of modeling them individually, Mufasa uses scaling relations from the Feedback In Realistic Environment (FIRE, Hopkins et al., 2014) simulations analysed in Muratov et al. (2015) that combine their effects. The outflowing rate η of the gas elements and their outflowing speed v_w read

$$\eta = 3.55\left(\frac{M_*}{10^{10}\,\mathrm{M_\odot}}\right)^{-0.35} \quad \text{and} \quad v_w = 2v_c\left(\frac{v_c}{200\ \mathrm{km\,s^{-1}}}\right)^{0.12} \tag{3.2}$$

where M_* and v_c are the stellar mass and circular velocity of the galaxy which the gas volume element belongs to. We eject the outflowing gas perpendicular to the galaxy plane.

To prevent the overproduction of massive ellipticals, Mufasa uses a heuristic prescription based on the ideas in Gabor & Davé (2015) to quench the star formation in massive structures. The idea is to stop further supply of gas inflow for galaxies within host haloes above a threshold quenching mass of $M_q = (0.96 + 0.48z) \times 10^{12}\ \mathrm{M_\odot}$ (Mitra et al., 2015). This is done by heating all the gas to the virial temperature $T_{vir} = 9.52 \times 10^7 M_h^{2/3}$ (Voit, 2005) of the haloes. The self-shielded gas in the interstellar medium of the galaxies is kept unaffected by this recipe. One significant difference with respect to the implementation

[1] https://grackle.readthedocs.io/en/grackle-2.1/genindex.html

in Mufasa is that we only heat the gas particles within the inner 25% of the virial radius of the haloes to mitigate the overproduction of quenched satellite galaxies found in Rafieferantsoa & Davé (2018).

3.2.2 Primary box

The cosmological parameters values used in this work are $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, $\Omega_b = 0.048$, $H_0 = 68\ \mathrm{km\,s^{-1}\ Mpc^{-1}}$, $\sigma_8 = 0.82$ and $n_s = 0.97$, corresponding to the Planck et al. (2016) inference.

Our primary run has $50h^{-1}$Mpc on a side with 512^3 dark matter particles only. The simulation starts at $z = 249$ with an initial condition generated with Music (Hahn & Abel, 2011). Once at $z = 0$, the haloes are identified using the friends-of-friends algorithm with a linking length of 2% of the mean inter-particle distance.

3.2.3 Refined regions

We apply a zoom-in technique for 2 carefully chosen haloes within our primary box. To investigate the effect of our quenching prescription on our satellite galaxies, we selected halo 10 (H10, hereafter) and halo 43 (H43, hereafter) to have $10^{12}\ \mathrm{M_\odot} < \mathrm{M_{halo}} < 10^{14}\ \mathrm{M_\odot}$. This is to ensure that our central galaxies passed through the quenching processes, which is typically around $\sim 10^{12}\ \mathrm{M_\odot}$, before $z = 0$. Our halo sample is re-simulated with dark matter particles 64 times smaller than those from the primary box. The refined dark matter particles are then split into gas volume elements and dark matter particles following our adopted baryon fraction (see §3.2.2). Our choice of haloes to be refined is based on their evolution history. Apart from the mass criterion, we make sure that the haloes did not have any major fly-by or merger during their evolution. This is to ensure that the haloes are self-evolved and did not change their physical properties by interacting with other structures with similar size. When resimulating, the center of mass of the highly resolved particles is shifted at the center of the cubical box to avoid any boundary effects. Properties of the haloes presented in this work are sumarized in Table 3.1. We show in Figure 3.1

the 2D distributions of the gas temperature (left panels), the HI number density (middle panels) and the stellar number density of H10 at different redshift (rows).

Figure 3.2 shows the total dark matter (full lines) and stellar (dashed lines) masses evolution of H10 (dark blue) and H43 (light green). For illustration, Figure 3.A show 3 haloes selected at $z = 0$ and satisfying our conditions. The upper panels are 2D distributions of the dark matter particles viewed along 3 different axes. The lower panels are the distributions of the corresponding particles within each $z = 0$ halo at $z = 250$.

TABLE 3.1: Zoomed-in haloes

haloes	original $M_{halo}(M_\odot)$	refined $M_{halo}(M_\odot)$	$SFR_{cen}(M_\odot/yr)$	# of satellites
H10 ($z = 0$)	5.578×10^{13}	6.6248×10^{13}	0	182
H43 ($z = 0.15$)	1.681×10^{13}	1.422×10^{13}	96.634	20

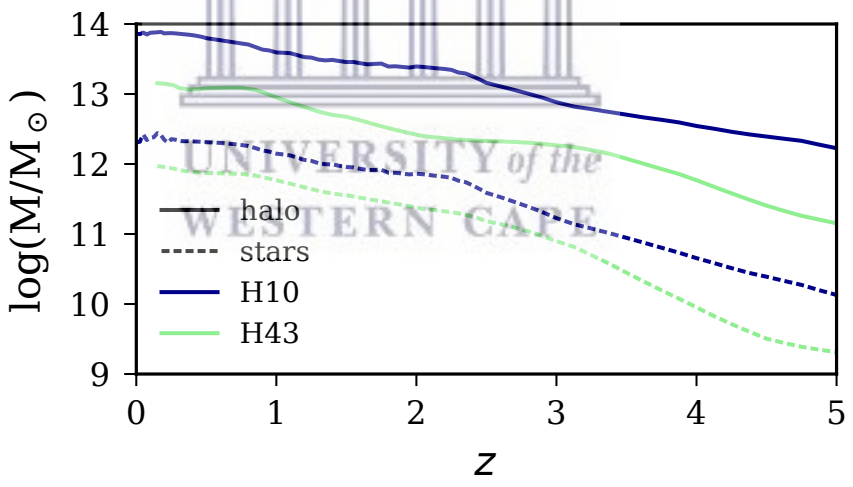


FIGURE 3.2: Total masses of the dark matter (full lines) and star (dashed lines) particles *vs* redshift z of H10 (dark blue) and H43 (light green).

Similarly to other MUFASA papers, we use SKID[2] galaxy finder. The properties of galaxies and haloes are obtained from CAESAR[3], which uses YT[4] simulation analysis suite as backend.

[2] http://www-hpcc.astro.washington.edu/tools/skid.html
[3] https://bitbucket.org/laskalam/caesar
[4] https://yt-project.org/

3.3 Galaxy properties

We are mostly interested in the gas content and star formation history of the galaxies in our host haloes. Figure 3.3 show the stellar mass (first columns), the HI mass (2nd columns), the H_2 mass (3rd columns) and the star formation rate (*right* columns) distribution of the galaxies in H10 (*lower* panels) and in H43 (*upper* panels). The *left*most panels show the growth of galaxy population inside the main halo as we go to lower redshift. Neutral (HI) and molecular (H_2) hydrogen content of galaxies are higher at higher redshift and those members of the host haloes slowly lose their gas towards $z = 0$. Galaxies are given lower limit gas content values once their HI (or H_2) mass goes below the x-axes limit of our figure. With such distributions of galaxy properties in our haloes, we can look at the evolution of HI and H_2 content modulo the required timescale for their consumption.

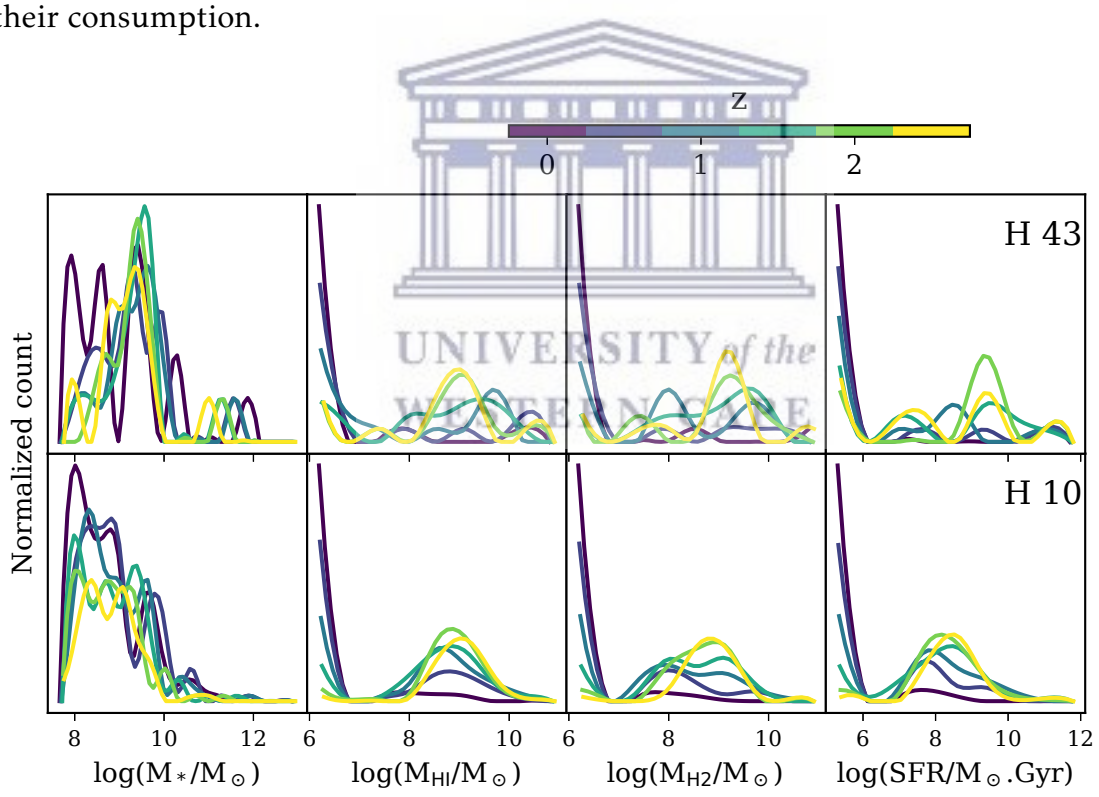


FIGURE 3.3: Evolution of the galaxy property distributions of our selected haloes. *Upper* panels are for H43 and *lower* panels for H10. Different colors represent the distributions at different redshift. Each columns show distributions for different galaxy properties.

With our simulated outputs spaced by ~ 300 Myr at the lowest redshift, and smaller time intervals at higher redshift, we can track individual galaxies based

on their stellar masses. We match the galaxies between two successive snapshots and identify the galaxies with the maximum number of similar star particles. With this method, the progenitor of a smaller galaxy flying-by a bigger galaxy would stay the progenitors of that bigger galaxy. To solve that problem, we also match the galaxies between snapshots at two time intervals.

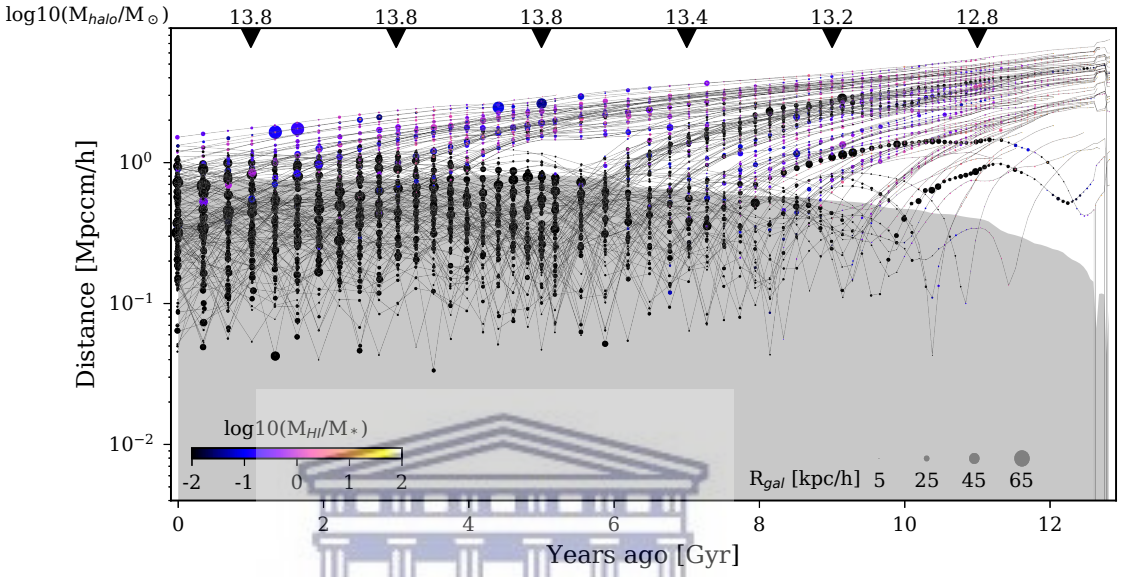


FIGURE 3.4: HI evolution of galaxies in H10. Each line shows the path of a galaxy relative to its distance to the center of main halo. The galaxies shown here are the galaxy members of the main halo at $z = 0$. The shaded area is the region inside the virial radii of the host. The evolution of the mass of the host is shown on the top of the panel.

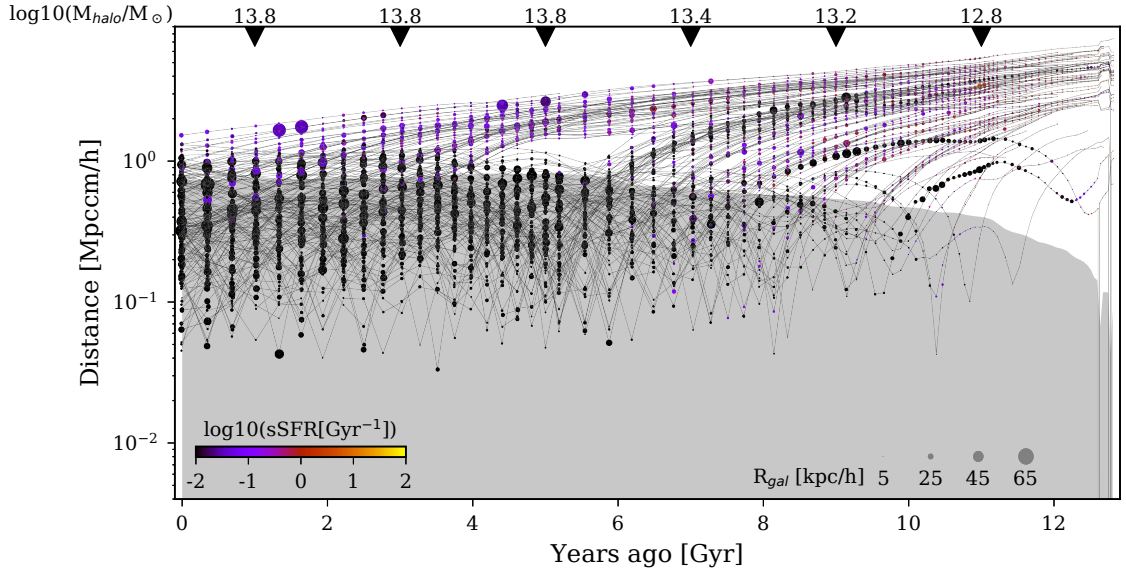


FIGURE 3.5: Similar to Figure 3.4 but showing for SFR (star formation rate) of the galaxies.

Figures 3.4 and 3.5 show the distances of the galaxies present in the host halo (H10) at $z = 0$ and their progenitors (connected with lines) earlier on. The size of the circles show the total size of the galaxies (gas+stars), and the color of the circles the properties of the galaxies (color coded). In this work, we are only concerned about the progenitors of the galaxies present in the main halo at $z = 0$. The shaded region is the area inside the host halo. We can clearly see that when galaxies cross the virial radius of the halo of interest, their HI-richness (or sSFR) is considerably depleted and the only growth of galaxies is via dry mergers. The following sections try to quantify the timescale required for such event. Througout this work, we use H10 and H43 data except for the latest redshift where we only use H10 because H43 is limited to $z \geq 0.15$.

3.4 Satellite quenching time scale

The loss of gas through the galaxy motions within their corresponding virialised regions and the consumption of gas in the interstellar medium via star formation resulted in estimated timescales t_q of 5−7 Gyr (Wetzel et al., 2013) or 1 – 2 Gyr (Oman & Hudson, 2016), depending on the environment they live in, before the galaxies cease to form stars. We can further quantify these timescales

using our highly resolved groups of galaxies to emphasise on the differences between the gradual starvation of the galaxies by lack of inflowing cold gas, and the rapid drop of star formation rate due to the remaining molecular content being used up.

We define the starting time t_0 of the gas depletion and the inhibition of star formation to be when the galaxies start to be satellite. We only include galaxies that were always central prior to t_0. Fossati et al. (2017) used the same definition of t_0 and they found a 2-5 Gyr timescale for quenching the satellites. They found that the fraction of quenched satellites is higher at lower redshift and higher halo masses. Surprisingly, they found no dependence of the passive central galaxy fraction with halo masses (see their fig. 17).

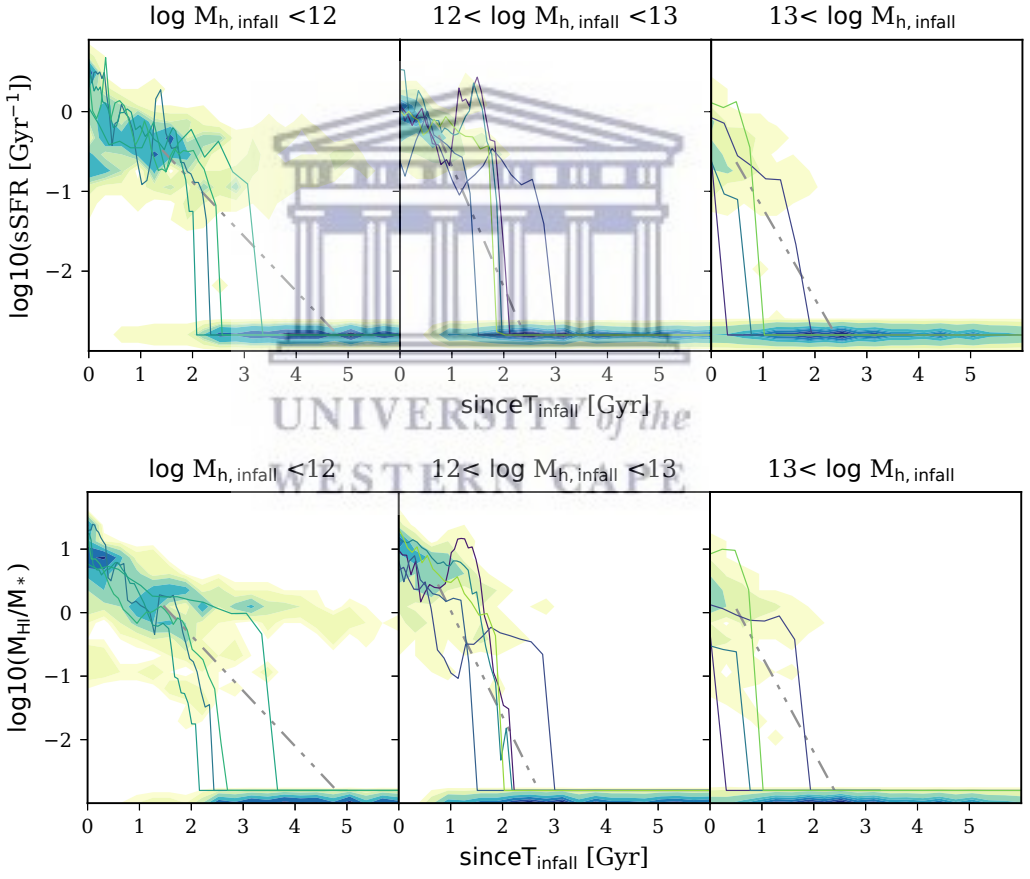


FIGURE 3.6: Specific star formation rate (*upper*) and HI-richness (*lower*) of galaxies since they become satellite. The mass ranges of the halo they fall into ($M_{h,infall}$) are shown on top of the panels. Note that all the galaxies belong to H10 at $z = 0$. It is clear that galaxies still form stars for ≥ 5 Gyr when $M_{h,infall} \leq 10^{12}$ $M_\odot$. That time decreases down to ~ 1 Gyr for $M_{h,infall} > 10^{13}$ $M_\odot$. The gray dashed lines connect the peak values of the upper and lower contours. The coloured lines show tracks of representative galaxies in each $M_{h,infall}$ range.

Figure 3.6 show our result with H10. Upper panels are for specific SFR and the lower panels for HI-richness of the galaxies. Each panel shows the 2D distribution of the galaxies and their progenitors since the time they became satellite (x-axes). The *left* panels are for galaxies with their first halo of infall to have $M_{halo} < 10^{12}$ $M_{\odot}$, the second panels with 10^{12} $M_{\odot} < M_{halo} < 10^{13}$ $M_{\odot}$, and the *right* panels for $M_{halo} > 10^{13}$ $M_{\odot}$. The dashed lines go from the mean values of the upper contours to the mean values of the lower countours. The dashed lines hint at a shorter (steeper) timescale towards higher host halo masses which is in disagreement with the halo mass independence found in Fossati et al. (2017).

To properly quantify the quenching timescale (t_q), we are splitting the latter into *Delay Time* – time during which the galaxy properties barely change – and Fading Time – time during which most of the depletion happens and the HI-richness (or specific SFR) crosses the threshold limit. We note that those definitions are similar to those used in Fossati et al. (2017). In our case the threshold below which the HI-richness (or sSFR) is considered insignificant is HI-richness $_{lim} = 10^{-2}$ (or sSFR $_{lim} = 10^{-2}$ Gyr^{-1}). The delay time is the duration between t_0 and the earliest snapshot before the exponential decay occured, whereas the fading time is the duration from the beginning of the exponential decay to the next time the quantity is below the threshold. The fading time typically happens within 2 successive snapshots in our simulation. In that case we interpolate the value based on an exponential fit.

