Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **455,** 642–658 (2015) doi:10.1093/mnras/stv2310

Machine learning and cosmological simulations – I. Semi-analytical models

Harshil M. Kamdar,[1,2★] Matthew J. Turk[2,3★] and Robert J. Brunner[2,3,4,5★]

[1]*Department of Physics, University of Illinois, Urbana, IL 61801, USA*
[2]*Department of Astronomy, University of Illinois, Urbana, IL 61801, USA*
[3]*National Center for Supercomputing Applications, Urbana, IL 61801, USA*
[4]*Department of Statistics, University of Illinois, Champaign, IL 61820, USA*
[5]*Beckman Institute For Advanced Science and Technology, University of Illinois, Urbana, IL 61801, USA*

Accepted 2015 October 1. Received 2015 September 30; in original form 2015 July 2

ABSTRACT

We present a new exploratory framework to model galaxy formation and evolution in a hierarchical Universe by using machine learning (ML). Our motivations are two-fold: (1) presenting a new, promising technique to study galaxy formation, and (2) quantitatively analysing the extent of the influence of dark matter halo properties on galaxies in the backdrop of semi-analytical models (SAMs). We use the influential Millennium Simulation and the corresponding Munich SAM to train and test various sophisticated ML algorithms (k-Nearest Neighbors, decision trees, random forests, and extremely randomized trees). By using only essential dark matter halo physical properties for haloes of $M > 10^{12}\,\mathrm{M}_{\odot}$ and a partial merger tree, our model predicts the hot gas mass, cold gas mass, bulge mass, total stellar mass, black hole mass and cooling radius at $z = 0$ for each central galaxy in a dark matter halo for the Millennium run. Our results provide a unique and powerful phenomenological framework to explore the galaxy–halo connection that is built upon SAMs and demonstrably place ML as a promising and a computationally efficient tool to study small-scale structure formation.

Key words: galaxies: evolution – galaxies: formation – galaxies: haloes – cosmology: theory – large-scale structure of Universe.

1 INTRODUCTION

In recent years, with the introduction of surveys such as SDSS,[1] DES,[2] and LSST,[3] the amount of data available to astronomers has exploded. These massive data sets have enabled astronomers to form and test sophisticated models that explain cosmic structure formation in the Universe. Cosmological simulations are a rich subset of these models and have consequently, also been on the rise; these simulations provide a concrete link between theory and observation. It has been argued that the Λ cold dark matter (ΛCDM) model (Peebles 1982; Blumenthal et al. 1984; Davis et al. 1985) is as widely accepted as it is today largely due to the emergence of these high-resolution numerical simulations (Springel 2005). However, modelling galaxy formation accurately by using numerical simulations remains an important problem in modern astrophysics, both scientifically and computationally.

The evolution of collisionless DM particles at large scales has been studied exhaustively at unprecedentedly high resolutions, given the meteoric rise in computational power and the relative simplicity of these simulations (Springel 2005; Springel et al. 2005; Klypin, Trujillo-Gomez & Primack 2011; Angulo et al. 2012; Skillman et al. 2014). The formation of structure on the scale of galaxies, however, has been incredibly difficult to model (Somerville & Davé 2015); the difficulty arises primarily because baryonic physics at this scale is governed by a wide range of dissipative and/or non-linear processes, some of which are poorly understood (Kang et al. 2005; Baugh 2006; Somerville & Davé 2015).