3.4.1 Delay Time

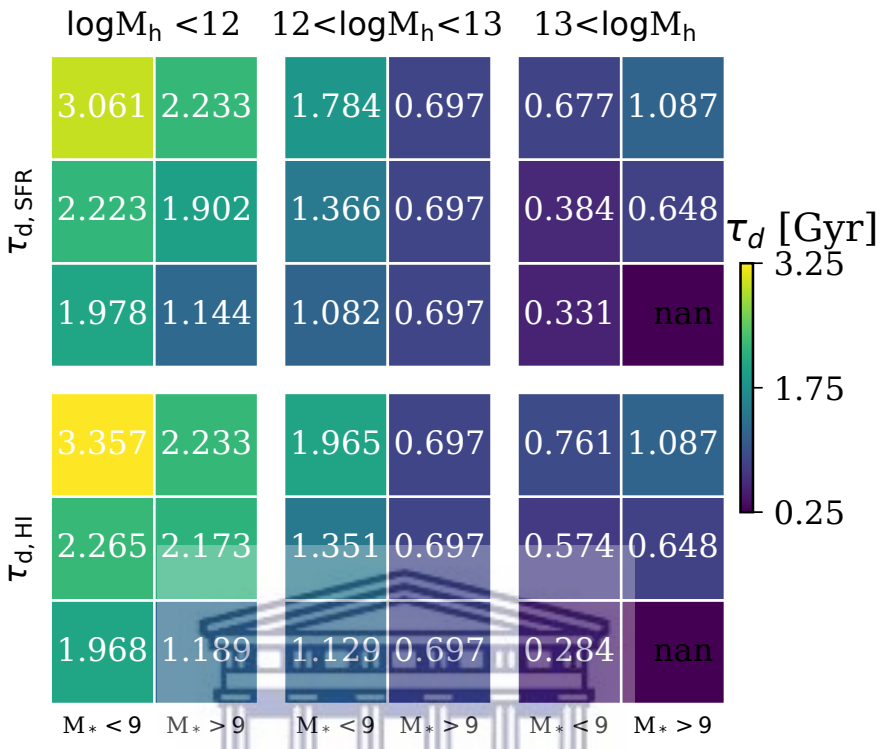


FIGURE 3.7: Delay time τ_d for sSFR (*upper* panels) and HI-richness (*lower* panels). *Left* panels are for galaxies becoming satellite in $M_{h,infall} < 10^{12} M_\odot$, *center* panels for $10^{12} \leq M_{h,infall} < 10^{13} M_\odot$ and *right* panels for $10^{13} M_\odot \leq M_{h,infall}$. Each panel is further divided into 6 area: 2 columns for stellar mass bins at infall ($M_{*,infall} < 10^9 M_\odot$ *left* and $M_{*,infall} \geq 10^9 M_\odot$ *right*) and 3 rows for different redshifts (*upper* for $z = 0$, *middle* for $z = 0.5$ and *lower* for $z = 1$). See Figure 3.B for the distribution of galaxies in terms of τ_d.

Figure 3.7 shows the sSFR and HI-richness delay times τ_d: *upper* and *lower* panels respectively. *Left* to *Right* panels show the results for increasing halo masses within which the galaxies first fall into. Each panel is divided into 6 area: 2 column-wise for binned stellar masses of the infalling galaxies at t_0 (*left*: $M_{*,infall} < 10^9 M_\odot$ and *right*: $M_{*,infall} \geq 10^9 M_\odot$) and 3 row-wise for different redshift (*upper* for $z = 0$, *mid* for $z = 0.5$ and *lower* for $z = 1$). The numbers of the figure show the average τ_d. We note that at higher redshift, we only track the galaxies present in the main halo progenitors at that redshift, *i.e.* we do not include the progenitors of galaxies that are present in the $z = 0$-main halo but did not cross the virial radius at the given redshift.

The mean of the delay times is the longest for the least massive host halo at $z = 0$ ($\sim$ 3 Gyr), and decreases down to $\sim$ 1 Gyr by $z \sim 1$. The galaxies falling in the most massive halo of infall take the shortest delay time: $\sim$ 1.5 Gyr faster than those falling in $M_{h,infall} \leq 10^{12}$ $M_\odot$ (*i.e.* $\leq$ 1 Gyr). Less massive satellite galaxies have longer delay time (1 Gyr more) than their massive counterparts, except for those falling in the most massive structure where this scenario is flipped. HI-richness and sSFR delay times show similar behaviour. We note that the similarity in values on the middle right column is due to low number of galaxies in those bins.

3.4.2 Fading Time

In the delayed-then-rapid scenario, the fading time τ_f is expected to be much shorter than the delay time. Fossati et al. (2017) computed the delay time by matching the star formation efficiency, quantified by the main sequence parameterisation of Wisnioski et al. (2015), between the 3D-HST sample and their mock catalogue. They estimated $\tau_f = t_q - \tau_d < 0.6$ Gyr.

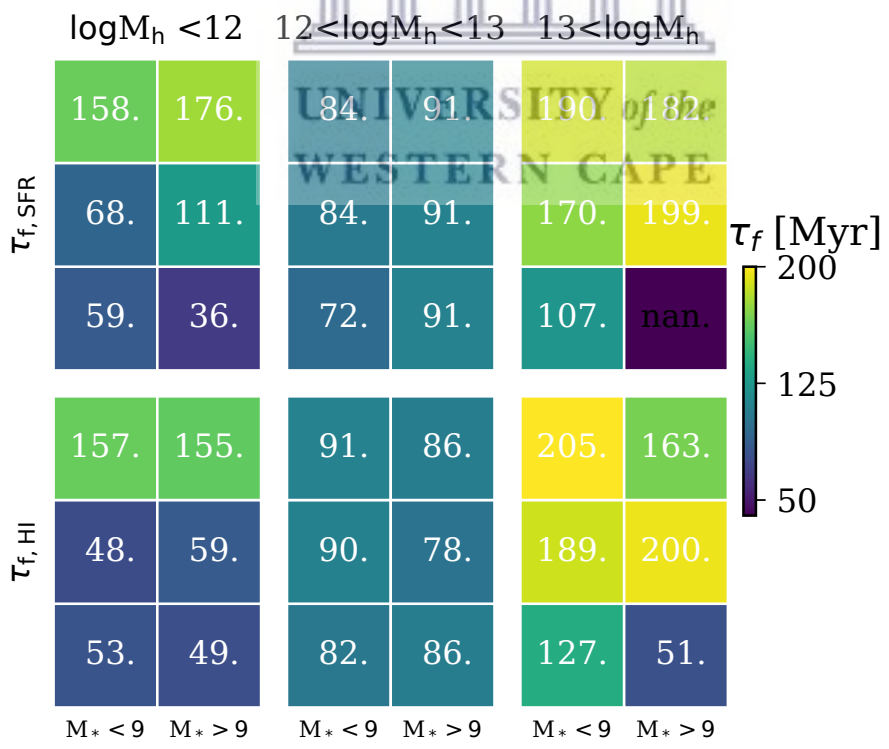


FIGURE 3.8: Similar to Figure 3.7 but showing for the Fading time τ_f binned by stellar mass. See Figure 3.C for the distribution of galaxies in terms of τ_f.

Figure 3.8 shows the mean fading times τ_f of the HI-richness (*lower* panels) and sSFR (*upper* panels). All scenarios show mean fading time of $\sim 50 - 200$ Myr. Generally, lower redshift (*upper* parts of each panel) infall requires longer τ_f to suppress the star formation and the neutral gas content compared to their counterparts at higher redshift (*lower* parts of each panel). Interestingly, galaxies falling into the most massive host tend to require more time to completely lose their gas and stop star formation: ~ 185 Gyr for sSFR and ~ 200 Myr for HI-richness for galaxies at $z = 0$, which might be due to our limited sample and requires further analysis. The stellar mass of the galaxies does not show any trend in the fading time, *i.e.* difference between the *left* and *right* areas in each panel.

In summary, the delay-time τ_d is longer for lower stellar and halo masses until the haloes exceed 10^{13} $M_\odot$. In the latter case, less massive galaxies have τ_d of ~ 400 Myr faster than more massive ones. Galaxies have longer τ_d when they become satellite recently. The case is the same for both HI and sSFR which suggests that the SFR is being reduced owing to the suppression of the extended gas reservoir. Fading-time τ_f shows the opposite scenario. Generally, galaxies falling in high halo masses have longer τ_f than those in low mass haloes. The case is less clear at $z = 0$ but become more apparent at higher redshift. The difference for lower and higher galaxy stellar masses is minor if nonexistent. At face value, we find similar trends with Fossati et al. (2017) in terms of stellar mass but our timescale dependence with halo mass as well as our shorter timescales by few Gyr were not found in their 3D-HST sample.

3.5 Relationship between gas content and Star formation

The cooling and coalescence of neutral hydrogen into dense molecular clouds are important steps before the star formation. However, the formation of stars can disrupt the collapse of cold gas and therefore reduce later star formation. In this section, we quantify the star formation activity given the amount of cold atomic and molecular gas.

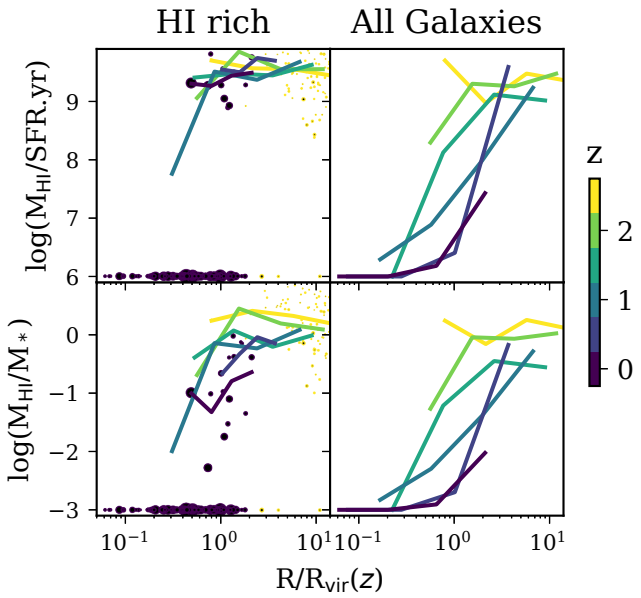


FIGURE 3.9: HI-consumption time (*Upper*, $\log(M_{HI}/SFR)$) and HI richness (*Lower*) of galaxies *vs* their radial distances from the center of the host halo. The circles show the galaxies at the lowest (purple) and highest (yellow) redshift presented here. The lines show the mean values, colour coded by the redshift. Higher redshift galaxies are the progenitors of the low redshit galaxies. The low redshift galaxies are only the members of the main halo. Galaxy properties below 6 (for HI-consumption time) and −3 (for HI-richness) are given those lower limit values to confine the figure. *Left* panels show the mean values for only gas rich galaxies but *Right* panels show the mean values for all galaxies.

Left panels of Figure 3.9 show the evolution of HI-richness (*lower* panel) and HI-consumption time (M_{HI}/SFR, *upper* panel) with respect to their distance to the main host and its progenitors center of masses. Note that we only follow all the progenitors of the galaxies present in the $z = 0$ main halo. The purple dots are the galaxies at $z = 0$, with the purple lines showing the mean values. Yellow dots are the progenitors of the $z = 0$ galaxies at $z = 2.5$, with the yellow lines the mean values. Lines with different colors are the mean values at different redshift as shown in the colorbar. We did not show the galaxies themselves to avoid cluttering of the figures. Galaxies without cold gas content are shown at the bottom of each panel but not included in the calculation of the mean quantities. This is because we are only interested in those that still have HI. The virial radii used in the x-axes is for the progenitors of the main halo not for the individual galaxies.

A first look at the *left* panels of Figure 3.9 indicates that our galaxy sample has higher gas content at higher redshift, they increasingly lose their gas until today (*lower* panel), and galaxies that crosses the virial radius contain less HI. The HI-consumption time (*upper* panel) is barely a function of redshift and distance to the center of the main halo progenitors. This reinforces the idea of gas consumption via star formation, and the quenching of the latter by depletion of the former. This scenario remains the same at different cosmic times. We speculate that the dichotomy present in HI-consumption time being analogous to that in HI-richness indicates the existence of HI-richness threshold below which further star formation is unlikely to happen without any extra supply of cold gas and above which the star formation rate correlates directly with the HI content, even though the star formation is actually occuring within the H_2 region.

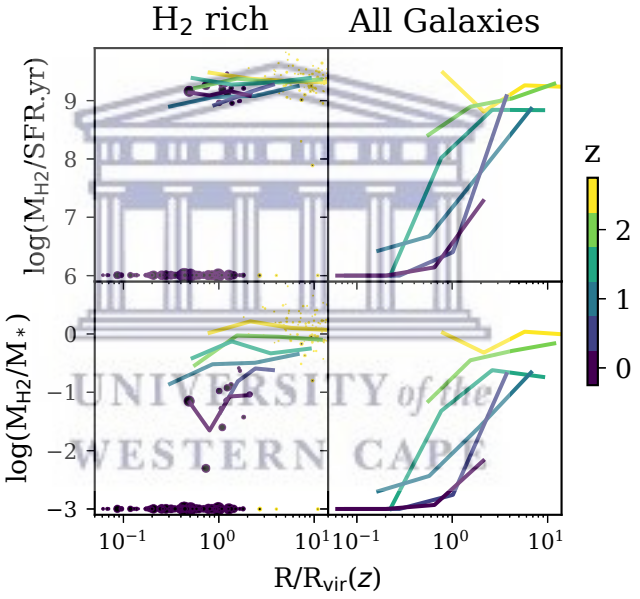


FIGURE 3.10: Similar to Figure 3.9 except showing for H_2 content.

Left panels of Figure 3.10 show the H_2-richness and H_2-consumption time to further analyse the gas consumption. The relationship is cleaner compared to the previous one, mainly due to our star formation prescription scaling with the H_2 fraction (f_{H2}, see equation 3.1). Again, the pattern shows that galaxies contain less gas in molecular form at lower redshift than at higher redshift (*lower* panel) and their star formation rate proportionally declines with it (*upper* panel), *i.e.* unchanged H_2-consumption time at different radial distance and different redshift. The similarity between the HI-consumption time and

the H_2-consumption time suggests the HI to be a direct tracer of star formation rate above certain mass threshold. To test this, we can increase our group sample and vary the gas element density thresholds for self-shielding and star formation. We plan to investigate on this in future work.

Statistically, we can do the same exercise but including all the galaxies regardless of their gas content. In this case, we are looking at the general trend rather than the process towards being quenched as we argued previously. *Right* panels of Figures 3.9 and 3.10 are similar to the *left*, except that we now show the mean values for all the galaxies. The trend remains the same at the highest redshift, but the growth of no-cold-gas galaxies at lower redshift shifted down the mean values.

In short, the gas consumption timescale is neither a function of the group-centric distance nor the cosmic time. This is seen in both atomic and molecular hydrogen contents of the galaxies suggesting that above a certain mass threshold, HI alone can be a direct indication of star formation.

3.6 Relationship between gas content and Halo of infall

We showed previously the quicker delay/fading times when the galaxies fall into more massive hosts. In that vein, we look at the change of gas-richness and gas-consumption time depending on the first halo of infall.

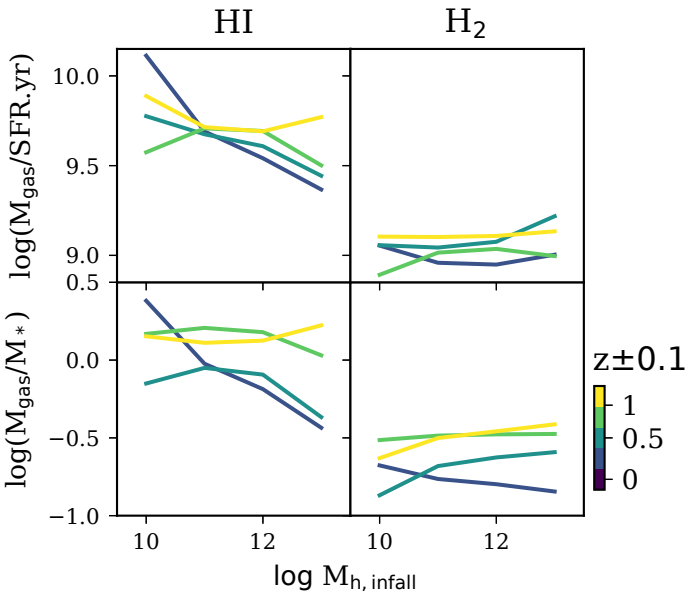


FIGURE 3.11: Gas-consumption time (*Upper*) and gas-fraction (*Lower*) of galaxies *vs* the halo mass they first fall into. *Left* (*Right*) panels are for HI (H_2).

Figure 3.11 shows the gas-richness (*lower* panels) and gas-consumption time (*upper* panels) of galaxies *vs* the mass of the haloes they fall into. The *left* panels show for HI content whereas the *right* panels for H_2 content. Generally, the decrease of HI-richness is faster when it falls in more massive haloes (*lower* panel) similar to what was found in Rafieferantsoa et al. (2015). However, such attributes are only true at lower redshift (purple, $z \sim 0$) but is not seen at the higher redshift explored here (yellow, $z \sim 1$). The *upper* panel shows that galaxies form star less efficiently early on for a given halo mass of infall: *i.e.* yellow > green > blue. The trend is very robust at higher halo mass of infall while at lower host halo masses, the trend becomes less coherent. This is due to the decreasing number of galaxies above our HI richness threshold at later time as infall in less massive host mostly happened earlier than infall in more massive ones. The HI-consumption time is independent of halo mass at infall at higher redshift but anti-correlates strongly at lower redshift.

In terms of molecular content, H_2-richness is less of a function of halo mass of infall than HI-richness and the correlation is barely present if not missing at any $z \lesssim 1$. H_2-richness at high halo mass of infall is higher at higher redshift, while it is less conclusive at the lowest masses. H_2-consumption time is

flat with respect to the halo mass of infall. We barely see any evidence of consumption time difference at different redshift that was relatively apparent with HI-consumption time at higher halo mass of infall.

Overall, HI, H_2 and SFR generally trace each other during infall. H_2-consumption time is not related to the size of the structure the galaxies fall into while that of HI is mildly anti-correlated to.

3.7 Conclusion

We used two zoomed-in groups of galaxies to study the timescales at which the HI content and the sSFR deplete when the galaxies become satellite. We also quantified the efficiency of gas content to form stars as well as the evolution of gas content depending on the halo mass of infall. We summarize our finding as follows

- Halo masses of infall have reversed effects on the delay timescale (τ_d) and fading timescale (τ_f) of the infalling galaxies. While τ_f increases with halo masses, τ_d decreases. Less massive galaxies have higher τ_d of ~ 1 Gyr slower than higher mass galaxies, while the stellar mass dependence with τ_f is less conclusive. Higher redshift galaxies spend $\gtrsim 50\%$ less τ_d for both the HI-content and the star formation rate.

- Higher redshift galaxies have higher HI-richness. It is independent of the radial distance to the center of mass of the main halo at higher redshift but starts to be a function of it at lower redshift such that closer to the center the galaxies are less HI-rich. There is little to no radial dependence of the HI and H_2 contents of galaxies and their efficiency to convert into stars, and the gas consumption time remains unchanged with respect to the distance to the main halo at different cosmic times explored here ($z \leq 2$).

- Galaxies can more efficiently convert HI-content into stars at lower redshift for more massive halo of infall. This scenario is not present for lower halo mass of infall. However, a general trend of higher effiency of converting HI-content into stars in more massive halo of infall is apparent at all redshift except perhaps for the highest redshift ($z \sim 2$) that displays independent behaviour with respect to how big the virialised structure

the galaxy falls in. H_2-richness as well as H_2-consumption time are not a function of the halo mass of infall. The higher H_2-richness and lower H_2-consumption time for higher redshift are barely visible at higher halo mass and even less at lower mass.

- This work contradicts with the halo mass independence of quenching timescales found in Fossati et al. (2017). Despite the improvement of the quenching prescription in Mufasa, the quenching timescales of the satellite galaxies are still relatively shorter by a factor of ~ 0.5 compared to the 3D-HST sample with $\sim 2-5$ Gyr (Fossati et al., 2017). Recent work by Foltz et al. (2018) suggests a halo mass dependence of the quenching timescales with a range of < 1.5 Gyr which are more inline with what we found considering they only looked at cluster member galaxies.

The environmental condition of galaxies to mostly be either in groups or clusters indicate the values of these findings for the upcoming extragalactic surveys such as those to be done on MeerKAT and ultimately on SKA. The results presented in this study argue for more detailed and computationally expensive simulations. Our future work will be extended to multiple groups and clusters with which we can verify our finding about the halo masses to have opposite effects on the fading timescale and the delay timescale. They will additionally clarify whether τ_f is really independent of the stellar mass of the galaxies at infall.

3.8 Appendix

To illustrate the zooming technique described in §3.2.3, we present in Figure 3.A the 2D density distributions of dark matter particles showing 3 selected haloes and their corresponding refined regions at the initial conditions. The upper panels are the dark matter distributions of the simulated 50 Mpc/h box at $z = 0$. Different columns show different projections of the box. The selected haloes are encircled with different colours. The white circles correspond to the virial radii of the haloes whereas 2.5× those radii are shown by the coloured circles. The circular boundaries shown in this figure translate to spherical boundaries in our calculations. We refine the particles within the coloured circles (spheres). At initial conditions (lower panels), the particles to be refined occupy

less of a spherical shape. This is due to the non-linear growth of structures. The lower panels show the 2D distributions of the refined particles at the start of the resimulation of $z = 250$. The colour of the countours correspond to the upper panels.

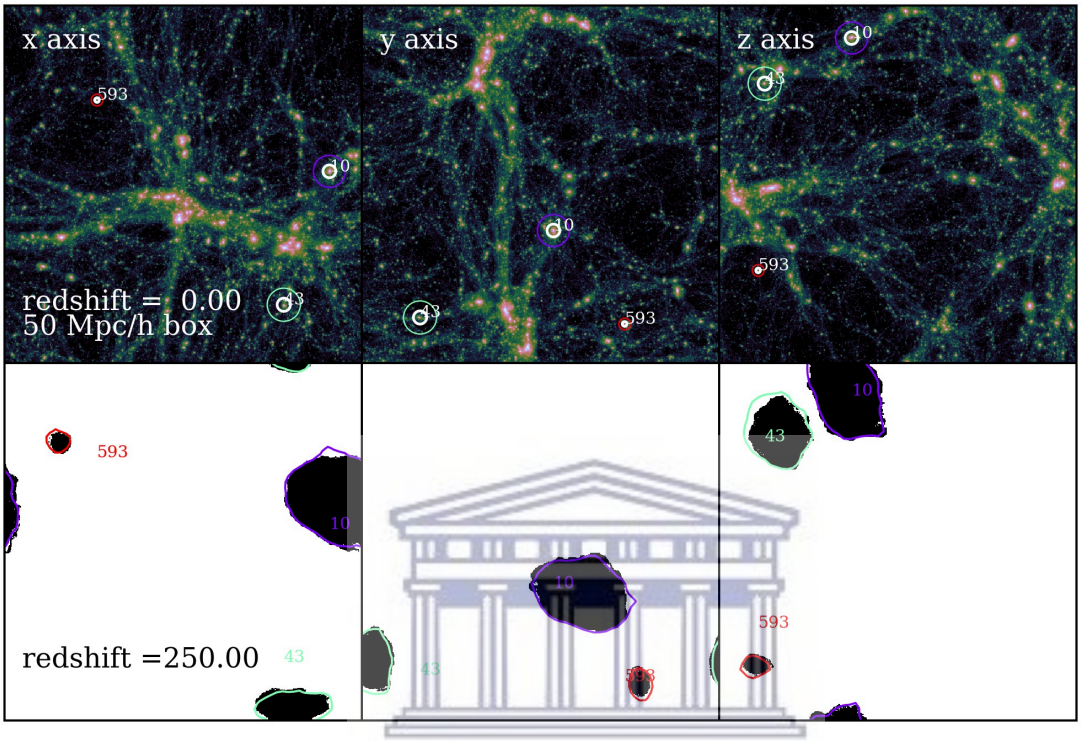


FIGURE 3.A: 3 selected haloes from the primary box. *Upper* panels show the distribution of dark matter particles at $z = 0$ for different projection axes (left to right). The white circles are the circles of the maximum radii of the haloes and the colored circles the regions containing the particles to be refined. *Lower* panels only show the original positions of the particles included in the colored circles in the upper panels at the initial condition ($z = 250$). The small numbers next to the haloes are identification numbers.

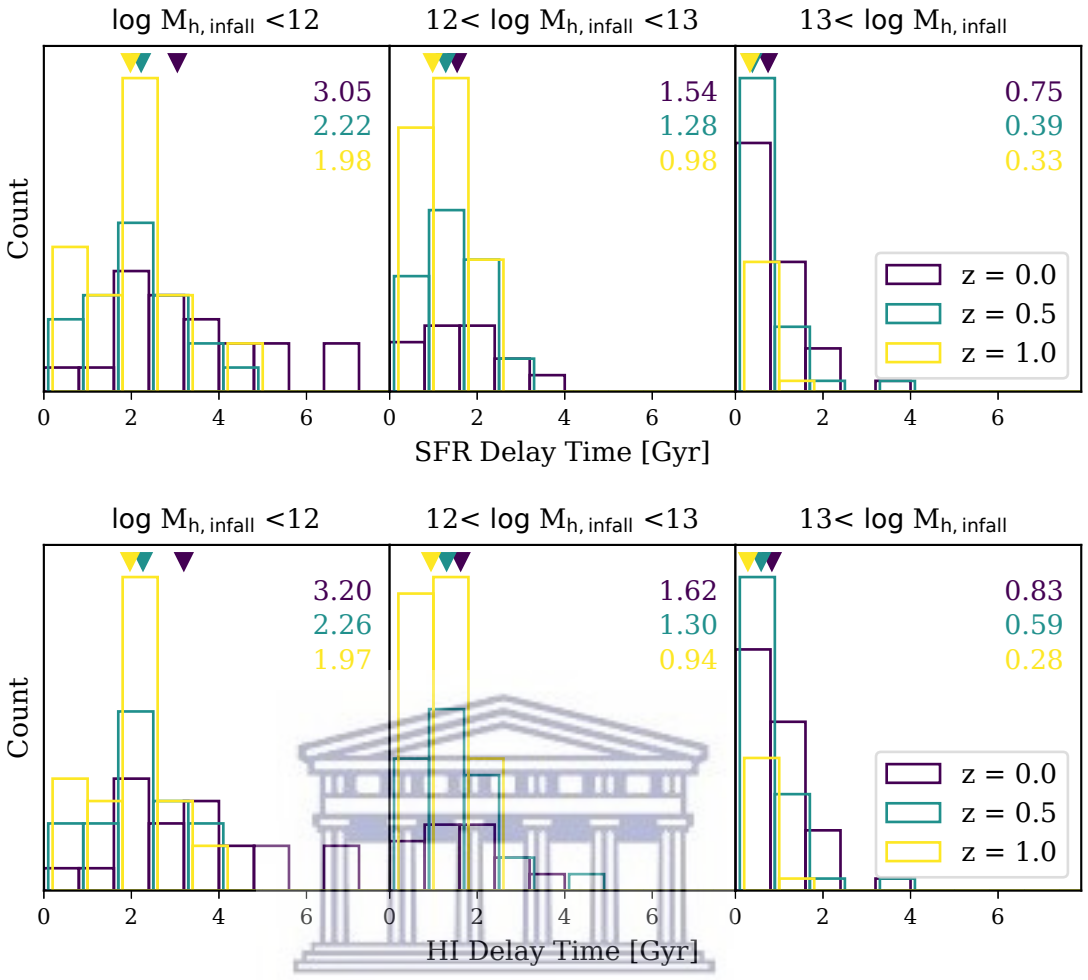


FIGURE 3.B: *Delay Time*. *Upper* (*Lower*) panels show the sSFR (HI-richness) delay time distributions. *Left* panels are for the galaxies first falling in $\log M_{h,infall} < 12$, the *Middle* panels for $12 < \log M_{h,infall} < 13$ and the *Right* panels for $13 < \log M_{h,infall}$. The downwards triangles on the top of the panels point to the mean values of the delay times where the corresponding values are shown on the top rights (color coded).

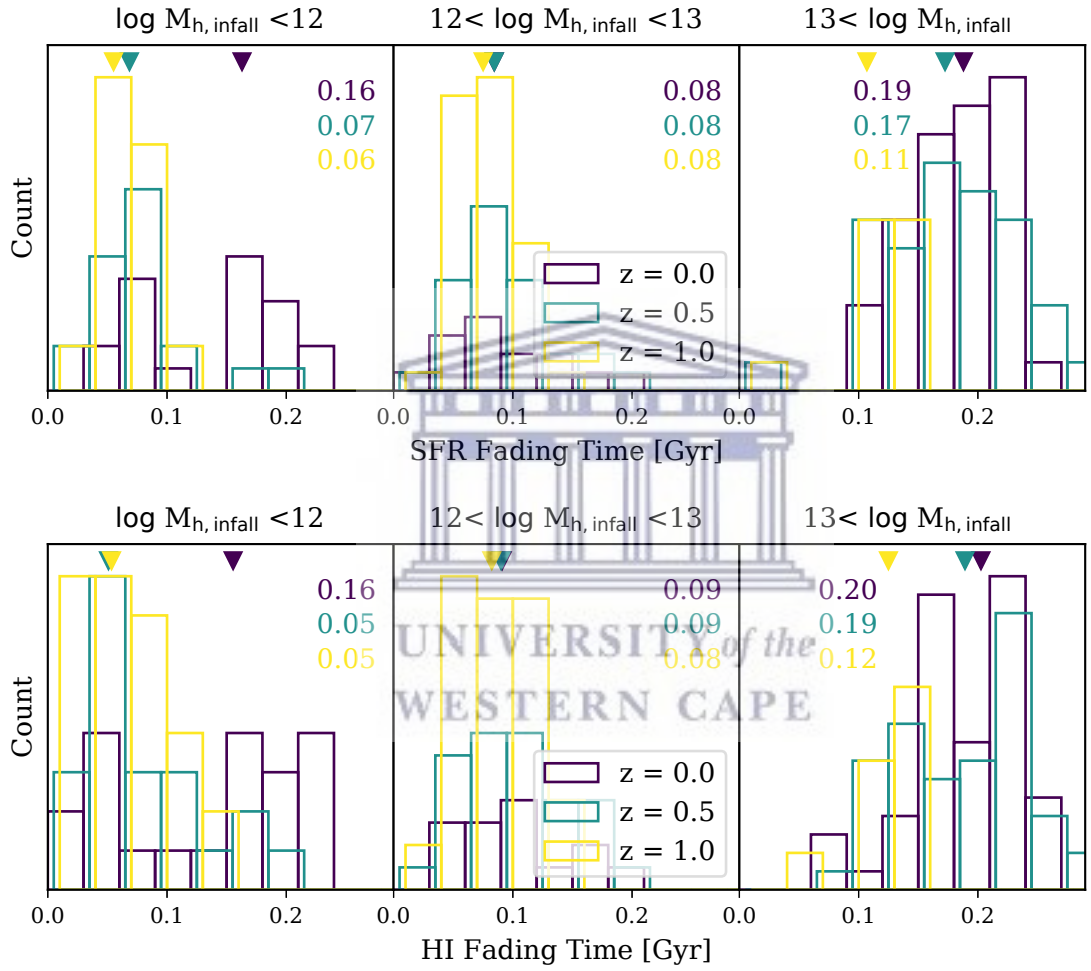


FIGURE 3.C: Similar to Figure 3.B but showing for the Fading Time τ_f.

CHAPTER 4

Predicting the Neutral Hydrogen Content of Galaxies From Optical Data Using Machine Learning

Abstract

We develop a machine learning-based framework to predict the HI content of galaxies from optical photometry and environmental parameters. We train the algorithm on $z = 0 - 2$ outputs from the MUFASA cosmological hydrodynamic simulation, which includes star formation, feedback, and a heuristic model to quench massive galaxies that yields a reasonable match to a range of survey data including HI. We employ a variety of machine learning methods (regressors), and quantify their performance using the slope of the predicted *vs.* true relation, its root mean square error (RMSE) and Pearson correlation coefficient (**r**). Training on only SDSS photometry, all regressors give **r**> 0.8 and RMSE$\sim$ 0.3 at $z = 0$, led by Random Forests with **r**= 0.91, and a Deep Neural Network (DNN) with comparable accuracy (**r**= 0.9). Adding near-IR photometry improves all regressors. All regressors perform worse with redshift, particularly at $z \gtrsim 1$. Slope values are generally sub-linear, so that we overpredict HI in HI-poor galaxies and underpredict HI-rich, because the regressors do not fully capture the scatter in the data. We test our framework on RESOLVE and ALFALFA survey data. Training on a subset of the observations, we find that our machine learning method can reasonably predict HI-richnesses in the remaining data (RMSE$\sim$ 0.28 for RESOLVE and $\sim$ 0.25 for ALFALFA). Training on mock data from MUFASA to predict observed data is worse (RMSE$\sim$ 0.45 for RESOLVE and 0.31 for ALFALFA), with DNN well outperforming other regressors. Our method will be useful for making galaxy-by-galaxy survey predictions and incompleteness corrections for upcoming HI 21cm surveys on SKA precursors such as MeerKAT, over regions where photometry is already available.

4.1 Introduction

One of the most important science goals of the Square Kilometre Array (SKA) project is to provide us with more insights into the growth and fueling of galaxies. A particular focus is on the evolution of their atomic neutral hydrogen, or HI content, which constitutes a major part of the gas content of galaxies, as traced by 21cm radio emission. HI gas represents the dense gas reservoir that will eventually form stars after passing through a molecular phase, and is thus a key and so far underexplored aspect of the baryon cycle governing galaxy evolution (Somerville & Davé, 2015). Hence upcoming surveys with SKA precursors MeerKAT and ASKAP aim to expand the depth and area of 21cm surveys out to $z \sim 1$, with the SKA potentially reaching even higher redshifts.

Much work has been done on studying the HI content of galaxies in the nearby Universe. The Arecibo Legacy Fast ALFA (ALFALFA; Giovanelli et al., 2005) blindly observed about 7000 deg^2 of the Arecibo sky and was completed in 2012. It has enabled a precise study of the distribution of galaxies in the local Universe based on their HI mass. For instance, Jones et al. (2016) studied the environmental effects on the HI content of galaxies using the Arecibo Legacy Fast ALFA survey $\alpha.70$ (70% of the final data). They found a shift of the Schechter function knee towards higher value in higher density environments. Due to ALFALFA's high positional accuracy of < 20 arcsec, they could explore the optical counterparts and extend the understanding of the stellar mass growth based on HI content. The GALEX Arecibo SDSS Survey (GASS; Catinella et al., 2010) used a complementary approach by selecting ~ 800 L^* galaxies from the Sloan Digital Sky Survey (SDSS; York et al., 2000) and observed their HI-line spectra until detection. Catinella et al. (2010) found that the *detected* (60% of the 20% observed) HI richness (M_{HI}/M_*) does not go below 40% even for the highest stellar masses explored ($\sim 10^{11}$ $M_\odot$). Using the full GASS dataset, Catinella et al. (2013) found an environment dependance of the gas fraction, such that galaxies in higher host halo masses have lower HI than those in less dense environments, confirming the idea that galaxy gas content and environment are tightly connected. The REsolved Spectroscopy Of a Local VolumE (RESOLVE; Kannappan et al., 2011) survey adopted yet another approach by observing ~ 1500 galaxies with ranges of stellar and gas masses within a volume-limited 53,000 Mpc^3 in the nearby Universe. Stark et al. (2016) used the RESOLVE data, targeting an

area within the SDSS redshift survey, and found that decreasing HI richness in galaxies is related to increasing host halo mass for a given stellar content. These data set the stage for explorations to lower masses and higher redshifts to be achieved with next-generation surveys.

Theoretical studies on the evolution of HI content of galaxies have also been expanding. Cunnama et al. (2014) predicted from the Galaxies-Intergalactic Medium Interaction Calculation (GIMIC) suites of hydrodynamical simulations (Crain et al., 2009), a tight dependence of galaxies' HI column density and environment: Galaxies in groups possess extended HI radial profiles compared to field galaxies. The extended radial profiles originate from the ram pressure redistribution which they found to dominate over the gravitational restoring forces. Although their findings are physically grounded, disentangling such processes remain a challenge for observers. Related results were found using a different galaxy formation model from Davé et al. (2013), where Rafieferantsoa et al. (2015) found a faster depletion of HI content once galaxies fall in a more massive haloes. The specific star formation rate of those galaxies also decreases but at rate less than that of the HI, indicating gas stripping from the outskirt of the galaxies inward. Quilis et al. (2017) studied the effects of ram pressure stripping. They used a cosmological simulated box to select a sample of galaxies residing in clusters to do their analysis. They found that galaxies below 10^{10} $M_{\odot}$ in stellar mass are often located at the outskirts of the clusters and have high eccentricity. Their interactions with the environment are more violent resulting in faster change of the gas contents and morphologies of the galaxies. More massive galaxies are situated closer to the cluster centers, and the gas removal is less dominant. The major change in those galaxies is caused by inflowing gas from the intercluster medium. Using the MUFASA data (Davé et al., 2016), Rafieferantsoa & Davé (2018) found a weak but extended galactic conformity in HI richness for galaxy members of low-mass haloes. Bigger host-halo galaxies tend to have stronger but less extended conformity. These studies demonstrate that the HI content of galaxies is impacted by their environment, but the exact nature of that dependence is not entirely clear.

Hence observational surveys suggest that understanding the baryon cycle requires precise measurements of the HI content of the galaxies, which at times might be affected by observational artifacts. Theoretical works, on the other

hand, predict physical results that are currently difficult to observe, which argues for larger and deeper HI surveys to improve our current understanding of the evolution of gas content and hence galaxy growth overall.

Although considerable efforts have gone into studying the gas phase properties of galaxies with the help of the currently available HI data, *e.g.* ALFALFA and RESOLVE, the understanding of HI evolution still lags behind the understanding of their stellar components. The main reason is that photometric data can be directly related to the stellar population of galaxies, and such optical and near-infrared data is currently technologically able to reach deeper levels than radio data. For the promise of multi-wavelengths surveys to be fully realised into the radio regime, it is important to be able to relate gas and stellar properties accurately. However, this is not straightforward. There have been attempts that have been proposed to estimate gas-phase properties of galaxies from their stellar masses obtained from spectral energy distributions (SED) fitting to photometrical properties. For instance, Kannappan (2004) found a correlation between $u-K$ colours and HI richness which they dubbed *photometric gas fractions*. The correlation was shown to be valid for galaxies with atomic gas fraction ranging from 1% to 10× the stellar masses. Zhang et al. (2009) developed a similar method by using the i-band surface brightness and the $g-r$ colour to estimate the HI richness of the galaxies. They found a tighter scatter compared to previous estimations. The HI scaling relations found by Zhang et al. (2009) were further improved upon by Wang et al. (2013) by introducing a form of correction to account for the fact that HI rich galaxies have more active star formation on the outer discs (bluer) (see Wang et al., 2011). Still with the standard approach by first establishing correlation between the gas fraction and other galaxy properties, Catinella et al. (2010) prescribed another relation $\log_{10}(\mathrm{M_{HI}/M_*}) = -0.332\log_{10}(\mu_*) - 0.240(\mathrm{NUV-r}) + 2.856$ which was also tested by Wang et al. (2015) with their samples to estimate the gas fraction as a function of stellar mass surface density (μ_*) and observed $NUV-r$ colour. Teimoorinia et al. (2017), in a work most similar to ours, used machine learning which was trained on ALFALFA data to predict the HI content of half a million SDSS-galaxies. In addition to direct photometric data, they considered 14 other derived galaxy properties as input parameters, and attained their best performance of only ~ 0.2 dex off the observed quantities. From these studies it is clear that developing ways to connect optical/NIR information with HI is an

important task, which affords many applications such as to estimate the HI content of certain galaxies based solely on their available photometry information, to enable larger statistics, and to assess incompleteness in surveys.

In this work, we explore a more general approach compared to previous studies by investigating the feasibility of predicting the HI richness of galaxies from directly observed optical properties of galaxies, particularly the photometric magnitudes and environmental quantities, using machine learning. The main idea is that machine learning can synthesise all the photometric data in order to optimally predict HI, rather than trying to isolate particular combinations that work best. The advantage of using machine learning techniques is mainly the capability of the model to learn peculiar aspects human might have overlooked, with the downside that such a method does not provide a direct physical interpretation of the result. The choice of only using directly observed quantities avoids introducing systematic uncertainties arising from the estimation of derived quantities such as stellar mass and star formation rate. For this paper we focus on galaxies having at least some HI content; future work will explore identifying which galaxies are gas-free.

Another key difference with respect to previous works is that we test the efficacy of using cosmological hydrodynamic simulations as a training set. The advantage of this is that, if our model is a sufficiently good representation of the real universe, then it can be used to explore regimes where HI data do not currently exist, such as those at higher redshifts. This provides a more physically-motivated prediction of HI content than using locally-calibrated relations. Furthermore, by using simulated galaxies to train and calibrate the method, connections can be made between the obtained correlations and the underlying physics, at least within the context of the given model.

Our best machine learning algorithms, random forest and deep learning, are able to predict the HI richness of simulated galaxies to within < 0.3 dex from their real values using only the photometric properties of the simulated galaxies. Testing this on RESOLVE and ALFALFA survey data, the prediction of the observed data from simulation-trained models yield less precise results. Generally, random forest is our optimal machine learning algorithm, but the neural network's performance becomes better when observational data are used.

Our method has numerous applications. Current data as well as future surveys will benefit from this method by providing ways to more accurately correct observations for incompleteness and confusion. For instance, the upcoming Looking At the Distant Universe with the MeerKAT Array (LADUMA; Holwerda et al., 2012) survey aims to directly detect and use different techniques to stack multiple objects to be able to measure HI fluxes out to $z \sim 1$ for the first time, to enable a deeper understanding of the fueling processes of galaxies and study the cosmic evolution of their HI content. But at higher redshift, confusion can become dominant especially when sources are located in groups. Meanwhile, ASKAP HI All-Sky Survey (WALLABY) which will cover two third of the sky will probe HI gas of 6×10^5 galaxies up to $z = 0.26$; DINGO, up to $z = 0.43$, will probe about 10^5 galaxies within 4×10^7 Mpc^3 cosmological volume (Duffy et al., 2012). These HI surveys will provide a wealth of information on galaxy evolution, but it is important to be able to accurately measure and understand the observations, which is where our method can provide insights.

§4.2 briefly reviews the MUFASA simulation used for this work. The approach we use in this study is detailed in §4.3, and we present the techniques utilized in order to achieve our goal in §4.4. §4.5 presents our findings and §4.6 shows a preliminary application of our method. We expand on the limitations of our method in §4.7 and finally conclude in §4.8.

4.2 Simulations

4.2.1 Galaxy formation models: MUFASA

For our training set we make use of the outputs of the MUFASA simulation model, which is fully described in Davé et al. (2016). We only present the key prescriptions in the model that are particularly relevant for this work.

MUFASA is implemented in the GIZMO cosmological hydrodynamics code, including a tree-particle-mesh gravity code based on GADGET (Springel, 2005), topped with a meshless finite mass hydrodynamic algorithm (Hopkins, 2015). The model uses radiative cooling and heating implemented with the GRACKLE

2.1 library[1]. Star formation follows a Schmidt (1959) law, based on a sub-grid prescription that determines the molecular gas content of each gas particles (Krumholz & Gnedin, 2011), and occurs only in gas elements above a hydrogen number density threshold of $n_H > 0.13\text{cm}^{-3}$.

MUFASA uses a kinetic gas outflow prescription to model star-formation driven winds, following scalings from the Feedback in Realistic Environments (FIRE; Muratov et al., 2015) zoom simulations. MUFASA also contains a heuristic prescription for star formation quenching whereby it heats the gas volume elements within a host halo that are above a halo mass threshold of $M_{\text{halo}} > (1 + 0.48z) \times 10^{12} M_{\odot}$ (Gabor & Davé, 2015; Mitra et al., 2015). This model is intended to mimic radio mode feedback from active galactic nuclei (Croton et al., 2006) in massive halos.

4.2.2 Galaxy sample

The galaxy sample used for our analysis is obtained by simulating a cube of $50h^{-1}$Mpc on a side with 512^3 dark matter particles and 512^3 gas volume elements. The initial conditions are generated at redshift $z = 249$ using MUSIC (Hahn & Abel, 2011) with Planck et al. (2016)-concordant cosmological parameters, namely $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, $\Omega_b = 0.048$, $H_0 = 68\ \text{km s}^{-1}\ \text{Mpc}^{-1}$, $\sigma_8 = 0.82$ and $n_s = 0.97$.