Broadly speaking, there are two prevalent techniques used to understand galaxy formation and evolution: semi-analytical models (SAM) and simulations that include both hydrodynamics and gravity. The former is a post de facto technique that combines DM-only simulations with approximate physical processes at the scale of a galaxy (Baugh 2006). The SAM used in this work is detailed in Croton et al. (2006), De Lucia et al. (2006), De Lucia & Blaizot (2007, hereafter DLB07), and Guo et al. (2011, hereafter G11). For a general, exhaustive review of the motivation of SAMs and a comparison of different SAMs, the reader is referred to Baugh (2006), Somerville & Davé (2015), and Knebe et al. (2015). *N*-body + hydrodynamical simulations (NBHS) evolve baryonic

★ E-mail: hkamdar2@illinois.edu (HMK); mjturk@illinois.edu (MJT); bigdog@illinois.edu (RJB)
[1] www.sdss.org
[2] www.darkenergysurvey.org
[3] www.lsst.org

components using fluid dynamics alongside regular DM evolution. The biggest advantage of NBHS over SAMs is the self-consistent way in which gaseous interactions are treated in the former. However, NBHS are incredibly computationally expensive to run and also require some approximations at the subgrid level similar to those applied in SAMs. Promising new NBHS are outlined in Vogelsberger et al. (2014) and Schaye et al. (2015). For an extensive comparison of SAMs and NBHS, the reader is referred to Benson et al. (2001), Yoshida et al. (2002), Monaco et al. (2014), and Somerville & Davé (2015).

DM plays an integral role in galaxy formation; broadly speaking, DM haloes are 'cradles' of galaxy formation (Baugh 2006). It is well established that gas cools hierarchically in the centres of DM haloes through mergers; the evolution of galaxies, however, is dictated by a wide variety of baryonic processes that are discussed later in this paper. While baryonic physics plays a crucial role in the outcome of gaseous interactions, the story always starts with gravitational collapse. However, no simple mapping has been found between the internal DM halo properties and the final galaxy properties because of the sheer complexity of the baryonic interactions. For instance, in Contreras et al. (2015), a systematic study of the relationship between the host halo mass and internal galaxy properties is performed. They conclude that no simple mapping was found between the cold gas mass or the star formation rate and the host halo mass. The lack of a relatively simple mapping between internal halo properties and the galaxy properties motivates many of the approximations that SAMs and NBHS make.

Moreover, the computational costs associated with both standard galaxy formation models are incredibly high. The Illustris simulation (an NBHS) used a total of around 19 million CPU hours to run.[4] SAMs, while significantly faster than NBHS, still require an appreciable amount of computational power. For instance, consider the open source GALACTICUS SAM put forth in Benson (2012); in GALACTICUS, a halo of mass $10^{12}\,M_{\odot}$ is evolved (with baryonic physics) in around 2 s and a halo of mass $10^{15}\,M_{\odot}$ is evolved in around 1.25 h. Thus, a very rough order of magnitude estimate can be made for the approximate runtime for GALACTICUS. For about 500 000 DM haloes, and an average evolution time of approximately 2 min (corresponding to about $10^{13}\,M_{\odot}$), the time taken for GALACTICUS to build merger trees to $z = 0$ is $O(15\,000)$ CPU hours. The lack of a simple mapping between DM haloes and the properties of galaxies, the computational costs associated with the popular galaxy formation models, and the highly non-linear nature of the problem make galaxy formation incredibly hard to model, leaving room for new exploration.

While SAMs, not limited to DLB07 and G11, have been incredibly successful in reproducing a lot of observations (White & Frenk 1991; Kauffmann, White & Guiderdoni 1993; Cole et al. 1994, 2000; Somerville & Primack 1999; Kang et al. 2005; Bower et al. 2006; DLB07; Monaco, Fontanot & Taffoni 2007; Lagos, Cora & Padilla 2008; Somerville et al. 2008; Weinmann et al. 2010; De La Torre et al. 2011) and produce similar results to NBHS (Benson et al. 2001; Somerville & Davé 2015), there still exist a few deficiencies in the general methodology of SAMs. Most importantly, the degeneracy inherent to most SAMs is concerning (see, for e.g. Henriques et al. 2009; Bower et al. 2010; Neistein & Weinmann 2010). SAMs (including DLB07 and G11) use simple yet powerful, physically motivated analytical relationships for most processes that play a role in galaxy formation; these processes have several free parameters that are 'tuned' to match up with observations.