MUFASA evolves these initial conditions to $z = 0$, outputting 135 snapshots. For each snapshot, we identify galaxies, with SKID[2] as gravitationally bound collections of stars and star-forming gas. In our analysis, we will only use $z \leq 2$ sample, which, in total, is made of 50 snapshots. Each snapshot contain typically around 8000 resolved galaxies (> 64 star particle masses or $M_* > 1.16 \times 10^9 M_{\odot}$).

4.2.3 Galaxy properties

Our simulated galaxy properties are calculated with a modified version of CAESAR[2], which is an add-on package for the YT simulation analysis suite. The stellar mass of a galaxy, or M_*, is the total mass of the stellar particles within

[1] https://grackle.readthedocs.io/en/grackle-2.1/genindex.html
[2] https://bitbucket.org/laskalam/caesar

it. The atomic neutral hydrogen content, M_{HI}, of the galaxy is the summation of all HI from the gas particles. For each gas volume element, we account for the self-shielding from the metagalactic UV background radiation, by using a fitting formula for the effective optically-thin photoionization rate as a function of density (Rahmati et al., 2013). The unbound particles are also accounted and assigned to the respective galaxy with the closest mass weighted distance. The galaxy peculiar velocity v_{gal} is the 1-D mass-weighted average of all the particle velocities contained in it, along each of the (x, y, z) directions. We use the projected nearest neighbour density Σ_3 to quantify the galaxy environment such that:

$$\Sigma_3 = \frac{3}{\pi R_3^2} \tag{4.1}$$

where R_3 is the distance of the galaxy to its 3rd closest neighbour, projected along the line of sight.

The magnitudes of the galaxies are obtained using the Line Of Sight Extinction by Ray-tracing LOSER[3] (see Davé et al., 2017a, for a fuller description) package (not CAESAR) but still using the groups identified by SKID. We first use the Flexible Stellar Population Synthesis (FSPS; Conroy & Gunn, 2010) library to derive the stellar spectra of each star particle based on its age and metallicity, summing to obtain the stellar spectrum for that galaxy. Every stellar spectrum is attenuated by the line of sight dust extinction obtained by scaling the metal column density along the given line of sight; this results in each of 6 lines of sight $(\pm x, \pm y, \pm z)$ having different extinction and thus different spectra. We obtain all magnitudes by applying the appropriate filters. We computed (*u*,*g*,*r*,*i*,*z*) SDSS magnitudes, (*U*,*V*) Johnson magnitudes, *NUV* GALEX magnitude, and the (J,H,K_s) 2MASS magnitudes.

4.3 Machine Learning Setup

The goal is to predict the HI richness (M_{HI}/M_*) from other properties of a given galaxy. We use the supervised learning paradigm which consists of training the algorithm to estimate the desired label when fed with a corresponding input. Through a learning process, the best model parameters that minimize a defined cost function, which we choose to be the mean squared errors (MSE),

[3] https://bitbucket.org/romeeld/closer

are solved. Sets of training datasets drawn from our simulated sample are used to train our learners to predict the target (M_{HI}/M_*) from the features $\{u, g, r, i, z, U, V, J, H, K_s, \Sigma_3, v_{gal}\}$ of our galaxies.

It is noted that v_{gal} indicates line of sight velocity, and our models will predict the HI richness (M_{HI}/M_*) of the galaxies rather than their M_{HI} due to the less constrained correlation between the latter and the galaxy stellar masses. In addition, we take the logarithmic values of the target due to its large dynamic range which can cause the learning process to fail. First of its series, this work focuses only on the prediction of the HI richness of HI rich galaxies and to do so, we only select galaxies with $M_{HI}/M_* > 10^{-2}$, which decreases the size of our sample. To counteract, we increase our data by calculating the galaxy properties along all the 6 projections axis of the simulated cubical box, resulting in 6× more data for our analysis.

We assume we have all photometric magnitudes for all available bands, covering a wide range of spectrum including SDSS magnitudes, Johnson magnitudes and 2MASS magnitudes, which we can compute from our simulated galaxies. Although this scenario is ideal for our analysis, it is not so realistic. We can expect observed galaxies to only have $\{u,g,r,i,z\}$ magnitudes at best. In this regard, we examine different possibilities in our analysis. All the setups considered in this work are listed in Table 4.1, where `color indices` denotes all possible pairwise combination (*e.g.* $g - r$) of all the magnitudes in the surveys considered in one setup.

We train our model in two different ways. First is the "f-training", which considers all the galaxies from all the $z \leq 2$ outputs (with f leading the setup names, see first column of Table 4.1). Second is the "z-training", in which we build a regressor at each redshift bin (with z leading the setup names). In both approaches, we randomly choose 75% of the data as the training set and 25% as testing set. We do the training 10 times with 10 different random batches to get the uncertainty of our results: at each iteration, the dataset is randomly shuffled and new batches of training and test sets are generated.

To this end, we make use of 6 different machine learning techniques that we describe in the following.

TABLE 4.1: List of all the setups that are considered in the analysis. For easy reference, each setup has been given a name.

Name	Surveys	Features	Target	Description
fSMg	SDSS	u, g, r, i, z, v_{gal}, Σ_3	$\log(M_{HI}/M_*)$	redshift information not required
fSClr	SDSS	`color indices`, v_{gal}, Σ_3	$\log(M_{HI}/M_*)$	redshift information not required
fSCmb	SDSS	`color indices`, u, g, r, i, z, v_{gal}, Σ_3	$\log(M_{HI}/M_*)$	redshift information not required
fAMg	SDSS+Johnson+2MASS	H, J, Ks, U, V, u, g, r, i, z, v_{gal}, Σ_3	$\log(M_{HI}/M_*)$	redshift information not required
fAClr	SDSS+Johnson+2MASS	`color indices`, v_{gal}, Σ_3	$\log(M_{HI}/M_*)$	redshift information not required
zSMg	SDSS	u, g, r, i, z, v_{gal}, Σ_3	$\log(M_{HI}/M_*)$	prediction at a given redshift bin
zSClr	SDSS	`color indices`, v_{gal}, Σ_3	$\log(M_{HI}/M_*)$	prediction at a given redshift bin
zSCmb	SDSS	`color indices`, u, g, r, i, z, v_{gal}, Σ_3	$\log(M_{HI}/M_*)$	prediction at a given redshift bin
zAMg	SDSS+Johnson+2MASS	H, J, Ks, U, V, u, r, r, i, z, v_{gal}, Σ_3	$\log(M_{HI}/M_*)$	prediction at a given redshift bin
zAClr	SDSS+Johnson+2MASS	`color indices`, v_{gal}, Σ_3	$\log(M_{HI}/M_*)$	prediction at a given redshift bin

4.4 Machine Learning Algorithms

We use `TensorFlow` to build the DNN model and `scikit-learn` (Pedregosa et al., 2011) package for the remaining methods.

4.4.1 Linear regression (LR)

Linear regression model (along with kNN, see §4.4.3) is the simplest amongst those we use in this work. Its simplicity, hence its great speed during training, provides quick insights into the relationship between the features ($\mathbf{x}$) and the corresponding target (y). The latter is defined as a linear combination of all the features, $y = \mathbf{w} \cdot \mathbf{x}$, and the idea consists of finding the weights $\mathbf{w}$ that minimize the mean squared error (MSE)

$$\text{MSE} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{w} \cdot \mathbf{x}_n - y_n)^2. \tag{4.2}$$

Here the bias is absorbed into the weights $\mathbf{w}$.

4.4.2 Ensemble learning methods: Random forest (RF) and Gradient Boosting (GRAD)

To understand both RF and GRAD algorithms one needs to first look at their base estimators, the Decision trees (Hastie et al., 2009), which will be clarified below.

In a simple one dimensional problem, we assume a dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ of length N ($(x, y) \in \mathbb{R} \times \mathbb{R}$). The first step of the algorithm is to split the training set at a split point s that minimizes the cost function

$$J = \min_{c_1} \left\{ \sum_{x_i \in R_1(s)} (y_i - c_1)^2 \right\} + \min_{c_2} \left\{ \sum_{x_i \in R_2(s)} (y_i - c_2)^2 \right\}, \tag{4.3}$$

where $R_1 = \{x_i | x_i \leq s\}$ and $R_2 = \{x_i | x_i > s\}$ are the two regions (also called nodes) resulting from the split. The values c_1 and c_2 that minimize each term in Eq. 4.3 are simply the averages of the labels y_i in R_1 and R_2 respectively; *i.e.*

$$c_1 = \frac{1}{m_1} \sum_{x_i \in R_1(s)} y_i,$$
$$c_2 = \frac{1}{m_2} \sum_{x_i \in R_2(s)} y_i, \tag{4.4}$$

where m_1 and m_2 are the number of inputs x_i found in R_1 and R_2 respectively. To grow the tree, each resulting node from the root is further split recursively (known as greedy algorithm) until a fixed maximum depth (or size) of the tree is reached. The nodes at the bottom of the tree are called the leaf nodes. To predict a new label y_{new} from a new input x_{new}, one simply walks through the tree from the root to reach a leaf node which then estimates y_{new} by averaging the corresponding labels y_i of the inputs x_i whithin it according to

$$\hat{y}_{\text{new}} = \frac{1}{m} \sum_{x_i \in \mathcal{L}} y_i, \tag{4.5}$$

where $\mathcal{L}$ indicates the leaf node and m the number of points x_i within it. Decision trees are prone to overfitting but there exist various techniques of regularization.

Random forest (Breiman, 2001), known to be a powerful machine learning algorithm, is composed of a given number (among the hyper-parameters of the model) of decision trees (base estimators) which are individually trained with a random subset of the dataset. To do a prediction, RF simply averages the predictions of its decision trees.

Another well known ensemble learning model that we use is gradient boosting (Friedman, 2000). Its base learner is also a decision tree but instead of simply aggregating the predictions of its regressors like in the case of RF, the training is carried out in a sequence. Except for the first regressor, which is trained with the dataset, each next regressor in the sequence – set by the number of the base estimators – fits the residual errors of its predecessor and so on. The resulting estimator, is then of the following form

$$\mathcal{E}(x) = \mathcal{E}_1(x) + \sum_{i=2}^{N} \gamma_i e_i(\epsilon_i), \tag{4.6}$$

where $\mathcal{E}_1(x)$ is the first estimator, ϵ_i the residual errors from the $i-1^{\text{th}}$ learner used as inputs of the i^{th} learner to fit a predictor e_i and γ_i is a coupling parameter which is optimized such that the error from the combined system at each iteration (*i.e.* $\mathcal{E}_{i+1}(x) = \mathcal{E}_1(x) + \sum_{k=2}^{i} \gamma_k e_k(\epsilon_k)$) is minimized. N is the number of base regressors (equal to the number of iteration) that form the ensemble.

4.4.3 k-Nearest Neighbor (kNN)

k-Nearest Neighbour (Altman, 1992) is a flexible non-parametric regression algorithm. Considering a set of instances $\mathbf{x}_n$ (in general $\mathbf{x}_n \in \mathbb{R}^d$ but for the sake of simplicity we let $\mathbf{x}_n \in \mathbb{R}$) with their corresponding label y_n ($y_n \in \mathbb{R}$), to predict a new label y_{new} given a new instance $\mathbf{x}_{\text{new}}$, the estimate of y_{new} is simply the weighted average of targets of the $k-$closest neighbours of $\mathbf{x}_{\text{new}}$. The principle is generalised for $d-$dimensions in feature space.

4.4.4 Support Vector Machine (SVM)

Given a set of training data consisting of examples $\mathbf{x}_n$ ($\mathbf{x}_n \in \mathbb{R}^d$) and their labels y_n ($y_n \in \mathbb{R}$), the method aims at finding a linear function of the form $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$. This can be seen as a convex optimization which seeks to

- minimize $\frac{1}{2}\mathbf{w}^T\mathbf{w}$,
  subject to the constraint $|y_n - (\mathbf{w} \cdot \mathbf{x}_n + b)| \leq \epsilon$,

where ϵ denotes the residuals between estimates and the desired outputs. To deal with otherwise intractable optimization problem, Vapnik (1995) introduced some slack variables ξ_n^-, ξ_n^+ such that it now aims at

- minimizing $\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}(\xi_n^- + \xi_n^+)$
  subject to

$$\text{the constraints} \begin{cases} y_n - (\mathbf{w} \cdot \mathbf{x}_n + b) \leq \epsilon + \xi_n^- \\ \mathbf{w} \cdot \mathbf{x}_n + b - y_n \leq \epsilon + \xi_n^+ \\ \xi_n^-, \xi_n^+ \geq 0 \end{cases} \tag{4.7}$$

where C is a positive value used for regularization. For simplicity, we only present the linear case but to deal with non-linearities one can resort to a kernelized SVM. It is noted that SVM method is also used for classification problem (Cortes & Vapnik, 1995).

4.4.5 Artificial neural network

We dedicate this section for a rather extended description of the deep neural network used for this work. This is so due to its novel application in astronomy. This is not so much the case with other machine learning techniques described before, as they are at some point fully or partly used to analyze astronomical data.

Due to our hardly correlated features and target, the choice of model to learn the connection between them is very complex, though our maximum number of galaxy properties are limited to only 12 components. Figure 4.1 shows a summary of our multilayer perceptron model. The left nodes show our galaxies properties as input into our 3 hidden layers and the right most node is the output. $\overset{j}{y}_k$ represents the k^{th} neuron in the j^{th} layer and is the linear weighted sum of the preceding neurons as shown in equation 4.8, f_a being the activation function (see 4.4.5.2).

$$\overset{j}{y}_k = f_a\left(\sum_l w^j_{k,l} \times \overset{j-1}{y}_l + b^j_k\right) \tag{4.8}$$

$w^j_{k,l}$ and b^j_k are the weight and bias of $\overset{j-1}{y}_l$ on $\overset{j}{y}_l$.

A deep neural network (DNN) is then to learn the (close to the) correct values of w's and b's for the model to be able to reproduce the *target* given the *features*.

The choices for the number of the hidden layers, the activation functions between layers and the optimiser are described in the following subsections.

4.4.5.1 Hidden layers

One of the toughest step that one has to overcome in building a DNN model is the choice of the number of hidden layers and the respective number of neurons in each layer. The use of models with a single hidden layer or the so called

universal approximators has been advocated since the artificial neural network was used into solving physical problems. Cybenko (1989) stated that a single hidden layer in a feedforward – any connections between neurons do no form a cycle – neural network is enough to capture the continuous non-linearity between the inputs and the outputs. This conclusion was extended later on by Hornik (1991) that the nature of the feedforward structure drives its universality irrespective to the activation function as long as the latter is continuous, bounded and non-constant (see. §4.4.5.2). The *"universal approximation"* principle ended recently after the work done by Hinton et al. (2006). They explored the improvement of the multi-hidden-layer architecture and concluded the following. Although a single hidden layer with finite number of neurons can be enough to map the connection between the input(s) and the output(s), one extra layer is useful to increase the accuracy of the mapping. Any additional layer is only for the model to explore possible representations of the map and to decrease the learning time given a set of data.

For those reasons, and after a trial-and-error approach, we opt to use 3 hidden layers in our model. We use 100 neurons in each layer to correctly map the galaxy properties with all their possible combinations. We have extra nodes to account for some degrees of freedom for safety. Using different (simpler) configurations end up with similar results for some of the proposed setups (zSCmb for instance, see Table 4.1), but our choice for more complex network is driven by the need for more stable algorithm.

4.4.5.2 Activation function

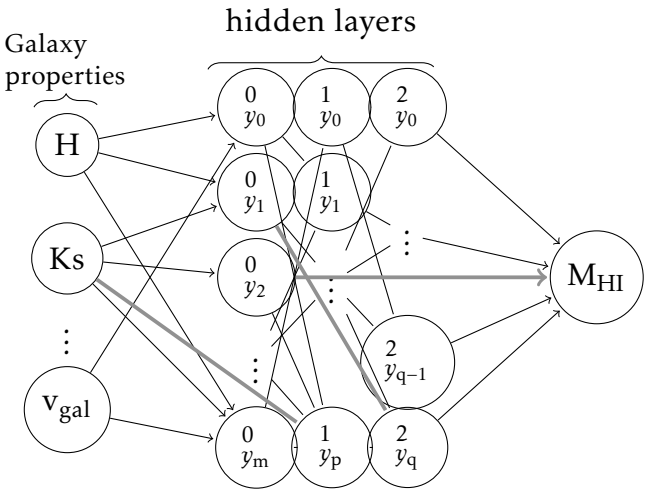


FIGURE 4.1: Network graph of our 4-layer perceptron with 1 output unit. The hidden layers contain m, p, q neurons respectively.

Given a set of values fed to one node in our model (see Figure 4.1), one has to decide how much of that information should be passed to the next connected node(s). This can be defined with an activation function. A sigmoid function was widely used in the past. Problems occur with that function when the input values of a node are high (or small in the negative end): that is the vanishingly small gradient at those ends. In our model, we use a rectified linear unit function (RelU, see eq. 4.9). It means that any negative values passing the nodes are set to zero (ignored).

$$f(x) = max(0, x) \tag{4.9}$$

We also tested the use of an exponential linear unit function (elU, see eq. 4.10). In this case, we allow a small fraction of the negative signal to go through the next connected node(s).

$$f(x) = \begin{cases} x, & \text{if } x \geq 0. \\ \exp\{(x)\} - 1, & \text{otherwise.} \end{cases} \tag{4.10}$$

Our test didn't result in any improvement (if not deterioration) in using eLU. Using different activation functions such as *hyperbolic tangent*, *gaussian* or *multiquadratics* are not favoured in our case.

4.4.5.3 Optimisation

After each step of calculations, the network should optimize the model based on its current and previous states to improve the subsequent mapping. Our model utilizes a computationally memory efficient optimization due to its dependancy to only the first order gradients, namely the "*adaptive moment estimation*" (or Adam). For more details we refer the readers to Kingma & Ba (2014). Adam optimization, compared to other gradient-based optimization, is very suitable for noisy and sparse gradients, and for simulated data which show very large scatter with respect to a given quantity of parameter (Kingma & Ba, 2014). With this optimizer, we have to decide few parameters in advance. The learning step α and the parameters controlling the moving averages of the 1st and 2nd order moments namely β_1 and β_2 (both $\in[0,1)$) respetively. For this purpose, we chose to minimize the mean squared error between the target and the prediction from the model: in what follows, we will alternatively call the mean squared error the "*objective function*" $f(\mathbf{x})$: with $\mathbf{x}$ the parameters of the model to be updated, such as weights and biases. At a given time $t \leq T$, where T is the maximal learning time step, we can update the parameters of the model as shown in the following.

$$g_t = \nabla_x f(\mathbf{x}_{t-1}) \tag{4.11}$$

$$\mu_{1,t} = \beta_1 \times \mu_{1,t-1} + (1-\beta_1) \times g_t \tag{4.12}$$

$$\bar{\mu}_{1,t} = \mu_{1,t}/(1-\beta_1^t) \tag{4.13}$$

$$\mu_{2,t} = \beta_2 \times \mu_{2,t-1} + (1-\beta_2) \times g_t^2 \tag{4.14}$$

$$\bar{\mu}_{2,t} = \mu_{2,t}/(1-\beta_2^t) \tag{4.15}$$

$$\mathbf{x}_t = \mathbf{x}_{t-1} - \alpha_t \times \bar{\mu}_{1,t}/(\sqrt{\bar{\mu}_{2,t}} + \epsilon) \tag{4.16}$$

where $\alpha_t = \alpha\sqrt{1-\beta_2^t}/(1-\beta_1^t)$ is the time-step at t. Equation 4.11 shows the gradients of the objective function at t with respect to the model parameters. Equations 4.12,4.14 update the estimations of the 1st and 2nd moments. Our moments are biased towards the initial values, thus we require equations 4.13,4.15 to account for the corrections. Finally, we update the model parameters with equation 4.16.

We do not claim that the choice of parameters implemented in our models as well as their configurations are the best to do similar work. We will likely continue to improve this method in subsequent papers.

The reasons we opted for such diversity of regressors in this study are as follows.

- to explore the linearity in the data by using LR, SVM,
- to explore the power of the ensemble learning with RF and GRAD,
- to explore the simplicity, versatility and speed of *k*NN,
- to explore the power of the more sophisticated DNN.

4.5 HI Prediction Using Machine Learning

Our goal is to predict the HI richness of a given galaxy based on its optical/near-IR photometry. We choose to predict HI richness and not HI mass as it is expected to correlate more with galaxy colours, with HI-poor galaxies being redder than HI-rich ones, so in some sense gives more physical information than just HI mass alone which approximately correlates with stellar mass. Nonetheless, our approach could equivalently be used for either, and we have tested that the resulting accuracy of the predictions is similar.

4.5.1 Quantifying the mapping accuracy

For a given trained model (see §4.3), we can predict the HI richness of a test set which contains the feature parameters, similar to those used during the training, and the real HI richness. One can then see for a given example (composed by the features) how the model estimates the corresponding HI richness and compare the predicted value of this latter with its real value.

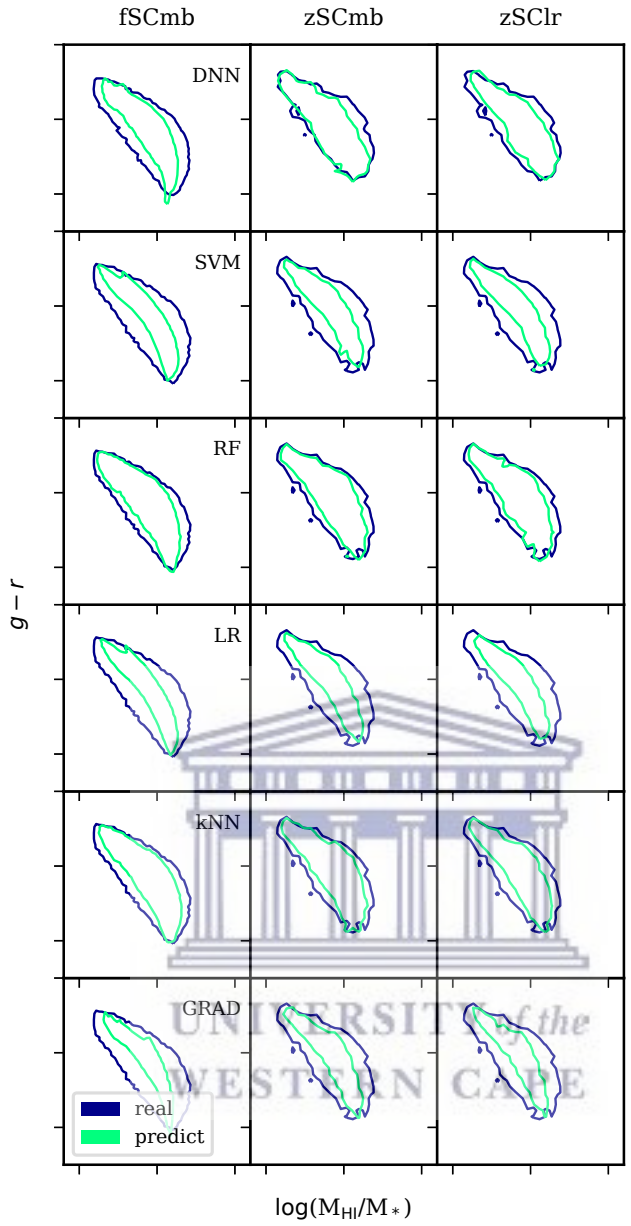


FIGURE 4.2: Superposition of the predicted (green) and the real (blue) HI richness of our galaxies (x-axes) *vs.* $g-r$ colour (y-axes). The contours are enclosing 2σ of the distributions. Each row shows different mapping corresponding a particular method and each column a different setup (see Table 4.1).

Figure 4.2 shows the galaxies' M_{HI}/M_*vs. a selected colour $g-r$, one of our input features. The simulated targets are shown with the blue contours and the predicted values with the green contours. Each column represents 3 selected setups (see Table 4.1) that only use SDSS magnitudes during the training whereas each row corresponds to one training model. The z-trained models shown here (two right columns: zSCmb, zSClr) are at $z = 0$.

Overall, the ML-predicted values follow the true values from the simulation, and show that galaxy colour is anti-correlated with M_{HI}/M_* as expected. The mean trend is always well recovered using any of the predictors. However, the scatter in the data is not fully captured by any of the models: The green contours are always inside the blue contours. Different ML algorithms perform differently in this regard: We see that for DNN, RF & kNN, the two contours are quite close. Only looking at the f-trained models (left column) where we train on all the data from $z = 0 - 2$ simultaneously, it is evident visually that RF maps $g - r$ best, kNN comes next followed by DNN. For the z-trained models where we train individually at various redshifts, DNN, RF & kNN do similarly well with zSCmb but the performance of RF is better with zSClr (where we add in the color indices). In contrast, SVM, LR and GRAD have difficulty to capture the scatter in the data, hence their predictions tend to be more tightly confined around the mean. While we have shown this specifically for $g - r$, the results for other colours are similar, and typically show that RF and DNN perform the best, with kNN not far behind.

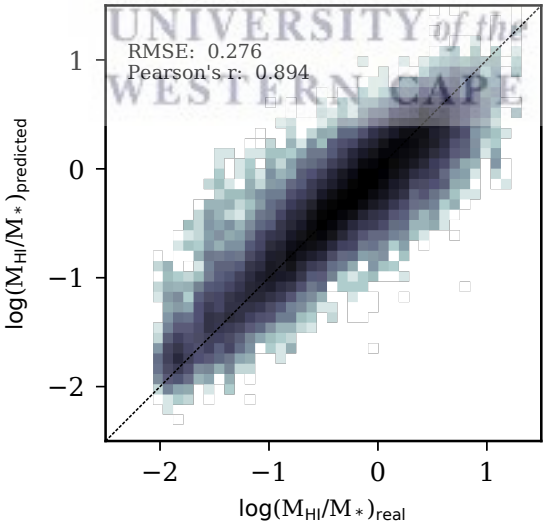


FIGURE 4.3: 2D distribution of the real (x-axis) *vs.* predicted (y-axis) HI richness with the $z = 0$-trained DNN model, using the zSCmb training set.

Figure 4.3 shows a direct comparison between the real and the predicted HI richness of the galaxies with the DNN models trained and tested with $z = 0$ simulated data. The dashed line shows the 1:1 line; if the ML algorithm were perfect, all points would lie along this line. The correlation is apparent and generally follows the identity line, indicating that the training performs reasonably well in the mean. However, there is a significant scatter, which degrades the performance on a galaxy-by-galaxy basis. The best-fit slope is also not identically unity, so the correlation is not perfect even in the mean. We thus would like to quantify our regressors' performance using the slope and tightness of the correlation.

To quantify the performance of our ML framework, we choose three metrics:

- The slope of the linear mapping $f : y \rightarrow \hat{y}$, where an ideal mapping would have a unity slope.
- RMSE (Root Mean Squared Error), given by

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

  where y and $\hat{y}$ are the real value and the estimate respectively, gives the average difference between the predicted and the real values. The square of this metric is also used as a cost function to be minimized in some methods for regression (*e.g.* deep neural network, linear regression). The lower the RMSE the better the performance of the model is.

- Pearson product-moment correlation coefficient (Pearson's **r**) which tells how scattered the predictions are compared to the true values. The closer to 1, the tighter (or better) the prediction is.

$$\text{Pearson's r} = \frac{\sum_{i=1}^{N}(y_i - Y)(\hat{y}_i - \hat{Y})}{\sqrt{\sum_{i=1}^{N}(y_i - Y)^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \hat{Y})^2}}$$

  where Y and $\hat{Y}$ are the mean values of y_i and $\hat{y}_i$ respectively.

In figure 4.3, we get RMSE= 0.276 and Pearson's **r**= 0.894 for the particular choice of the DNN regressor and the zSCmb training set; this is one of our best

cases, but RF is actually slightly better. Previous work by Zhang et al. (2009), estimating HI-to-stellar mass ratio using analytic equation leads to 1σ scatter > 0.3, which shows that our ML approach is more accurate.

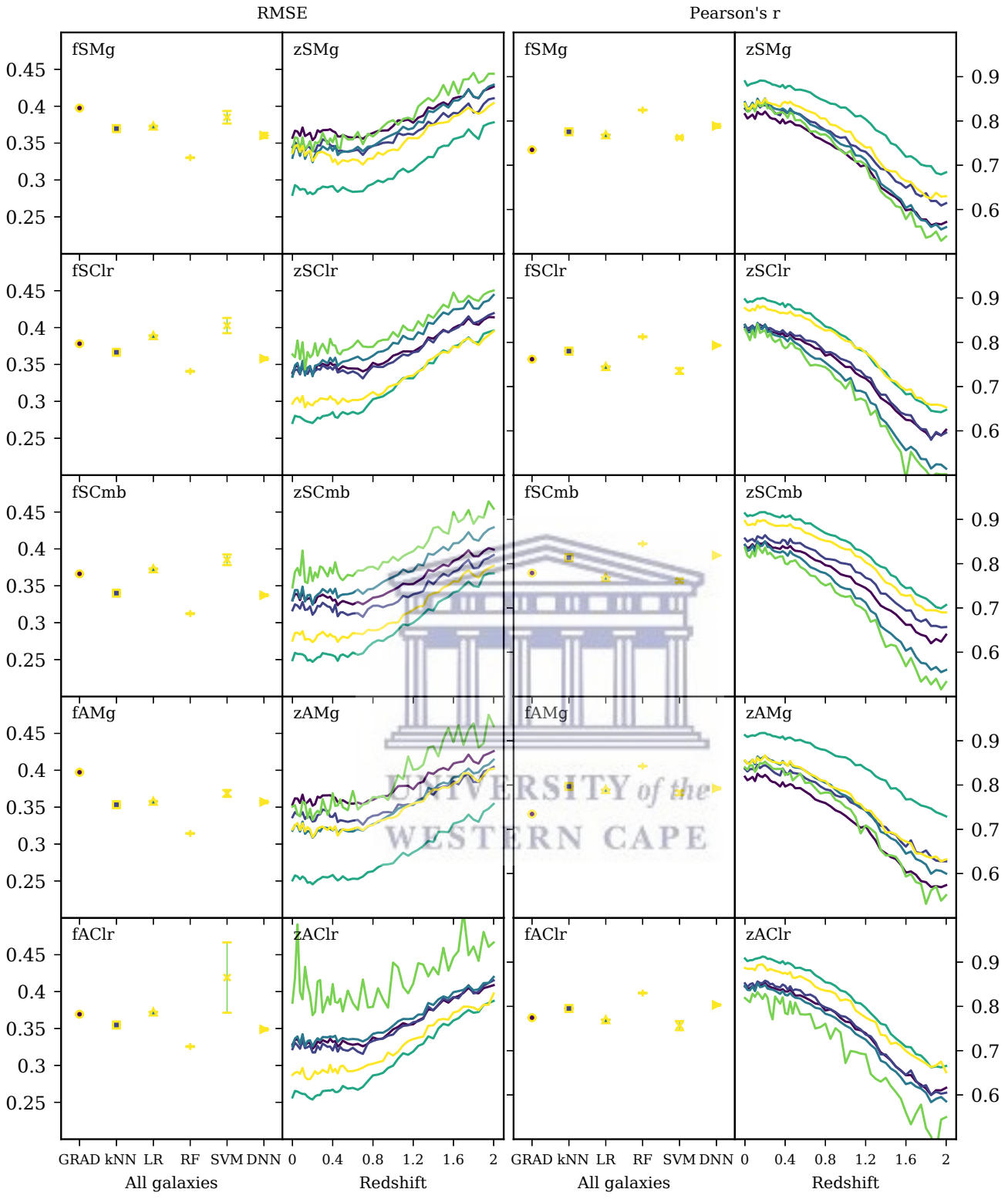


FIGURE 4.4: Root mean square errors RMSE and Pearson product-moment correlation coefficients **r** are shown on the 2 columns from the left and right, respectively. Models perform better if they show lower RMSE and higher **r**. The first on the left shows a mapping for all the galaxies, and the second for galaxies at different redshifts. The dots and lines are color coded by the training models we use. Each rows show different results for different setups. The RMSE values are shown on the left y-axes and the **r** values on the right y-axes.

Figure 4.4 shows the performance of the various models considering each setup in Table 4.1 using RMSE and Pearson's **r** coefficient. The 2 columns from the left are the RMSE and the 2 columns from the right are Pearson's **r**. Each row corresponds to the results from different features used in the training. The name of the setup is shown on the top left of each panel. Different results from different learning techniques are presented with the color coded lines (with distinctive markers). In the following subsections we discuss how well our various regressors perform when varying the training set and the training method.

4.5.2 Dependence on redshift

Examining the leftmost column in Figure 4.4, these are the RMSE's for various ML algorithms when training on the entire data set from $z = 0 - 2$ without any redshift information (f-training). The results bear out the trends noted in Figure 4.2: The RF method generally does the best (lowest RMSE) for any of the input data sets, while DNN and kNN follow, and then the remaining methods. The RF values are still typically above 0.3, with the lowest values for the fSCmb (SDSS colours, magnitudes, and environment) and perhaps marginal improvement in fAMg which adds the near-IR photometry.

The third column shows the corresponding Pearson's **r** values. The basic story is the same, that RF provides the best prediction, with values of $\mathbf{r} \approx 0.85$ in the best cases, with others down to $\mathbf{r} \approx 0.75$. The predictions from the aggregate dataset clearly contain significant information, but are perhaps not as optimal as one might get from including some redshift information.

The second and fourth columns show the result of training and testing at individual redshifts (z-training). It is clear that from $z \sim 0 - 0.5$, the z-training performs better than the aggregate (f) training, with lower RMSE around 0.25 in the best-case RF models (zSCmb and zAMg). The other ML algorithms are clearly poorer than RF, although DNN does reasonably well in the zSCmb case. Similarly, the fourth column showing the Pearson's **r** also is very good at $z = 0 - 0.5$, and here DNN in many cases does nearly as well as RF.

Beyond $z > 0.5$, all the regressors show degrading performance, with increasing RMSE and decreasing **r**. This increase in RMSE likely owes to the fact that at high-z, all galaxies are more HI rich ($M_{HI}/M_* > 10^{-2}$) Rafieferantsoa et al. (2015), with

fewer and fewer quenched galaxies with very low M_{HI}/M_*. Because the intrinsic M_{HI}/M_* vs. mass (and other properties) thus becomes fairly flat, it becomes increasingly difficult for the ML to pick out the correct M_{HI}/M_* based on other galaxy properties as would be reflected in the photometry. This is likely an intrinsic limitation of this method, owing to the evolution of HI in galaxies.

Redshift information can be obtained observationally, amongst other methods, from photometry or spectroscopy. The latter is still easier to retrieve than direct HI data, while the former typically obtains redshift errors of a few percent, which is still good enough to ascribe a training redshift. It is clear from the above results that redshift information is useful to improve the predictions. Even out to $z \sim 1$, the limit of currently planned surveys, the predictions do not degrade greatly, it is only at $z > 1$ that they become worse than the aggregate case. Hence from here on we will primarily discuss the z-training results.

4.5.3 Dependence on input features

The different rows in Fig. 4.4 show the impact of varying the input features into the ML framework. As we have seen, RF generally performs the best followed by DNN. GRAD, kNN, LR and SVM perform similarly poorly regardless of our setups (their RMSE's$\simeq$ 0.34), with perhaps GRAD performing the worst. For this reason, unless otherwise stated, we are only going to discuss RF and DNN in what follows.

At $z = 0$, using only SDSS magnitudes results in relatively poor performance, with RMSE $\approx$ 0.3 for RF and 0.35 for DNN and others. For RF, using either `color indices` instead of magnitudes (zSCls) or in addition to magnitudes (zSCmb), or including additional magnitudes into the near-IR (zAMg) improves this significantly, with RMSE as low as 0.25 and $\mathbf{r}$> 0.9. Thus it appears that providing colour information directly into the ML algorithm helps it determine a better mapping than only providing the magnitudes, even though in principle the magnitudes contain all the colour information. Also, providing additional near-IR bands seems to be advantageous.

For DNN, the story is slightly different. Again, only SDSS bands has the worst performance, but here, including the near-IR data does not improve things as much as providing `color indices`, and particularly providing both `color indices`

and magnitudes together (zSCmb), which achieves a performance approaching that of RF.

The redshift dependence of RMSE and **r** is similar among all these combinations of input datasets. The overall message is that providing more bands is better, which is unsurprising, but also that it is preferable to provide the colours directly rather than the magnitudes given the choice. In many cases, it is possible via SED fitting to obtain a galaxy colour that has uncertainties that are smaller than would be obtained by just subtracting magnitudes, so this may be a more valuable input for ML predictions.

4.5.4 The slope of the mean relation

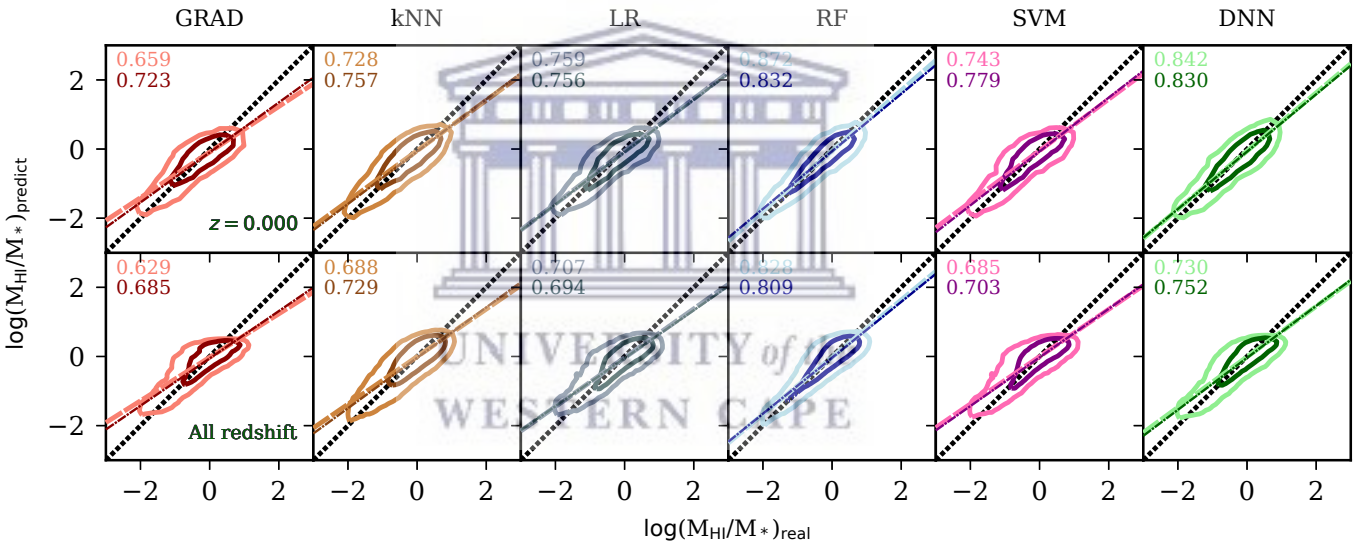


FIGURE 4.5: 2D representations of our real (x-axes) *vs.* predicted (y-axes) values of HI richness. Upper panels show for different models at $z = 0.0$, whereas the lower panels show for all redshift combined. We only show the results from our {f,z}SCmb features. The numbers with dark (light) colors on the left top corners show the slopes of the linear fit of the 1σ (2σ) subsample.

In Figure 4.5 we show linear fittings for the correlation between real (x-axis) and predicted (y-axis) values for M_{HI}/M_*. The top pannels are for the z-trained models at $z = 0$ and the lower panels for f-trained models. Each column corresponds to a given regressor as labeled on top. In each panel, the dark (light) lines represents the 1σ (2σ) contours between the targets and the predictions. The numbers on the top right are the slopes of the linear fits (color coded) for

the two contours. The thick dashed line shows the 1-to-1 relation, which would be the perfect prediction. We only show the SDSS combined setup (zSCmb) here, *i.e.* the features are SDSS magnitudes+`color indices` $+v_{gal}+\Sigma_3$, but the results from other setups are similar.

We can see that f-trained (lower panels) models tend to have slopes further away from unity compared to those from the z-trained ones. This confirms what we found previously with RMSE and Pearson's **r**, that at low redshifts, training on the smaller but more homogeneous sample at a given z provides a better prediction than training on a larger sample that conflates all the redshifts.

Among regressors, again we see that RF and DNN have slopes that are closest to unity, and thus perform better. All other methods have best-fit slopes below 0.8. However, all the slopes are < 1, which indicates an under-prediction of the HI richness for HI rich galaxies and over-prediction for HI-poor galaxies. This reflects the fact that, as seen in Figure 4.2, the true scatter in the M_{HI}/M_* around the mean is not fully reproduced in the predictions, such that all the regressors tend to fit galaxies closer to the mean. Hence at the lowest M_{HI}/M_*, they tend to fit slightly higher values, while at the highest M_{HI}/M_*, they tend to fit slightly lower values, resulting in a sub-unity slope: akin to an Eddington bias. The slope thus partly reflects a measure of how well the scatter around the mean is predicted. The fact that RF and DNN have the best slopes just quantifies the qualitative impression from Figure 4.2 that these regressors reproduce the extent of the scatter most closely.

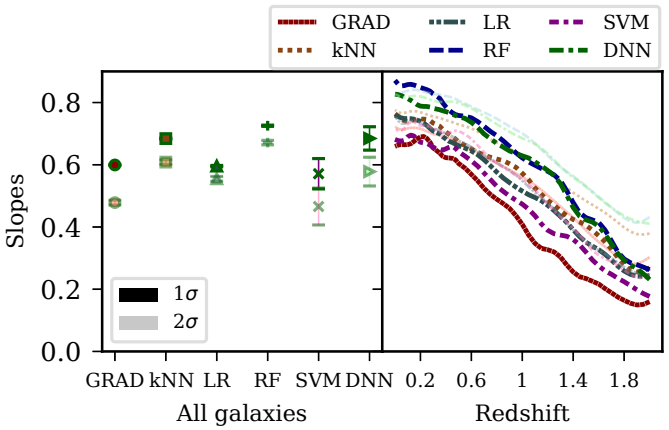


FIGURE 4.6: Slopes of the linear fit (y-axes) of the relationship between the predictions and the real HI richness of our simulated galaxies. The dark color (or thick lines) show the fit for the 1σ sample around the maximum and the light color (or thin lines) for 2σ. The left (x-axis showing the names of the models) is similar to what is shown in Figure 4.5 second row, the right (x-axis showing the redshift values) panel presents the evolution of slopes from our zSCmb features.

Figure 4.6 shows the comparison of slope values for the f-trained sample (left panel) and the redshift evolution of the z-trained sample (right panel) among the various regressors. The left panel effectively just shows a plot depicting the numbers in the bottom row of Figure 4.5. Here, RF performs the best but not so far from DNN (considering the variance among 10 subsamples), and the other models perform somewhat worse.

The right panel extend the values shown in the upper panel of Figure 4.5 to higher redshift. Dark colors (or/and thick lines) show the 1σ slopes and the light colors (or/and thiner lines) show the 2σ slopes. Looking at the z-training results (right panel), it is very clear that the slopes of RF and DNN are closer to unity than the other models, and that is true across all redshifts. The 2σ slopes (light color lines) are generally better than the 1σ's, except at the lowest redshifts. Slopes < 0.5 implies a weak correlation between the predicted and the real values of HI richness, so Figure 4.6 indicates that all regressors become unreliable beyond $z \gtrsim 1$.

In summary, k-NN, RF and DNN methods show better performance as compared to SVM, GRAD and LR (Figure 4.2). DNN and RF tend to perform better when providing galaxy colours as opposed to photometry, and when providing more bands. Among our tests, the best mapping of HI richness was achieved with RF at $z = 0$ using optical and near-IR bands, which gave RMSE's ≈ 0.25 and

r > 0.9. Using all data from $z = 0 - 2$ together did not provide as a good fit as training at individual redshifts, despite the smaller samples for the latter. The evolution of RMSE or Pearson's **r** shows a stronger redshift dependence beyond $z \sim 0.5 - 1$ making the prediction uncertain at higher redshift ($z > 1$, see Figure 4.4). Slopes of linear fits are generally less than unity owing to the fact that the true scatter is not fully spanned by the prediction; again, RF performs the best with DNN close behind, and the other regressors significantly poorer. All slopes move further from unity with increasing redshift, once again limiting applicability at $z \gtrsim 1$.