An alternative approach to model galaxy formation, that is physically much more transparent, was employed in Neistein & Weinmann (2010, hereafter NW). NW put forth a simple model that includes treatment of feedback, star formation, cooling, smooth accretion, gas stripping in satellite galaxies, and merger-induced starbursts with one key difference compared to conventional SAMs. In the NW model, the efficiency of each physical process is assumed to depend only on the host halo mass and the redshift, making it a much simpler model than G11, DLB07, and other SAMs. NW produces a very similar population of galaxies with similar physical properties to that of DLB07's (G11's predecessor). The success of NW raises an interesting question: could we go even further and try to learn more about the halo–galaxy connection using solely the halo environment and merger history, and would we be able to reproduce results found in conventional SAMs? However, attempting to do so is a non-trivial task for several reasons. First, the inputs for the exploratory model are not exactly clear. Secondly, the mappings between DM halo properties and galaxy properties are incredibly complex, as discussed earlier. NW, G11, and all other SAMs use simplified analytical relationships to capture complex baryonic processes; these relationships have a partial, non-trivial dependence on internal halo properties but it is not clear how these analytical relationships can be used to build a DM-only model to probe galaxy formation and evolution. Given their non-parametric nature and their ability to successfully model complex phenomena, machine learning (ML) algorithms provide an interesting framework to explore this problem.

A variety of statistical techniques, falling under the broad subfield of ML, are gaining traction in the physical sciences. The main goal of ML is to build highly efficient, non-parametric algorithms that attempt to learn complex relationships in and make predictions on large, high-dimensional data sets. Applications of ML to model highly complex physical models include pattern recognition in meteorological models (Liu & Weisberg 2011), particle identification (Roe et al. 2005), inferring stellar parameters from spectra (Fiorentin et al. 2007), photometric redshift estimation (Kind & Brunner 2013), and source classification in photometric surveys (Kim, Brunner & Kind 2015). ML techniques have been shown to be highly effective at picking up complex relationships in high-dimensional data (Witten & Frank 2005; Johnson & Zhang 2011; Graff et al. 2014). As discussed later, we find that ensemble techniques that use a combination of decision trees perform the best in the context of galaxy formation. The relative simplicity of some ML techniques, their high computational efficiency, and powerful predictive capabilities for complex models make the problem of galaxy formation well-suited problem for ML.

In this paper, we present a first exploration into using supervised ML techniques to model galaxy formation. For our analyses, we use the high-resolution Millennium Simulation (Springel 2005; Springel et al. 2005) performed by the Virgo Consortium. The Millennium Simulation is an extremely influential DM simulation that has motivated more than 700 papers in the study of large-scale structure and galaxy evolution. The nature of the Millennium Simulation makes it ideal for galaxy-scale studies. The internal halo properties and a partial merger history for each DM halo at $z = 0$ in the Millennium Simulation are extracted to be used as input features for our algorithms. We use the well-established Munich SAM, G11, for Millennium (Croton et al. 2006; De Lucia et al. 2006; DLB07; G11) for our training and testing. We use various ML algorithms to predict the cold gas mass, hot gas mass, stellar mass in the bulge,

[4] http://www.illustris-project.org/about/

total stellar mass, and the central black hole (BH) mass for the central galaxy of each DM halo in the Millennium Simulation. These components of the mass at $z = 0$ provide an extensive probe into how effective ML algorithms are at learning baryonic processes prescribed in SAMs by using only DM inputs.

It should be emphasized here that absolutely no baryonic processes are explicitly included in our analyses. Only the relevant internal DM halo properties, number of DM particles, spin, $M_{\rm crit200}$, velocity dispersion (σ_v), maximum circular velocity ($v_{\rm max}$), and a partial merger history of the DM halo in which the galaxy resides, are used as the inputs for the algorithms. The results of this study will shed valuable light on the halo–galaxy connection. We can quantitatively, admittedly phenomenologically and not physically, determine the extent of the impact that DM has on the structure formation at smaller scales in the Universe. We can also evaluate how well ML can learn the physical prescriptions that are used in SAMs. To reiterate, our model uses only the 'skeleton' of most galaxy formation models: the merger tree of each DM halo.