4.6 Application to observed data

We now apply and test our ML methodology against real observations from the RESOLVE and the ALFALFA data. RESOLVE survey provides both photometry and M_{HI}/M_*, so provides an ideal sample to test the efficacy of our predictions. There are two ways we will test this: First, we will train on the RESOLVE data itself, and predict the RESOLVE data, to test how well it works in the ideal circumstance of having the training and testing set be from the same sample. Second, we will train on the simulation and predict the RESOLVE data, which is more like the application envisioned for this technique, to see how much degradation there is when the training and testing sets are different. If the simulation was a (statistically) perfect representation of the RESOLVE data, we would expect the resulting RMSE and **r** to be similar, but given that we expect some differences, we aim to quantify the degredation in a real-world situation. We concurrently apply the above procedure to the ALFALFA data, in order to test sensitivity to different input data sets having different systematics.

4.6.1 Simulated *vs* observed data

4.6.1.1 RESOLVE data

We first describe the RESOLVE data. We make use of the photometry data (Eckert et al., 2015) as well as their corresponding HI-flux (Stark et al., 2016) from the Data Release II of the RESOLVE survey. We use the following standard

equation

$$M_{\rm HI} = 2.36 \times 10^5 \times D^2 \times F_{\rm Total} \tag{4.17}$$

to compute the HI mass in $M_\odot$, where D is the distance to the galaxy (Mpc) calculated from the apparent and absolute magnitude in **r** band given in the photometry data. $F_{\rm Total}$, provided by the RESOLVE data, is the total HI line flux ($Jy.\ {\rm km\,s^{-1}}$) of the galaxy. The RESOLVE photometric data release[4] contains SDSS (*u*,*g*,*r*,*i*,*z*), 2MASS (*J*,*H*,*K*), GALEX (*NUV*) and UKIDSS (*Y*,*H*,*K*) band magnitudes.

One immediate issue when comparing to simulations will be that stellar population models, initial mass function, etc. used to obtain M_* from the data (from which we compute $M_{\rm HI}/M_*$) is different between what we assume in LOSER versus what RESOLVE assumed to obtain their M_* values. Hence it turns out there is a small offset in M_* that we must first correct. We do so empirically, by using our ML framework to predict the M_* from the photometry in our simulations and from RESOLVE, and then comparing the M_* values.

[4] https://resolve.astro.unc.edu/data/resolve_phot_dr1.txt

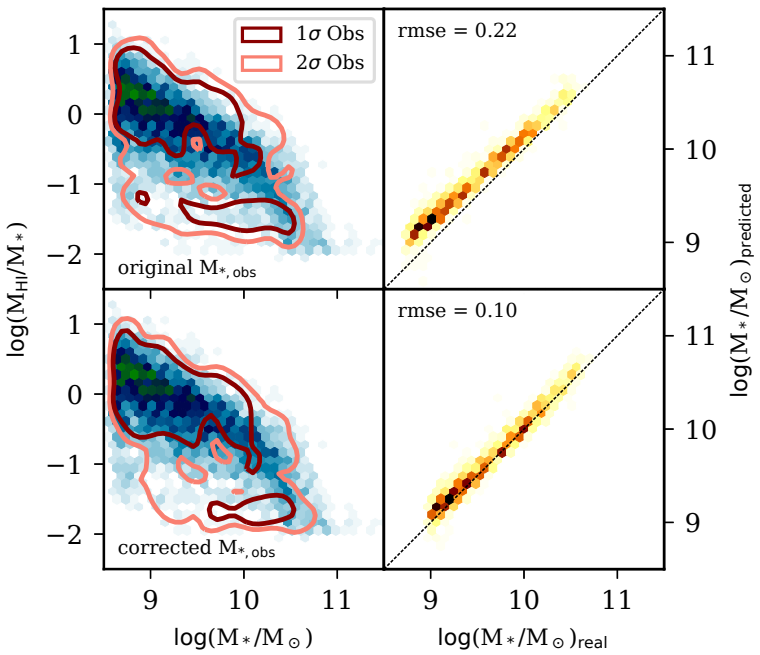


FIGURE 4.7: Left panels: The blue-green maps show the distribution of M_* (x-axes) *vs.* M_{HI}/M_* (y-axes) of the simulated galaxies, while the dark and light red contours show the 1 & 2 σ distributions of the RESOLVE data. Right panels: distributions of the real (x-axes) and predicted (y-axes) galaxy stellar masses of the RESOLVE galaxies. Upper panels show the distributions prior to the correction to observed stellar masses as described in the text, and lower panels after correction. The lack of bimodality in the simulated data (*right* panels) as seen in the data is mainly due to our cut to only include galaxies with $M_{HI}/M_* > 10^{-2}$.

Figure 4.7 (right panels), shows the difference between the original (top) and the corrected (bottom) M_* values from RESOLVE. The original RESOLVE data is offset by $\sim 0.1 - 0.2$ dex; this is within the uncertainties of typical M_* determinations from photometry. The correction we apply is a linear scaling of the stellar masses to match with MUFASA galaxies, obtained by training the DNN model with the simulation to predict the stellar mass of the RESOLVE data, and comparing the result with the real value from RESOLVE. We repeat the process 10× and take the average of the linear slopes and the intercepts, to obtain the following relation: $\log M_{*,corrected} = 0.920 \times \log M_{*,original} + 0.924$. It can be seen that M_* is predicted very tightly, with a scatter of RMSE= 0.1 once the correction is applied. Prior to the correction, the RMSE= 0.22 relative to the 1-to-1 line, which is dominated by the offset rather than the scatter itself. Note that scaling the simulated stellar masses would give the same results, but we don't use this option because we know exactly the stellar mass of the simulated galaxies.

We can also compare the trend of M_{HI}/M_* vs. M_* in the simulations and RESOLVE, which is done in the left panels of Figure 4.7, before (top) and after (bottom) the M_* correction. The green-blue distributions on the left panels are from MUFASA-galaxies whereas the contours are from the observational data. In general, particularly after the correction is applied, the simulations and observations agree quite well for the bulk of the galaxies. A clear trend is seen that lower-M_* galaxies have higher HI fractions. The mean trend of the galaxies with HI is in good agreement between RESOLVE and this simulation, which confirms the agreement versus other data sets shown in Rafieferantsoa et al. (2015). This indicates that MUFASA provides a generally viable model to predict observed HI from photometry.

There is a notable difference that the observational data shows a bimodal distribution that is not seen in the simulated data. This is because we have explicitly ignored galaxies from MUFASA with $M_{HI}/M_* < 0.01$. In MUFASA, we have many galaxies with no HI, while in the observations there is a distribution of low-M_{HI}/M_* values. We will leave more careful modeling of these low-M_{HI}/M_* objects for future work, but we note that the bimodality is going to degrade our results since the ML is unlikely to effectively predict galaxies with M_{HI}/M_* approaching ~ 0.01.

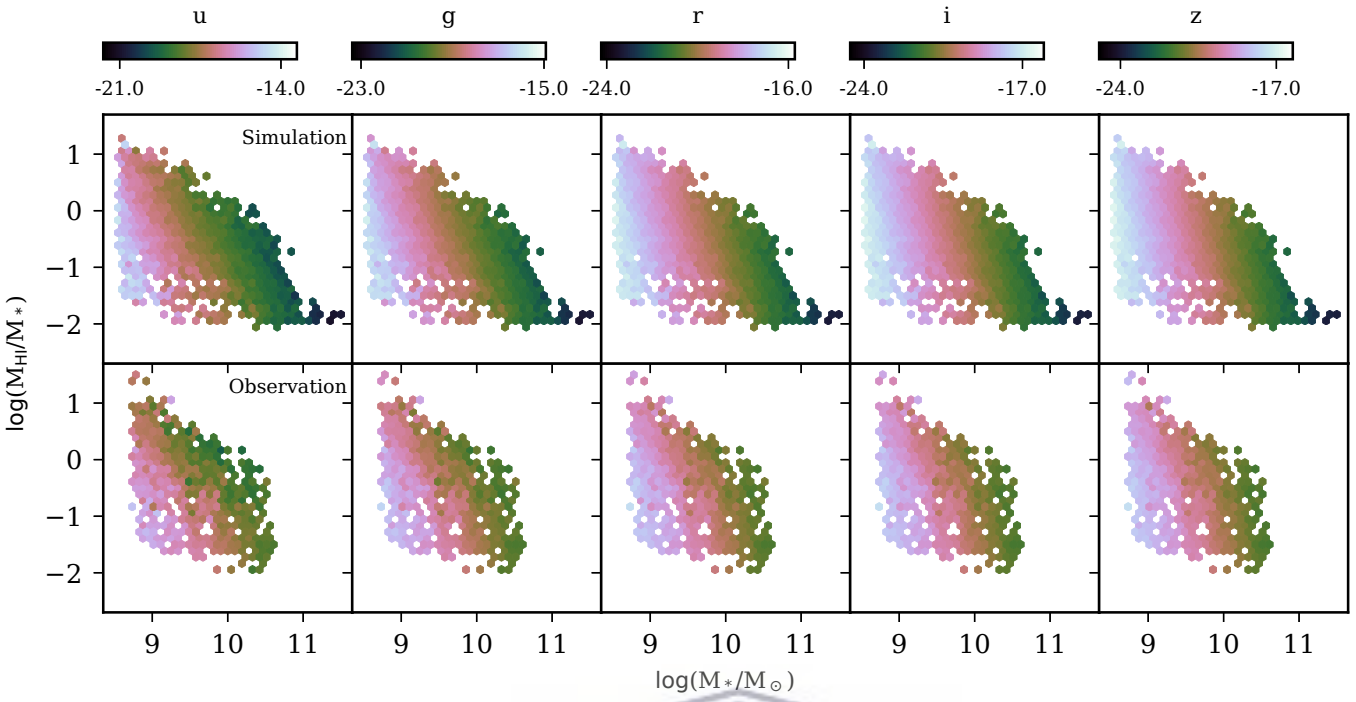


FIGURE 4.8: The x-axes and y-axes represent the stellar masses and HI richness of the galaxies, respectively. Similar to bottom left panel of Figure 4.7, but showing the mean magnitudes in each pixel for the SDSS passbands (columns). The top and bottom panels are for simulated and observational data respectively. The agreement between the two data is noticeable and the range of the observational data are well included in that of the simulated ones.

We also check if the range of magnitudes between the RESOLVE and MUFASA are in broad accordance. Figure 4.8 shows the same distributions as in Figure 4.7 lower left panel, except that now the colours in each hexagonal bin represent the mean magnitudes of the galaxies in that bin. We show *ugriz* magnitudes for illustration but we get similar results for other bands. Each column represents one band. Upper and lower panels are for simulated and observational data respectively. We can clearly see that the trends are consistent. Note that apart from SDSS magnitudes, we also use *NUV*, *J*, *H* and K_s magnitudes in the training. We however point that including all those bands decreases the size of the sample due to missing data in each band. The RESOLVE data contain 2159 galaxies with SDSS magnitudes. When accounting for *NUV*, *J*, *H* and K_s we end up with only 1017 galaxies.

4.6.1.2 ALFALFA data

We use the α.100 ALFALFA data (Haynes et al., 2018) which contain the derived HI mass and the position of the sources in RA and DEC. To obtain the photometric magnitudes (u, g, r, i, z) and the stellar masses, we cross-matched the sources to their SDSS counterpart based on the positions. We allow a 6″ maximum search radius. We end up with 16588 galaxies. We use the pre-built SDSS web interface[5] to do our crossmatching procedure. Similar to RESOLVE data, we need to account for the stellar mass correction for SDSS data using the same method described before and obtain $\log M_{*,\text{corrected}} = 1.022 \times \log M_{*,\text{original}} - 0.4534$.

4.6.2 Training on and predicting observed data

4.6.2.1 RESOLVE results

We first consider the case where we train the regressors using one subset of the RESOLVE data and test them using the other subset (the one which was not used for the training). Due to the relatively small sample in hand, we only use 10% of the data for testing. This case can be considered optimal in the sense that the training and testing sets are drawn from (different parts of) the same sample, so there are no systematic differences.

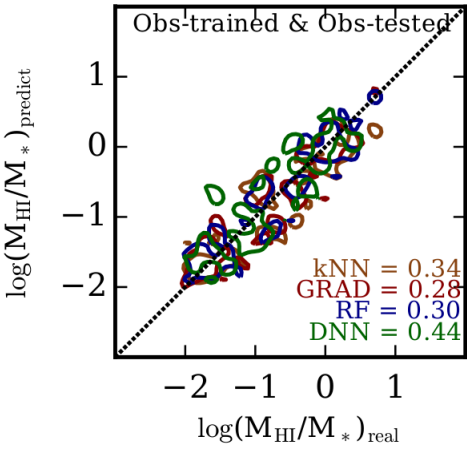


FIGURE 4.9: RESOLVE data. Predictions of the observed HI richness (y-axes) when the different mapping algorithms (colour coded lines) are trained with observational data. The contours correspond to 1σ distribution.

[5] http://skyserver.sdss.org/dr14/en/tools/crossid/crossid.aspx

Figure 4.9 shows our prediction using the test sets. Judging by the contours, it is clear that all the presented models here perform reasonably well, *i.e.* the distribution of the real vs predicted values lie along the identity line, and the predicted values (y-axis) covers all the range of the real values for all regressors. Comparing regressors, GRAD with RMSE= 0.28 performs best followed closely by RF, k-NN and lastly DNN with RMSE= 0.44. Now the trend is reversed such that DNN, which was among the best in the previous scenario becomes the worst in this case. DNN's typically require larger training samples to properly constrain the large number of layers, so it is likely its poor performance owes to the small sample of RESOLVE galaxies.

4.6.2.2 ALFALFA results

The previous results motivate us to test if our models produce similar outputs using a different set of data. Using the larger ALFALFA sample cross-matched with SDSS photometric data, we can equally perform the previous exercise. We train the regressors on the training sample (75% of ALFALFA) and predict the HI richness on the testing sample (remaining 25%).

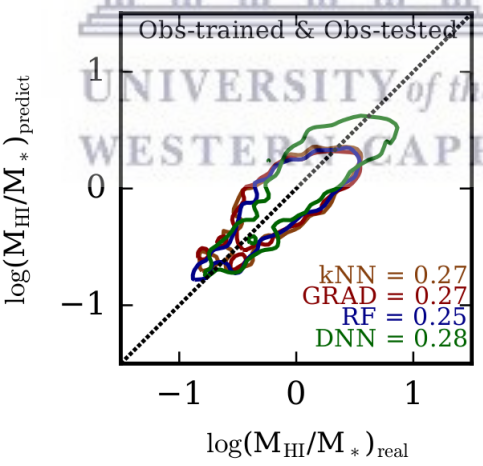


FIGURE 4.10: Similar to Figure 4.9 but with ALFALFA data instead.

Figure 4.10 shows our results. We attain RMSE< 0.3 for all regressors shown here, and none show any systematic bias, with a median lying on the one-to-one line across all HI-richness. In detail, RF performs slightly better than others whereas DNN slightly worse than others. The better performance, compared to the one corresponding to the RESOLVE sample, likely arises in part from the

almost eight times larger sample. To test this, we randomly select 8% of the ALFALFA data for training the DNN model and 2% for testing. We repeat the process 10× and obtain an average RMSE=0.38 which is > 30% higher (worse performance) than the DNN model trained with all the sample. We note that the difference in the ranges of real HI-richness used in DNN and the others is due to the different random splitting between `TensorFlow` and `scikit-learn`, but the resulting performance shows no qualitative difference.

These tests show that our ML framework is able to quantitatively predict observed data when (a distinct subset of) the same data set is used for training. Similar results has been shown in other works such as Teimoorinia et al. (2017), and in fact their predictions are formally even tighter, likely owing to the inclusion of derived galaxy properties in addition to purely photometric data. While these tests are encouraging, our broader aim is to utilize our ML framework to make predictions in regimes where no training set exists, by using our simulated galaxies as the training set. Thus we must now test the case where we train on the simulations, and test against the observations from RESOLVE and ALFALFA.

4.6.3 Training on MUFASA and predicting observed data

A more general application would be where we have no or very limited HI training data, and only photometric data. This might be the case at $z \sim 0.3-1$, where the HI data is almost nonexistent now and even future surveys will provide only a sparse sampling of the most HI-massive objects. In this case, we would like to be able to use the simulations to provide the training set. Naturally, this introduces more uncertainties and assumptions, because the simulations build in a specific physical model which likely is not exactly correct, and does not reproduce the real HI population in all its details. To test how much more uncertain the predictions would be, we can attempt this using RESOLVE and ALFALFA where we *know* what the correct answer is, and see how well the simulation recovers it relative to the case in the previous section where we used observed data themself to train.

In order to mitigate the effects of those uncertainties, one must carefully mimic the input features of the simulated data to encompass those from the observational data as discussed in the previous section. Given that MUFASA reproduces

several observables that are usually used as benchmark for simulation models, such as stellar mass function, HI mass function, specific star formation rate function, etc. (Davé et al., 2016, 2017b,a), we feel confident that it provides a state of the art approach to making predictions for upcoming surveys such as LADUMA or MIGHTEE, *i.e.* using simulated data for training the algorithms and applying it to available observational photometric data.

4.6.3.1 Simulation-trained ML applied to RESOLVE

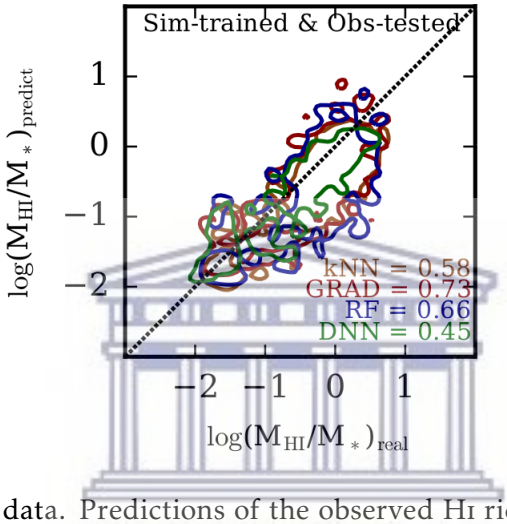


FIGURE 4.11: RESOLVE data. Predictions of the observed HI richness (y-axes) when the different mapping algorithms (colour coded lines) are trained with simulated data. The contours correspond to 1σ distribution.

Figure 4.11 shows the HI richness prediction of our four best models, training the regressors with the simulation data and predicted the HI richness of the RESOLVE data. The contours show the distributions of the RESOLVE HI richness (x-axis) *vs* the predicted HI richness (y-axis) from the models. The numbers on the bottom right of each panel show the RMSE of each model.

Overall, the predictions still lie along the one-to-one relation, indicating that using the simulations to train still provides an adequate prediction in the mean. However, the RMSE values are much higher here than in the right panel. This clearly shows that the simulated sample does not fully mimic the details of the observed sample. Given the discrepancies between simulation and observation, implying differences of the underlying distributions of the two samples, this is not surprising.

k-NN, GRAD and RF now all have RMSE values above 0.5, which is fairly poor. They estimate with larger scatter and a noticeable offset towards lower HI richness values, which is strongest at $\log_{10}(M_{HI}/M_*) \sim 0$ (lower contours are farther from 1:1 line than the upper ones).

Rather remarkably, DNN (green contour) now performs the best in this case, with RMSE = 0.45 and predictions extending to the lowest values ($-2 \leq$) following the 1:1 line. Although DNN was outperformed in Figure 4.4 using only simulated data for training and testing, we can clearly see here that its performance shines in a more difficult scenario, where now the training sample is much larger but the data is more complex. Indeed, the RMSE for DNN hardly changed at all when using the RESOLVE or MUFASA data to train, though this probably arises from the larger training sample offsetting the less homogeneous testing sample. Our results suggest that in this real-world application, DNN can learn better from the simulated data than simpler regressors. We note that the offset in the predictions will always be present in any pairwise set of different data, be it {simulation-observation} or {observation-observation}.

From the two approaches presented in Figures 4.9 and 4.11, we can see that DNN presents robust predictions regardless of the training setups. It is able to learn important features from the simulation and translate those into the observed data. kNN, GRAD and surprisingly RF are less efficient in doing so. The latter only performs best when the training and testing samples are drawn from the same main sample.

4.6.3.2 Simulation-trained ML applied to ALFALFA

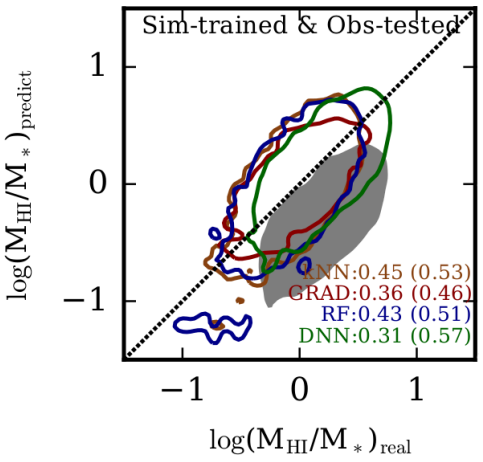


FIGURE 4.12: Similar to Figure 4.11 but with ALFALFA data instead.

Figure 4.12 shows the results when the training is done with MUFASA and the trained models predict the HI content of the ALFALFA sources. The contours show the distributions of the ALFALFA HI-richness (x-axis) *vs* the predicted HI-richness from the models. The numbers inside the brackets are the original RMSE before the linear shift (color coded). The grey shaded area corresponds to the DNN prediction before the shift is applied. Other regressors' (not shown for clarity) have lower offsets. The linear shift is the necessary amount of HI-richness (in dex) to minimize RMSE: *i.e.* the required *intercept* value to make the linear fit of the prediction coincide to the 1:1 line the closest. Such offset is expected due to the nature of the ALFALFA survey which was only targeting HI fluxes resulting in a HI biased-high sample. We did not account for such HI offset with the predictions for RESOLVE data due to the latter being highly complete and volume-limited survey data. The HI offsets are highest for DNN (= 0.43 dex) whereas the other predictions requires only ∼ 0.23 dex. We interpret the difference in those values to be a result of the different slopes of the linear fits between the prediction and the original values. They are 1.06, 1.13, 1.15 and 1.13 for DNN, RF, GRAD and *k*NN respectively. Not accounting for the offsets, one might read the best performance to be that of GRAD with RMSE = 0.46, followed by RF and *k*NN and finally DNN with RMSE = 0.57. We note that previous work by Rafieferantsoa et al. (2015) showed an offset of > 0.5 dex in HI richness between α.40 ALFALFA data cross-matched with SDSS and our previous galaxy formation model. In order to account for this properly,

we would have to create mock ALFALFA selection functions and "observe" our simulations appropriately, which is a challenging task that is beyond the scope of this work, and is not really even feasible given the small simulation volume compared to SDSS.

Figures 4.10 and 4.12 show that the predictions are remarkably good despite training and testing on different data sets. The RMSE~ 0.31 is minimized again for DNN, and other models perform less well but still have RMSE< 0.5. All distributions have medians that closely follow the identity line, showing no systematic offset. This accuracy is significantly better than that obtained using the RESOLVE data. It is unclear what aspects of RESOLVE versus ALFALFA make the latter more closely align with the simulation predictions, since they encompass a similar redshift and mass ranges.

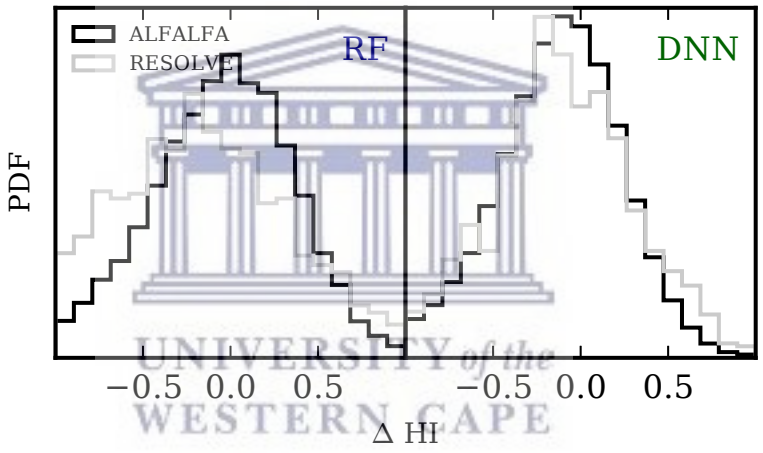


FIGURE 4.13: Log-scaled probability distribution function of the uncertainty of the HI predictions from simulation-trained regressors. *Left*: random forest. *Right*: deep neural networs. The black (grey) lines are the prediction uncertainties of the ALFALFA (RESOLVE) data.

Finally, Figure 4.13 shows the probability distribution functions of the deviation of the predictions from the true values for RESOLVE (grey line) and ALFALFA (black line) data. The regressors are trained from MUFASA data. The left panel is for RF and the right panel for DNN. We log-scale the y-axis to highlight the peculiarities between the distributions. For both DNN and RF, we notice that the peaks of the pdf are below the mean (ΔHI< 0) for the predicted HI values of RESOLVE data. That is already noted in Figure **??** but it is shown more quantitatively here. For ALFALFA, the peak is closer to the mean (~ 0).

Predictions of RESOLVE HI also have more extended tails probably due to the size of the data. The higher counts for under-predicted HI content from the RF regressor (RESOLVE, grey line at ΔHI< 0) is again already seen in Figure **??** (bent shape of the contour).

To summarize, we have shown that training on a subset of observational data can yield a reasonably tight prediction for a testing set taken from the same data. This provides a way to populate photometric surveys in scenarios where a sizeable HI training set is available, such as RESOLVE or ALFALFA; similar results have been shown in previous works (e.g. Teimoorinia et al., 2017). The new aspect in this paper involves training the models using simulated data, and predicting the observational targets. This can result in somewhat higher uncertainties, but still without any significant systematic offsets. In particular, testing a simulation-trained network on over 16,000 ALFALFA galaxies cross-matched to SDSS shows that the simulation-trained ML produces a scatter comparable to that obtained by training on the ALFALFA data itself, around 0.3 dex, when using DNN. In terms of regressors, while RF has been the best choice in more homogeneous training/test set situations (simulation-simulation and observation-observation), when applying the simulation training to observational data, the performance of our deep neural network clearly outshines the others.

4.7 Discussion

Extraction of information in a given set of data is a challenge in all models. Although RF and DNN are our best regressors, they still have difficulty in extracting all the necessary information. That being said, attaining an accuracy of $\mathbf{r} > 85\%$ is a non trivial success for both of regressors. In our training for the DNN, we make sure that the loss function stays unchanged for several training steps to make sure the network learns as much information as it needs but not as much as it might overfit the training data and lose the important information necessary for the prediction. It may be possible to tune this better.

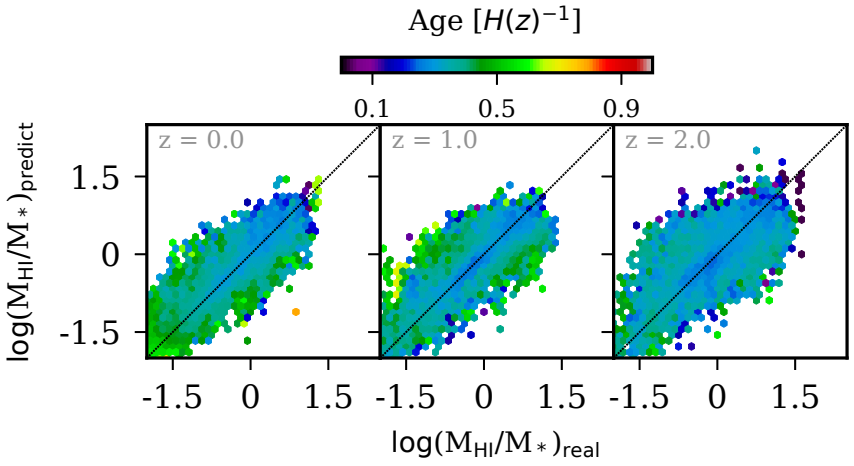


FIGURE 4.14: Mean galaxy age for each pixel in the distribution of the real (x-axes) and predicted (y-axes) HI richness of the simulated galaxies. This result is from the DNN-trained model. Different panels show for different redshift. We use the age of the galaxies at the given redshifts (shown on the top left corner in each panel).

It is possible that photometric surveys can yield other information such as the age, star formation rate, and (from a group catalog) halo masses, albeit with some uncertainties. It is interesting to ask whether providing such information would improve predictions. However, we find that this is not obviously the case. We illustrate this for the mean stellar age in Figure 4.14. Here we show the distribution of the galaxies based on their simulated (real) HI richness and the predicted values from the DNN model, with the colour of each hexagonal bin showing the mean age of galaxies falling in that bin (in unit of the Hubble time at the given redshift). Different panel show different redshifts: *left*, *center*, *right* for $z = \{0, 1, 2\}$ respectively. We can see that for a given HI richness value we cannot see any age gradient in the predicted values, and it remains the case up to $z = 2$. We interpret this to mean the ML model has learned about the age of the galaxies even though that information was not explicitly given in the training set. The same situation happens with the specific star formation rates and the halo mass of the galaxies. This is the case for all of our ML models. Hence providing such information, which introduces further uncertainties from their estimation, is unlikely to be helpful.

Then we might ask why do some regressors perform better than others? We believe that the design of the models themselves may lead to different mapping of the input-output, thus, to improved results depending on the data. Changing the layer structures in DNN or optimising the tree size (or the number of base

estimators) in RF might alleviate certain issues we encountered in our training. We are currently analysing such possibilities and might improve our model in that direction in upcoming work. Also DNN may particularly benefit from a larger simulation training sample with more dynamic range than available in MUFASA.

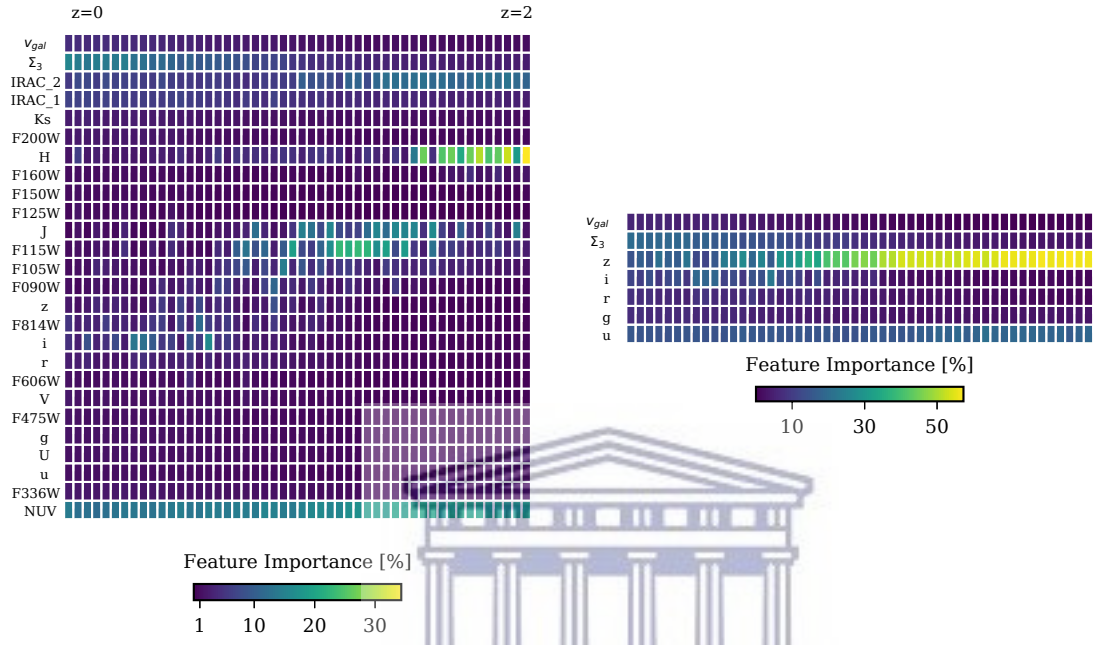


FIGURE 4.15: Evolution of the importance of the input features from the RF training. Each row represents one band with the filter name on the left, except for the 1st (2nd) row which show for the line of sight velocity (3rd nearest neighbour) feature. The bands from the bottom to the top are with increasing peak wavelengths. Left to the right shows the feature importance from $z = 0$ to $z = 2$.

One useful feature of RF is that it provides an estimate of the importance level of the input parameters, based on the rate of incidence that a given parameter is utilised in the decision trees. We show in Figure 4.15 the importance of parameters from RF training. The left subfigure shows the result when using all the available magnitudes from our simulation whereas the right subfigure represents the result when only using the SDSS magnitudes. The 1st (2nd) row in each subfigure show the importance of the line of sight velocity v_{gal} (3rd nearest neighbour Σ_3) from $z = 0$ (left) to $z = 2$ (right). The remaining rows show for bandpass filters (names on the left) with a wide range of peak wavelenghts from 2309Å (bottom row) increasing to 44630Å.

It is interesting to see that Σ_3 becomes increasingly important only at later epochs; this is physically expected since environment becomes an increasing

determinant of HI properties at lower redshifts (Rafieferantsoa & Davé, 2018, or Chapter 2). The line of sight peculiar velocities v_{gal} do not add value to the training, which is unsurprising since it is not obvious why the HI content should care about peculiar velocity (except perhaps through correlations of peculiar velocities and the large-scale potential well); this in a sense serves as a sanity check that our method is not finding physically implausible relationships. In the upper subfigure, the IRAC channels have some importance at higher redshift, particular IRAC 4.5μm while 3.6μm is less important. The H-band magnitude is very important at high redshift but contributes much less at low redshift. The importance of magnitudes between i (6250$\AA$) and J (12500$\AA$) bands move from low to higher peak wavelengths towards higher redshift. NUV magnitudes seem to exhibit relatively high importance at all redshift bins, highlighting the connection between HI and the gas that fuels star formation and hence UV light.

In the lower subfigure with a more restricted input set, z magnitude is very important at higher redshift but becomes less although still important at $z = 0$, whereas the importance of i magnitude increases towards the present day. The value that u magnitude adds to the accuracy of the prediction seems to be relatively constant at all redshifts, following NUV in the upper subfigure. It appears overall that the reddest available photometric band has the highest RF importance level at high redshifts but decreases in importance at lower redshifts, while Σ_3 increase in importance particularly at $z \lesssim 0.5$.

On the whole what the two panels in Figure 4.15 tell us is that given the features available in the data, the feature importance in principle allows one to select only a set of the most important ones in order to achieve a given accuracy. This, amongst other methods like Principal Component Analysis (PCA), is of a great value especially when reducing the dimensionality that might not be avoidable due to a limited computing power or when the dimension is as big as the size of the data (*i.e.* number of features is as large as the number of examples for the training). Also, the importance levels could be helpful in survey design, if a particular photometric band is more useful it might be regarded as higher priority to obtain. However, one must be aware that in many cases, RF importance levels do not truly reflect the necessity of a given data, in the sense that sometimes RF says a particular input is important, but the information from that input is actually encoded in the other inputs, so that removing it does not have as

detrimental effect as one might think (e.g. Agarwal et al., 2018). Likewise, one should not expect the importance of input parameters to be continuous across cosmic time because the algorithm can swap between parameters with similar importances. Properly assessing the importance level would involve re-training the entire data set removing each input in turn, to assess the increase in RMSE, which is highly computationally intensive. Nonetheless, RF importance level can at least provide a guide for this process.

4.8 Conclusion

We have investigated estimating the HI richness of galaxies based on their optical and near-IR survey properties, in particular SDSS $\{u, g, r, i, z\}$, Johnson $\{U, V\}$ and 2MASS $\{J, H, K_s\}$, line of sight velocities, and environmental measures, using machine learning (ML). For our analysis, the training data have been generated from the MUFASA simulation, which has been shown to provide a good description of the HI content of observed galaxies. We have tested various machine learning regressors including random forests and deep neural networks. We considered various input feature combinations, including only SDSS magnitudes and environmental properties, using galaxy colours instead of and in addition to magnitudes, and including 2MASS and Johnson magnitudes. We trained each model to predict M_{HI}/M_* based on an aggregate of all simulated galaxies at $z = 0 - 2$ (f-training), and in 50 individual redshift bins (z-training). As an example application, we applied this framework to the RESOLVE and ALFALFA+SDSS galaxy survey catalogs with HI and photometric data. To measure and compare the performance of each method, we used RMSE, Pearson correlation coefficient **r** , and the correlation slope.

We summarize our main findings as follows:

- By using 75% of the MUFASA data for training and testing on the remaining quarter, we find that all ML methods are able to approximately recover M_{HI}/M_* from galaxy photometry. The accuracy depends both on the input data set and the ML algorithm. Generally, random forests (RF) provides the best performance at $z = 0$, *i.e.* lowest RMSE ≈ 0.25, highest **r** ≈ 0.9, and slope closest to unity, with deep neural network (DNN) close behind.

- At $z \lesssim 1$, it is advantageous to do the ML training at a given redshift rather than aggregating all redshifts. The smaller number of galaxies available for training in the former is outweighed by the conflating of evolutionary trends when aggregating. The RMSE of all ML algorithms increases with redshift, with commensurately lowered **r** and a best-fit slope diverging from unity, though the effect is mild out to $z \sim 0.5$. Predictions at higher redshifts are more challenging owing to reduced trend in M_{HI}/M_* among high-z galaxies, since most galaxies at $z \gtrsim 1$ have similar M_{HI}/M_* prior to significant populations of quenched galaxies arising.

- Providing more input training data results in better predictive power, unsurprisingly. Using only SDSS data results in RMSE≈ 0.3 for RF at $z = 0$, while either including 2MASS data or training on both colours and magnitudes yields a more optimal RMSE. DNN has in the best case similar performance, but it is more strongly dependent on the selected input features.

- All the regressors tend to under-predict the high HI richness and over-predict the low HI richness, as shown by the slope (< 1) of the linear fits between the targets and the predictions. This owes to the regressors being unable to fully capture the scatter in the M_{HI}/M_* values at *e.g.* a given colour, instead tending to push the M_{HI}/M_* towards the mean. This raises the value of low M_{HI}/M_* objects and lowers it for high M_{HI}/M_* objects, resulting in a sub-unity slope. The under-prediction of the high HI richness is more severe at high redshift (Figure 4.6).

- By training our ML framework on a subset of the RESOLVE and ALFALFA+SDSS data and testing it on the remainder, we showed that it is possible to predict M_{HI}/M_* with RMSE $\lesssim 0.3$, which is comparable or better than what is obtained with scaling laws. RF again performs among the best, though GRAD and *k*NN also show slightly better performance. When training on MUFASA and testing on the observed data, we find the best regressor is DNN. The predictions are significantly degraded with RMSE ≈ 0.45 with RESOLVE data, but for the larger ALFALFA+SDSS sample the RMSE ≈ 0.31 which is only slightly worse than the case where we trained on ALFALFA+SDSS data itself. The worse performance in this case likely owes to subtle mismatches between simulation predictions and analysis procedures versus those from the observations. While the scatter

is substantial, the median trend remains well-matched, showing that the ML prediction introduces at most mild systematic biases.

We have shown through this study that it is clearly possible to estimate the HI richness of a galaxy by relying only on the information from photometric magnitudes. We considered various magnitudes from different surveys like SDSS, Johnson and 2MASS in this work, but including other bands is doable. The broadly successful test of training on simulated data and applying to observed data, particularly ALFALFA+SDSS, suggests that the estimation of HI gas at higher redshift using the methods presented here, even with the lack of testing data, is fruitful. With the advent of future surveys such as LADUMA and MIGHTEE, our ML framework constitutes an important new tool to aid studies of neutral hydrogen evolution in galaxies out to intermediate redshifts.

CHAPTER 5

Conclusion

5.1 Summary

We have studied the effects of environment on the gas content and growth of galaxies. We use the galaxy formation models Mufasa implemented in the hydrodynamic code Gizmo to conduct our cosmological scale simulations that treat collisionless dark matter and stellar particles and collisional gas volume elements. Once simulated, we extract our galaxies and haloes with our group finders and compute their properties. We started by looking at the environmental constraints on the evolution of galaxies in chapter 2. Particularly, we focused on the galactic conformity that quantifies the similarity between the neighbouring galaxy properties. We first test the consistency of the simulation models compared to the already quantified conformity signals from observed data then extend the study to galaxy properties that are more difficult to observe. The results suggest that the extended, but weak conformity for less massive haloes started since very recently but the restricted and strong conformity in more massive hosts began earlier. We continued by looking at the timescales for which the galaxies considerably deplete their gas content and star formation rate as a consequence of their properties in chapter 3. Internal properties such as the stellar mass of the galaxies anticorrelates with the timescales as much as the external properties such as the environment or the halo mass of the first infall. The evolution of gas related galaxy properties such as HI, H_2 and SFR are in conjunction at first infall. The gas consumption timescale is less of a function of the distance of the galaxy to the center of its host, although the gas content is strongly affected by the latter. Finally, we attempted to predict the HI content of galaxies using machine learning techniques in chapter 4. With our simulated data with galaxies evolving from $z > 5$ to $z = 0$, and complete samples of observed data limited to much lower redshift, we trained our machine learning models to map the photometric properties of those galaxies with their HI contents. With the testing sample, the predicted HI mass of the galaxies are $\sim 25\%$ better compared to pure statistical approaches.

5.2 Relevance to the Upcoming Surveys

This thesis interconnects the importance of galaxy formation models, such as Mufasa, as valuable tools for prediction and one of the important key sciences

of the Square Kilometer Array project to study the cosmic neutral atomic hydrogen.

The cosmic distribution of HI in galaxies or HI mass function (HIMF) can be used to quantify the relative importance of HI in the growth of galaxies. Understanding of different factors, particularly the environmental density, affecting the HIMF requires complete dataset of 21cm in emission. Surveys such as ALFALFA (Haynes et al., 2018) provide extensive extragalactic catalogs that allow the exploration of the environmental effect on the HIMF (Jones et al., 2018) but it is limited to only $z < 0.22$. The future LADUMA survey on MeerKAT (Holwerda et al., 2012) will probe as far as $z \sim 1.4$ to extend the current insight in the HIMF. We predicted previously (Rafieferantsoa et al., 2015) that HI-richness is largely affected by the galaxy environment quantified by the virialised halo mass. We then analysed the robustness of our models in terms of environmental effects by looking at the galactic conformity of our simulated sample compared to the observed data (Rafieferantsoa & Davé, 2018). The level of agreement confirms the quality of the prediction and the applicability of the results to tune for the modeled large survey programs such as LADUMA.

One of the key questions of the SKA project is also the connection between the HI content, stellar mass and halo mass of galaxies. The choice of the LADUMA survey, to be done on the SKA precursor MeerKAT, to point in the Chandra Deep Field South is the availability of multi-wavelength data in those regions. MUFASA produced simulated galaxies covering a wide range of stellar masses, HI masses and halo masses. Davé et al. (2017a), Davé et al. (2017b) and Rafieferantsoa & Davé (2018) provided benchmarks on the galaxy population with respect to those properties and chapter 3 (Rafieferantsoa et al. in preparation) looks at the connection between them and their effects on the gas consumption timescales. The current state of the MUFASA simulation offers a viable option to pre-analyse different aspects of this science question.

Furthermore, the MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) Survey (Jarvis et al., 2016) will observe 20 square degrees of the sky to cover the highest mass range of HI in galaxies. Together with the LADUMA survey that will provide ~ 5 times better sensitivity and thus the ability to observe lower HI masses, these surveys will remarkably advance our current understanding of the HI content in galaxies and their evolution. ALFALFA has initiated the effort by providing extended HI data sample for the local Universe.

We used the ALFALFA data to contrain the Mufasa models and predict higher redshift HI content self-consistently.

Those results are important in the field of galaxy evolution but numerous improvements are still expected.

5.3 Theoretical Prospects

Mufasa is one of the galaxy formation models that allow the study of the time related growth of both individual galaxies and the whole sample as a statistical collection. One particular aspect of the code requires improvements. The use of halo mass to trigger the star formation quenching is empirically reasonable but better models should be implemented to self-consistently suppress the star formation as a product of physical phenomena inside the galaxies. An attempt to do so is to the implementation of black hole feedback (Davé et al. in preparation). Chapter 2 provides a new approach to testing quenching models by examining neighbouring galaxies, as opposed to tests that focus on the properties of the central galaxies or their surrounding gas. It remains to be seen what the impact of a more physically-motivated quenching prescription using energy released from AGN would have on conformity; this will be explored in future work.

In chapter 3, the study was limited by only galaxy members of 2 groups. Extension of that sample as well as the use of better motivated physical models are expected in the future to further quantify what was found here.