The paper is organized as follows. In Section 2, we discuss the data extracted from the Millennium simulation and the basics of ML. More specifically, we outline the basics of the Millennium Simulation and G11 and our reasons for using the Millennium Simulation. We also discuss the basic principles of ML and outline and discuss the best algorithm that was found empirically. In Section 3, we outline our results. In this section, the results for the different components of the central galaxy mass are presented. We also discuss the drawbacks of our model and provide possible explanations for some of the discrepancies in our results and present an alternative ML approach to correct for these discrepancies. In Section 4, we discuss what our results imply about the halo–galaxy connection and SAMs. Finally, in Section 5, we conclude the paper with a summary of our findings and potential avenues for future research.

2 DATA AND BACKGROUND

In this section, we discuss the data set obtained by using the Millennium Simulation and the ML algorithms that were used for the analyses. First, we discuss general details of the Millennium Simulation and the cosmogony that was employed in carrying out the simulation. Next, we discuss the SAM used in Millennium (G11) and give a brief overview of how key physical processes are handled in the model. We also discuss our reasons for choosing the Millennium Simulation in place of a higher resolution simulation with a more accurate cosmogony; in particular, Millennium's σ_8 value is noticeably off from the most recent *Planck* results (Planck Collaboration XIII 2015). We also discuss the extraction of the data set and outline the challenges that were faced in constructing it. Finally, we briefly review how ML works, and outline the primary algorithms that were used in our analyses.

2.1 Millennium Simulation

The data for this project were extracted from the publicly available Millennium Simulation (Springel 2005; Springel et al. 2005). The Millennium Simulation was run with a custom version of GADGET-2, using the TREE-PM method (Xu 1995) to handle gravitational interactions. The Millennium halo catalogues were generated by using a friends-of-friends (FoF) algorithm with a linking length of 0.2 times the mean DM particle separation. The Millennium Simulation is run with 2160^3 DM particles in a $500\,h^{-1}$Mpc box from $z = 127$ to 0. The mass of each DM particle is $8.6 \times 10^8\,{\rm M}_\odot\,h^{-1}$ and the smallest subhalo has at least 20 particles. The cosmological model employed in the Millennium Simulation has $\Omega_{\rm m} = 0.25$, $\Omega_{\rm b} = 0.045$, $\Omega_\Lambda = 0.75$, $h = 0.73$, $n_{\rm s} = 1$, and $\sigma_8 = 0.9$, where the Hubble constant is parametrized as $H_0 = 100$ km s^{-1} Mpc^{-1}.

The raw simulation was sampled in the form of a snapshot 64 times, with FoF group catalogues and their substructures, identified by using SUBFIND, which is discussed in Springel et al. (2001). With SUBFIND, each FoF group is decomposed into a set of subhaloes by identifying locally overdense, gravitationally bound regions. The merger tree organization of the DM haloes is shown in fig. 11 of Springel et al. (2005).

2.2 G11

The SAM used to populate the DM haloes in Millennium with galaxies is described extensively in Springel et al. (2005), Croton et al. (2006), De Lucia et al. (2006), DLB07, and G11. We only provide a brief overview here. G11 includes ingredients and methodologies originally introduced by White & Frenk (1991) and later refined by Springel et al. (2001), De Lucia, Kauffmann & White (2004), and Croton et al. (2006). G11, like most SAMs, has simple, yet physically powerful prescriptions for gas cooling, star formation, supernova feedback, galaxy mergers, and chemical enrichment that are tuned by using observational data. G11 uses the Chabrier (2003) IMF. Additionally, G11 takes into account the growth and activity of central BHs and their effect on suppressing the cooling and star formation in massive haloes. Morphological transformation of galaxies and processes of metal enrichment are also modelled. For a more thorough description of the physical prescriptions used in G11, the reader is referred to the set of papers referenced above. Liu et al. (2010), De La Torre et al. (2011), and Cucciati et al. (2012) show where DLB07 (G11's predecessor) agrees with observational data and also show some weaknesses of DLB07. Furthermore, Knebe et al. (2015) discuss how L-GALXIES (the code behind G11; Henriques et al. 2013) and DLB07 perform against other recent SAMs (e.g. Galacticus, GalICS, etc.) In this section, we provide an overview of how G11 handles the cold gas mass (Section 2.2.1), central BH mass (Section 2.2.2), total stellar mass (Section 2.2.3), bulge mass (Section 2.2.4), and the hot gas mass (Section 2.2.5).

2.2.1 *Cold gas mass*

We outlined in the Introduction how the DM merger history forms the 'skeleton' of SAMs. However, the structure of DM haloes and their internal properties are also important in determining the rate at which gas cools and the dynamics of the galaxies in the halo (Baugh 2006). The cooling of gas in G11 is computed using the growth of the cooling radius $r_{\rm cool}$ as defined in Croton et al. (2006) and G11, which describes the maximum radius at which the hot gas density is still high enough for the cooling to occur within the halo dynamical time $t_{\rm h}$, following the simple model presented in Springel et al. (2001). In G11, it is assumed that infalling gas is shock-heated to the virial temperature ($T_{\rm vir}$) of the DM halo at an accretion shock. The cooling time ($t_{\rm cool}$) at each radius is given by

$$t_{\rm cool} = \frac{3}{2}\frac{\mu m_{\rm p} kT}{\rho_{\rm hot}(r)\Lambda(T_{\rm hot}, Z_{\rm hot})}. \quad (1)$$

Here, $\mu m_{\rm p}$ denotes the mean particle mass, k is the Boltzmann constant, $\rho_{\rm hot}(r)$ is the hot gas density, $\Lambda(T_{\rm hot}, Z_{\rm hot})$ is

the cooling function, that depends on temperature and metallicity (Sutherland & Dopita 1993). $T_{\rm hot}$ is given by

$$T_{\rm hot} = 35.9 \left(\frac{V_{\rm vir}}{\rm km\,s^{-1}}\right)^2. \quad (2)$$

The cooling radius is the point where the local cooling time is equal to $t_{\rm h}$, where

$$t_{\rm h} = \frac{R_{\rm vir}}{V_{\rm vir}} = 0.1 H(z)^{-1}. \quad (3)$$

Therefore, the cooling radius can be written as

$$r_{\rm cool} = \left[\frac{t_{\rm dyn} m_{\rm hot} \Lambda(T_{\rm hot}, Z_{\rm hot})}{6\pi\mu m_{\rm h} k T_{\rm vir} R_{\rm vir}}\right]^{\frac{1}{2}}. \quad (4)$$

The cooling rate can then be written through a simple continuity equation, assuming an isothermal distribution:

$$\dot{M}_{\rm cool} = \frac{1}{2}\frac{M_{\rm hot} r_{\rm cool} V_{\rm vir}}{R_{\rm vir}^2}. \quad (5)$$

A major modification in the cooling rate in equation (5) comes about through 'radio mode' AGN feedback. AGN feedback becomes especially important in haloes with larger masses. G11 follows the prescription laid out in Croton et al. (2006) for the suppression of this cooling rate; the modified rate is given by

$$\dot{M}_{\rm cool}^{'} = \dot{M}_{\rm cool} - 2\frac{\dot{E}_{\rm radio}}{V_{\rm vir}^2}, \quad (6)$$

where

$$\dot{E}_{\rm radio} = 0.1\,\dot{M}_{\rm BH}\,c^2 \quad (7)$$

$$\dot{M}_{\rm BH} = \kappa \left(\frac{f_{\rm hot}}{0.1}\right)\left(\frac{V_{\rm vir}}{200\,\rm km\,s^{-1}}\right)^3 \left(\frac{M_{\rm BH}}{10^8\,{\rm M}_\odot\,h^{-1}}\right) {\rm M}_\odot\,{\rm yr}^{-1}. \quad (8)$$

Here, $f_{\rm hot}$ for a main subhalo is given by the ratio of the hot gas mass to the subhalo mass ($\frac{M_{\rm hot}}{M_{\rm DM}}$) and κ sets the efficiency of the accretion of the hot gas. A more detailed explanation can be found in G11.