In chapter 4, we have only selected galaxies that are observable in HI, with a threshold of $M_{HI}/M_*> 10^{-2}$. This raises a key question: *"Would a model still generalize well if one also included the* HI*-depleted galaxies in the dataset for the training?"*. There are two ways to address this question. First, we can simply add the HI-deficient galaxies in the dataset and redo the fitting procedure prescribed in this work, although from the standpoint of observations, predicting the HI richness of a HI-depleted or gas-starved galaxy is not really meaningful. The more elegant approach would be to first use machine learning to classify galaxies based on their observable features whether they are HI deficient or not, then only estimate its HI richness (based on the same features) in the case

it would potentially contain observable HI. Of course, the minimal value of observed HI can be a free parameters in our model but in reality that should depend on the telescope capabilities. Future work will discuss these solutions, provide more tailored predictions for upcoming surveys, utilise larger training samples that could particularly help improve the current *deep neural network* results, and make this tool available to the community.

Bibliography

Abadi, M. et al., 2015, *TensorFlow: Large-scale machine learning on heterogeneous systems*. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Abel, T., Bryan, G. L. & Norman, M. L., 2002, Science, vol. 295:pp. 93. doi:10.1126/science.295.5552.93.

Abolfathi, B. et al., 2018, ApJS, vol. 235:42. doi:10.3847/1538-4365/aa9e8a.

Agarwal, S., Davé, R. & Bassett, B. A., 2018, MNRAS. doi:10.1093/mnras/sty1169.

Agertz, O. et al., 2007, MNRAS, vol. 380:pp. 963. doi:10.1111/j.1365-2966.2007.12183.x.

Altman, N. S., 1992, j-AMER-STAT, vol. 46(3):pp. 175. ISSN 0003-1305 (print), 1537-2731 (electronic).

Anglés-Alcázar, D. et al., 2015, ApJ, vol. 800:127. doi:10.1088/0004-637X/800/2/127.

Ann, H. B., Park, C. & Choi, Y.-Y., 2008, MNRAS, vol. 389:pp. 86. doi:10.1111/j.1365-2966.2008.13581.x.

Bagla, J. S. & Ray, S., 2003, New Astronomy, vol. 8:pp. 665. doi:10.1016/S1384-1076(03)00056-3.

Baker, A. J. et al., 2018, *In American Astronomical Society Meeting Abstracts #231*, vol. 231 of *American Astronomical Society Meeting Abstracts*, p. 231.07.

Baldry, I. K. et al., 2004, ApJ, vol. 600:pp. 681. doi:10.1086/380092.

——, 2012, MNRAS, vol. 421:pp. 621. doi:10.1111/j.1365-2966.2012.20340.x.

Balogh, M. L. et al., 2004, ApJL, vol. 615:pp. L101. doi:10.1086/426079.

Benson, A. J. et al., 2003, ApJ, vol. 599:pp. 38. doi:10.1086/379160.

Berti, A. M. et al., 2017, ApJ, vol. 834:87. doi:10.3847/1538-4357/834/1/87.

Bigelow, B. C. & Dressler, A. M., 2003, *In* Iye, M. & Moorwood, A. F. M., eds., *Instrument Design and Performance for Optical/Infrared Ground-based Telescopes*, vol. 4841 of *Proc. SPIE*, pp. 1727–1738. doi:10.1117/12.461870.

Binney, J., 1977, ApJ, vol. 215:pp. 483. doi:10.1086/155378.

Birnboim, Y. & Dekel, A., 2003, MNRAS, vol. 345:pp. 349. doi:10.1046/j.1365-8711.2003.06955.x.

Blumenthal, G. R., Pagels, H. & Primack, J. R., 1982, Nature, vol. 299:p. 37. doi:10.1038/299037a0.

Bond, J. R., Szalay, A. S. & Turner, M. S., 1982, Physical Review Letters, vol. 48:pp. 1636. doi:10.1103/PhysRevLett.48.1636.

Brammer, G. B. et al., 2012, ApJS, vol. 200:13. doi:10.1088/0067-0049/200/2/13.

Bray, A. D. et al., 2016, MNRAS, vol. 455:pp. 185. doi:10.1093/mnras/stv2316.

Breiman, L., 2001, Mach. Learn., vol. 45(1):pp. 5. ISSN 0885-6125. doi:10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Catinella, B. et al., 2010, MNRAS, vol. 403:pp. 683. doi:10.1111/j.1365-2966.2009.16180.x.

——, 2013, MNRAS, vol. 436:pp. 34. doi:10.1093/mnras/stt1417.

Conroy, C. & Gunn, J. E., 2010, *FSPS: Flexible Stellar Population Synthesis*. Astrophysics Source Code Library.

Cortes, C. & Vapnik, V., 1995, Machine Learning, vol. 20(3):pp. 273. ISSN 1573-0565. doi:10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

Crain, R. A. et al., 2009, MNRAS, vol. 399:pp. 1773. doi:10.1111/j.1365-2966.2009.15402.x.

Croton, D. J. et al., 2006, MNRAS, vol. 367:pp. 864. doi:10.1111/j.1365-2966.2006.09994.x.

Cunnama, D. et al., 2014, MNRAS, vol. 438:pp. 2530. doi:10.1093/mnras/stt2380.

Cybenko, G., 1989, Mathematics of Control, Signals and Systems, vol. 2(4):pp. 303. ISSN 1435-568X. doi:10.1007/BF02551274. URL https://doi.org/10.1007/BF02551274.

Davé, R., Rafieferantsoa, M. H. & Thompson, R. J., 2017a, MNRAS, vol. 471:pp. 1671. doi:10.1093/mnras/stx1693.

Davé, R., Thompson, R. & Hopkins, P. F., 2016, MNRAS, vol. 462:pp. 3265. doi:10.1093/mnras/stw1862.

Davé, R. et al., 2013, MNRAS, vol. 434:pp. 2645. doi:10.1093/mnras/stt1274.

——, 2017b, MNRAS, vol. 467:pp. 115. doi:10.1093/mnras/stx108.

De Lucia, G. et al., 2012, MNRAS, vol. 423:pp. 1277. doi:10.1111/j.1365-2966.2012.20983.x.

De Vaucouleurs, G., de Vaucouleurs, A. & Corwin, J. R., 1976, *In Second reference catalogue of bright galaxies, Vol. 1976, p. Austin: University of Texas Press.*, vol. 1976.

Duffy, A. R. et al., 2012, MNRAS, vol. 420:pp. 2799. doi:10.1111/j.1365-2966.2011.19894.x.

Eckert, K. D. et al., 2015, ApJ, vol. 810:166. doi:10.1088/0004-637X/810/2/166.

Fabian, A., 2012, Annual Review of Astronomy and Astrophysics, vol. 50(1):pp. 455. doi:10.1146/annurev-astro-081811-125521. URL https://doi.org/10.1146/annurev-astro-081811-125521.

Faucher-Giguère, C.-A. et al., 2010, ApJ, vol. 725:pp. 633. doi:10.1088/0004-637X/725/1/633.

Foltz, R. et al., 2018, ArXiv e-prints.

Fossati, M. et al., 2017, ApJ, vol. 835:153. doi:10.3847/1538-4357/835/2/153.

Friedman, J. H., 2000, Annals of Statistics, vol. 29:pp. 1189.

Gabor, J. M. & Davé, R., 2012, MNRAS, vol. 427:pp. 1816. doi:10.1111/j.1365-2966.2012.21640.x.

——, 2015, MNRAS, vol. 447:pp. 374. doi:10.1093/mnras/stu2399.

Gabor, J. M. et al., 2010, MNRAS, vol. 407:pp. 749. doi:10.1111/j.1365-2966.2010.16961.x.

Geha, M. et al., 2012, ApJ, vol. 757:85. doi:10.1088/0004-637X/757/1/85.

Giavalisco, M. et al., 2004, ApJL, vol. 600:pp. L93. doi:10.1086/379232.

Giovanelli, R. et al., 2005, AJ, vol. 130:pp. 2598. doi:10.1086/497431.

Grogin, N. A. et al., 2011, ApJS, vol. 197:35. doi:10.1088/0067-0049/197/2/35.

Gunn, J. E. & Gott, J. R., III, 1972, ApJ, vol. 176:p. 1. doi:10.1086/151605.

Haardt, F. & Madau, P., 2012, ApJ, vol. 746:125. doi:10.1088/0004-637X/746/2/125.

Hahn, O. & Abel, T., 2011, MNRAS, vol. 415:pp. 2101. doi:10.1111/j.1365-2966.2011.18820.x.

Hartley, W. G. et al., 2015, MNRAS, vol. 451:pp. 1613. doi:10.1093/mnras/stv972.

Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 ed., 2009. URL http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

Hatfield, P. W. & Jarvis, M. J., 2017, MNRAS, vol. 472:pp. 3570. doi:10.1093/mnras/stx2155.

Haynes, M. P. et al., 2018, ApJ, vol. 861:49. doi:10.3847/1538-4357/aac956.

Hearin, A. P., Behroozi, P. S. & van den Bosch, F. C., 2016, MNRAS, vol. 461:pp. 2135. doi:10.1093/mnras/stw1462.

Hearin, A. P., Watson, D. F. & van den Bosch, F. C., 2015, MNRAS, vol. 452:pp. 1958. doi:10.1093/mnras/stv1358.

Hinton, G. E., Osindero, S. & Teh, Y.-W., 2006, Neural Computation, vol. 18(7):pp. 1527. doi:10.1162/neco.2006.18.7.1527. URL https://doi.org/10.1162/neco.2006.18.7.1527. PMID: 16764513.

Hirschmann, M. et al., 2014, MNRAS, vol. 444:pp. 2938. doi:10.1093/mnras/stu1609.

Holwerda, B. W., Blyth, S.-L. & Baker, A. J., 2012, *In* Tuffs, R. J. & Popescu, C. C., eds., *The Spectral Energy Distribution of Galaxies - SED 2011*, vol. 284 of *IAU Symposium*, pp. 496–499. doi:10.1017/S1743921312009702.

Hopkins, P. F., 2015, MNRAS, vol. 450:pp. 53. doi:10.1093/mnras/stv195.

Hopkins, P. F. et al., 2014, MNRAS, vol. 445:pp. 581. doi:10.1093/mnras/stu1738.

Hornik, K., 1991, Neural Networks, vol. 4(2):pp. 251 . ISSN 0893-6080. doi:https://doi.org/10.1016/0893-6080(91)90009-T. URL http://www.sciencedirect.com/science/article/pii/089360809190009T.

Hunter, J. D., 2007, Computing In Science & Engineering, vol. 9(3):pp. 90. doi:10.1109/MCSE.2007.55.

Jarvis, M. et al., 2016, *In Proceedings of MeerKAT Science: On the Pathway to the SKA. 25-27 May, 2016 Stellenbosch, South Africa (MeerKAT2016). Online at ¡A href="href="¿href="https://pos.sissa.it/cgi-bin/reader/conf.cgi?confid=277¡/A¿, id.6*, p. 6.

Jones, M. G. et al., 2016, MNRAS, vol. 457:pp. 4393. doi:10.1093/mnras/stw263.

——, 2018, MNRAS, vol. 477:pp. 2. doi:10.1093/mnras/sty521.

Kannappan, S. J., 2004, ApJL, vol. 611:pp. L89. doi:10.1086/423785.

Kannappan, S. et al., 2011, *In American Astronomical Society Meeting Abstracts #217*, vol. 43 of *Bulletin of the American Astronomical Society*, p. 334.14.

Kauffmann, G., 2015, MNRAS, vol. 454:pp. 1840. doi:10.1093/mnras/stv2113.

Kauffmann, G. et al., 2003, MNRAS, vol. 341:pp. 33. doi:10.1046/j.1365-8711.2003.06291.x.

——, 2004, MNRAS, vol. 353:pp. 713. doi:10.1111/j.1365-2966.2004.08117.x.

——, 2013, MNRAS, vol. 430:pp. 1447. doi:10.1093/mnras/stt007.

Kawinwanichakij, L. et al., 2016, ApJ, vol. 817:9. doi:10.3847/0004-637X/817/1/9.

Kennicutt, R. C., Jr., 1998, ApJ, vol. 498:pp. 541. doi:10.1086/305588.

Kereš, D. et al., 2005, MNRAS, vol. 363:pp. 2. doi:10.1111/j.1365-2966.2005.09451.x.

Kernighan, B. W. *The C Programming Language*. Prentice Hall Professional Technical Reference, 2nd ed., 1988. ISBN 0131103709.

Kingma, D. P. & Ba, J., 2014, CoRR, vol. abs/1412.6980.

Klypin, A. A., Trujillo-Gomez, S. & Primack, J., 2011, ApJ, vol. 740:102. doi:10.1088/0004-637X/740/2/102.

Koekemoer, A. M. et al., 2011, ApJS, vol. 197:36. doi:10.1088/0067-0049/197/2/36.

Krumholz, M. R. & Gnedin, N. Y., 2011, ApJ, vol. 729:p. 36. doi:10.1088/0004-637X/729/1/36.

Lacerna, I. et al., 2018, MNRAS, vol. 475:pp. 1177. doi:10.1093/mnras/stx3253.

Lanson, N. & Vila, J., 2008a, SIAM Journal on Numerical Analysis, vol. 46(4):pp. 1912. doi:10.1137/S0036142903427718. URL https://doi.org/10.1137/S0036142903427718.

——, 2008b, SIAM Journal on Numerical Analysis, vol. 46(4):pp. 1935. doi:10.1137/S003614290444739X. URL https://doi.org/10.1137/S003614290444739X.

Lara-López, M. A. et al., 2010, A&A, vol. 519:p. A31. doi:10.1051/0004-6361/200913886.

Larson, R. B., Tinsley, B. M. & Caldwell, C. N., 1980, ApJ, vol. 237:pp. 692. doi:10.1086/157917.

Lawrence, A. et al., 2007, MNRAS, vol. 379:pp. 1599. doi:10.1111/j.1365-2966.2007.12040.x.

Lilly, S. J. et al., 2007, ApJS, vol. 172:pp. 70. doi:10.1086/516589.

Lintott, C. J. et al., 2008, MNRAS, vol. 389:pp. 1179. doi:10.1111/j.1365-2966.2008.13689.x.

Madau, P. & Dickinson, M., 2014, ARA&A, vol. 52:pp. 415. doi:10.1146/annurev-astro-081811-125615.

Mannucci, F. et al., 2010, MNRAS, vol. 408:pp. 2115. doi:10.1111/j.1365-2966.2010.17291.x.

McCracken, H. J. et al., 2012, VizieR Online Data Catalog, vol. 354.

McGee, S. L. et al., 2009, MNRAS, vol. 400:pp. 937. doi:10.1111/j.1365-2966.2009.15507.x.

McNamara, B. R. & Nulsen, P. E. J., 2007, ARA&A, vol. 45:pp. 117. doi:10.1146/annurev.astro.45.051806.110625.

Mitra, S., Davé, R. & Finlator, K., 2015, MNRAS, vol. 452:pp. 1184. doi:10.1093/mnras/stv1387.

Morganti, R. et al., 2006, MNRAS, vol. 371:pp. 157. doi:10.1111/j.1365-2966.2006.10681.x.

Morris, J. P., 1996, PASA, vol. 13:pp. 97.

Moster, B. P., Naab, T. & White, S. D. M., 2013, MNRAS, vol. 428:pp. 3121. doi:10.1093/mnras/sts261.

Muratov, A. L. et al., 2015, MNRAS, vol. 454:pp. 2691. doi:10.1093/mnras/stv2126.

Murray, N., Quataert, E. & Thompson, T. A., 2005, ApJ, vol. 618:pp. 569. doi:10.1086/426067.

Newburgh, L. B. et al., 2016, *In Ground-based and Airborne Telescopes VI*, vol. 9906 of *Proc. SPIE*, p. 99065X. doi:10.1117/12.2234286.

Odekon, M. C. et al., 2016, ApJ, vol. 824:110. doi:10.3847/0004-637X/824/2/110.

Oman, K. A. & Hudson, M. J., 2016, MNRAS, vol. 463:pp. 3083. doi:10.1093/mnras/stw2195.

Oppenheimer, B. D. et al., 2010, MNRAS, vol. 406:pp. 2325. doi:10.1111/j.1365-2966.2010.16872.x.

Pedregosa, F. et al., 2011, J. Mach. Learn. Res., vol. 12:pp. 2825. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1953048.2078195.

Peebles, P. J. E., 1968, ApJ, vol. 153:p. 1. doi:10.1086/149628.

——, 1982, ApJL, vol. 263:pp. L1. doi:10.1086/183911.

Peng, Y.-j. et al., 2010, ApJ, vol. 721:pp. 193. doi:10.1088/0004-637X/721/1/193.

——, 2012, ApJ, vol. 757:4. doi:10.1088/0004-637X/757/1/4.

Penzias, A. A. & Wilson, R. W., 1965, ApJ, vol. 142:pp. 419. doi:10.1086/148307.

Planck et al., 2016, A&A, vol. 594:A13. doi:10.1051/0004-6361/201525830.

Price, D. J., 2012, Journal of Computational Physics, vol. 231:pp. 759. doi:10.1016/j.jcp.2010.12.011.

Quilis, V., Planelles, S. & Ricciardelli, E., 2017, MNRAS, vol. 469:pp. 80. doi:10.1093/mnras/stx770.

Rafieferantsoa, M. & Davé, R., 2018, MNRAS, vol. 475:pp. 955. doi:10.1093/mnras/stx3293.

Rafieferantsoa, M. et al., 2015, MNRAS, vol. 453:pp. 3980. doi:10.1093/mnras/stv1933.

Rahmati, A. & Schaye, J., 2014, MNRAS, vol. 438:pp. 529. doi:10.1093/mnras/stt2235.

Rahmati, A. et al., 2013, MNRAS, vol. 430:pp. 2427. doi:10.1093/mnras/stt066.

Rossum, G. *Python reference manual*. Tech. rep., Amsterdam, The Netherlands, The Netherlands, 1995.

Schlafly, E. F. & Finkbeiner, D. P., 2011, ApJ, vol. 737:103. doi:10.1088/0004-637X/737/2/103.

Schmidt, M., 1959, ApJ, vol. 129:p. 243. doi:10.1086/146614.

Shao, Z. et al., 2007, ApJ, vol. 659:pp. 1159. doi:10.1086/511131.

Sijacki, D. et al., 2012, MNRAS, vol. 424:pp. 2999. doi:10.1111/j.1365-2966.2012.21466.x.

Simha, V. et al., 2009, MNRAS, vol. 399:pp. 650. doi:10.1111/j.1365-2966.2009.15341.x.

Sin, L. P. T., Lilly, S. J. & Henriques, B. M. B., 2017, MNRAS, vol. 471:pp. 1192. doi:10.1093/mnras/stx1674.

Somerville, R. S. & Davé, R., 2015, ARA&A, vol. 53:pp. 51. doi:10.1146/annurev-astro-082812-140951.

Springel, V., 2005, MNRAS, vol. 364:pp. 1105. doi:10.1111/j.1365-2966.2005.09655.x.

——, 2010, MNRAS, vol. 401:pp. 791. doi:10.1111/j.1365-2966.2009.15715.x.

Springel, V., Di Matteo, T. & Hernquist, L., 2005, MNRAS, vol. 361:pp. 776. doi:10.1111/j.1365-2966.2005.09238.x.

Stark, D. V. et al., 2016, ApJ, vol. 832:126. doi:10.3847/0004-637X/832/2/126.

Straatman, C. M. S. et al., 2014, ApJL, vol. 783:L14. doi:10.1088/2041-8205/783/1/L14.

Tal, T. et al., 2014, ArXiv e-prints.

Teimoorinia, H., Ellison, S. L. & Patton, D. R., 2017, MNRAS, vol. 464:pp. 3796. doi:10.1093/mnras/stw2606.

Thompson, R., 2014, *pyGadgetReader: GADGET snapshot reader for python*. Astrophysics Source Code Library.

Tinker, J. L. et al., 2018, MNRAS, vol. 477:pp. 935. doi:10.1093/mnras/sty666.

Tomczak, A. R. et al., 2014, ApJ, vol. 783:85. doi:10.1088/0004-637X/783/2/85.

Toro, E. F. *Riemann Solvers and Numerical Methods for Fluid Dynamics*. Springer-Verlag, Berlin Heidelberg, 333 ed., 2009.

van Gorkom, J. H., 2004, Clusters of Galaxies: Probes of Cosmological Structure and Galaxy Evolution, p. 305.

van Gorkom, J. H. et al., 2003, Ap&SS, vol. 285:pp. 219. doi:10.1023/A:1024607103456.

Vapnik, V. N. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8.

Vogelsberger, M. et al., 2014, MNRAS, vol. 444:pp. 1518. doi:10.1093/mnras/stu1536.

Voit, G. M., 2005, Advances in Space Research, vol. 36:pp. 701. doi:10.1016/j.asr.2005.02.042.

Wagoner, R. V., Fowler, W. A. & Hoyle, F., 1967, ApJ, vol. 148:p. 3. doi:10.1086/149126.

Wang, J. et al., 2011, MNRAS, vol. 412:pp. 1081. doi:10.1111/j.1365-2966.2010.17962.x.

——, 2013, MNRAS, vol. 433:pp. 270. doi:10.1093/mnras/stt722.

——, 2015, MNRAS, vol. 453:pp. 2399. doi:10.1093/mnras/stv1767.

Weinmann, S. M. et al., 2006a, MNRAS, vol. 366:pp. 2. doi:10.1111/j.1365-2966.2005.09865.x.

——, 2006b, ArXiv Astrophysics e-prints.

Wetzel, A. R., Tollerud, E. J. & Weisz, D. R., 2015, ApJL, vol. 808:L27. doi:10.1088/2041-8205/808/1/L27.

Wetzel, A. R. et al., 2013, MNRAS, vol. 432:pp. 336. doi:10.1093/mnras/stt469.

White, S. D. M. & Rees, M. J., 1978, MNRAS, vol. 183:pp. 341. doi:10.1093/mnras/183.3.341.

Wisnioski, E. et al., 2015, ApJ, vol. 799:209. doi:10.1088/0004-637X/799/2/209.

York, D. G. et al., 2000, AJ, vol. 120:pp. 1579. doi:10.1086/301513.

Zhang, W. et al., 2009, MNRAS, vol. 397:pp. 1243. doi:10.1111/j.1365-2966.2009.15050.x.

Zu, Y. & Mandelbaum, R., 2018, MNRAS, vol. 476:pp. 1637. doi:10.1093/mnras/sty279.