The rate of gas cooling is an integral part of galaxy formation because it determines the rate at which stars form in a galaxy (Baugh 2006). As we can see, G11 has a complex recipe to incorporate gas cooling and, while we can see some halo dependence in equations (5) and (6), this is not a trivial mapping. In the NW model discussed in the Introduction, because of the complexity inherent to gas cooling, the coefficient for the cooling rate was empirically determined by running DLB07 on the milli-Millennium Simulation since no fitting function in terms of the host halo mass and redshift was found.

2.2.2 *Central black hole mass*

G11, like Croton et al. (2006), splits AGN activity into 'quasar' mode and 'radio' mode. The formation and evolution of the BH is dominated by the 'quasar' mode feedback, where the central BH grows through major and/or gas-rich mergers. During a merger, the central BH of the larger progenitor absorbs the minor progenitor's BH, and cold gas is accreted on to the central BH. The evolution of the BH mass through 'quasar' mode feedback is given by

$$\delta M_{\rm BH} = M_{\rm BH,min} + f\left(\frac{M_{\rm minor}}{M_{\rm major}}\right)\left(\frac{M_{\rm cold}}{1+\frac{280\,\rm km\,s^{-1}}{V_{\rm vir}}}\right). \quad (9)$$

Here, $M_{\rm BH,min}$ is the mass of the BH in the minor progenitor, f is a free parameter that is set to 0.03 to reproduce the observed BH mass–bulge mass relation, $M_{\rm minor}$ and $M_{\rm major}$ are the total baryonic masses of the minor and major progenitors, and $M_{\rm cold}$ is the total cold gas mass.

2.2.3 *Stellar mass*

G11 follows the Kauffmann (1996) recipe in assuming that star formation is proportional to the mass in cold gas above a certain threshold. A threshold surface density

$$\Sigma_{\rm crit} = 12 \times \left(\frac{V_{\rm vir}}{200\,\rm km\,s^{-1}}\right)\left(\frac{R}{\rm kpc}\right)^{-1} {\rm M}_\odot\,{\rm pc}^{-2} \quad (10)$$

is set for cold gas below which stars do not form and above which stars do form (Kennicutt 1998). Furthermore, it is assumed that this cold gas mass is distributed uniformly over the disc, giving us the following for the critical mass:

$$M_{\rm crit} = 11.5 \times 10^9 \left(\frac{V_{\rm max}}{200\,\rm km\,s^{-1}}\right)\left(\frac{r_{\rm disc}}{10\,\rm kpc}\right) {\rm M}_\odot. \quad (11)$$

Therefore, when the mass of the cold gas in the galaxy is greater than $M_{\rm crit}$, the stars form at the following rate per unit time:

$$\dot{M}_\star = \alpha_{\rm sf}\frac{M_{\rm cold} - M_{\rm crit}}{t_{\rm dyn,disc}}, \quad (12)$$

where the disc dynamical time is given by $t_{\rm dyn,disc} = \frac{r_{\rm disc}}{V_{\rm vir}}$ and $\alpha_{\rm sf}$ is the star formation efficiency, which is manually set between 5 and 15 per cent. In the NW model, a modified star formation rate is used:

$$\dot{M}_\star = f_{\rm s}(M_{\rm cold} - M_{\rm crit}), \quad (13)$$

where $f_{\rm s}$ is the star formation efficiency and has the units of ${\rm Gyr}^{-1}$. The analytic fitting function that NW found for $f_{\rm s}$ was given as

$$f_{\rm s} = M_{\rm h}^{1.04} t^{-0.82} 10^{-6.5-0.0394(\log(M_{\rm h}))^2}, \quad (14)$$

and $M_{\rm crit}$ was parametrized in terms of the host halo mass as

$$M_{\rm crit} = f_{\rm s}^{-1} 10^{-8.61} M_{\rm h}^{0.68} t^{-0.515}. \quad (15)$$

$M_{\rm h}$ has the units of ${\rm M}_\odot\,h^{-1}$ and t is in Gyr. The results presented in NW for stellar mass show good agreement with DLB07 results. The above parametrization of the star formation efficiency and $M_{\rm crit}$ in terms of solely the host halo mass and time and its success offer motivation for our study.

2.2.4 *Bulge mass*

In G11, bulge growth is modelled in three ways: minor mergers, major mergers, and disc instabilities. In the case of minor mergers (i.e. a satellite merging with a central galaxy), the total stellar mass of the satellite galaxy is added to the bulge of the central galaxy and the disc of the larger progenitor remains intact. The cold gas of the satellite galaxy is added to the disc of the central galaxy and a fraction of the combined cold gas from both galaxies is turned into stars as a result of the merger. In the case of major mergers, the discs of both merging galaxies are destroyed to form a spheroid to which the combined stellar mass of the two progenitors is assigned.

G11 uses energy conservation and the virial theorem to calculate the change in size:

$$C\frac{GM_{\rm new,bulge}^2}{R_{\rm new,bulge}} = C\frac{GM_1^2}{R_1} + C\frac{GM_2^2}{R_2} + \alpha\frac{GM_1M_2}{R_1+R_2}. \quad (16)$$

Here, C is a parameter relating the binding energy of a galaxy to its mass and radius, and α is a parameter signifying the effective interaction energy in the stellar components. Furthermore, in G11 a prescription for bulge formation through disc instability is also included. Following the framework set up in Mo, Mao & White (1998), it is assumed that a stellar disc becomes unstable when

$$\frac{V_c}{(\frac{G\,m_d}{r_d})^{\frac{1}{2}}} \leq 1, \tag{17}$$

where V_c is approximated as V_{vir}. At each time step, this inequality is checked and if it is not satisfied, some stellar mass is transferred from the disc to the bulge until stability is restored.

2.2.5 *Hot gas mass*

The total hot gas mass is a result of various physical processes. If we ignore supernovae feedback and gas stripping, we obtain the following amount of hot gas for a halo at each snapshot available for cooling:

$$M_{hot} = f_b M_{vir} - \sum_i M^{(i)}_{cold}. \tag{18}$$

Here, the cold gas mass is summed over in all the galaxies in the FoF group and f_b is the universal baryon fraction, given by 0.017.

Supernova feedback is the main source of reheating incorporated in G11. Supernovae feedback, based on Martin (1999), is modelled in G11 as

$$\delta M_{reheat} = \epsilon_{disc} \times \delta M_\star, \tag{19}$$

where δM_* is the stellar mass of the newly formed stars over some finite time interval. Unlike DLB07, G11 has a variable ϵ_{disc} and is modelled as follows:

$$\epsilon_{disc} = \epsilon \times \left[0.5 + \left(\frac{V_{max}}{70\,km\,s^{-1}}\right)^{-\beta}\right], \tag{20}$$

where both ϵ and β are free parameters which were set in G11 based on the observed stellar mass function.

In most SAMs, when a merger happens, all the hot gas is assumed to be transferred instantaneously from the smaller halo to the larger halo; however, this rapid transfer has been shown to cause a rapid decline in star formation (Baldry et al. 2006; Wang et al. 2007). G11 implements a gas stripping model that includes both the instantaneous stripping and tidal, more gradual stripping in their treatment.

There are two reasons why we chose to use the Millennium Simulation over more recent, higher-resolution simulations with a more accurate ΛCDM cosmogony. First, Millennium is a state-of-the-art cosmological simulation that is uniquely linked to the concurrent development and refinement of two cutting edge SAMs (Bower et al. 2006; Croton et al. 2006). Secondly, and perhaps more importantly, the Millennium Simulation provides a readily accessible data set. The publicly available simulation data enables reproducibility and consistency. In the growing climate of scientific reproducibility, this approach is becoming increasingly important; and we follow this trend and release all our data and code at https://github.com/ProfessorBrunner/ml-sims. Next, we describe how the data set was obtained.

2.3 Data extraction

We used the online SQL data base hosted by GAVO (Lemson et al. 2006) to construct our data set. Using the queryable SQL data base for the Millennium Simulation, we extracted 365 361 DM haloes at $z = 0$. Only DM haloes with masses larger than $10^{12}\,h^{-1}\,M_\odot$ were used in our analysis. For each DM halo, we extracted the following physical properties: number of DM particles ($\mathcal{N}$), spin, $M_{crit200}$, maximum circular velocity (v_{max}), and velocity dispersion (σ_v). For haloes at $z = 0$, we also include the virial mass M_{virial}, the half-mass radius R_{half}, virial velocity v_{virial}, virial radius r_{virial}, and $r_{crit,\,200}$. Furthermore, we extracted the cold gas mass, total stellar mass, stellar mass in the bulge mass, central BH mass, and the hot gas mass of the primary galaxy for each DM halo and matched them likewise. As discussed in the Introduction, the merger history of each DM halo plays an important role in how SAMs populate DM haloes with galaxies. However, sampling the merger history sufficiently and translating that into a well-defined set of inputs for our ML algorithms turns out to be a difficult task.

The naive way to extract the merger history would be to take all the progenitor haloes of each DM halo and use the internal properties of each progenitor halo and the current halo as the inputs to our ML algorithms. However, in the Millennium Simulations, some massive DM haloes have tens of progenitors just one snapshot back, potentially making the number of inputs in the hundreds. Moreover, going just one snapshot back is not sufficient to truly capture the merger history of a DM halo. Consequently, going to higher redshifts would easily result in the number of inputs being in the thousands. Our algorithms' runtimes are directly dependent on the dimension of the input data, and, therefore, employing the naive approach would severely impact the efficiency of our ML algorithms.

The merger tree for the Millennium Simulation includes a descendantID and a firstProgenitorID for each halo. We are left with two ways to sample the merger history while retaining computational efficiency. We can either go top-down (i.e. start at a high redshift and go to $z = 0$) by using descendantID or we could go bottom-up (i.e. starting at $z = 0$ and proceed to higher redshifts) by using the firstProgenitorID. In the Millennium documentation,[5] the first progenitor is defined as 'the main progenitor of the subhalo'. The first progenitor is simply defined as the most massive of each DM halo's progenitors. And thus the firstProgenitorID tracks the main branch of a merger tree. We chose the latter approach for two reasons. The first approach results in fewer haloes and, consequently, galaxies, being examined at $z = 0$. Secondly, the 'main progenitor' approach encodes information about the main branch of a progenitor of each halo, implying that this may provide more valuable information about the DM halo's internal history. Using the firstProgenitorID from $z = 0$ to 5.724 (from snapshot 63 to 19), we extract the five physical properties mentioned above for 365 361 DM haloes.

DM haloes below $10^{12}\,M_\odot$ were not considered in our analyses because the computational cost associated with our technique (given how far back we go in the merger tree) would rise significantly since the number of haloes between just 10^{11} and $10^{12}\,M_\odot$ is a factor of $O(100)$ to $O(1000)$ greater than what we have considered in this work. This amounted to a trade-off between a higher range of masses explored and how much deeper we can go into the merger tree. Since the main point of the paper is to roughly explore the applicability of ML in reproducing a reasonable population of galaxies, we decided against including haloes of lower masses. However, lower mass haloes are included in our analyses in the next paper (Kamdar, Turk & Brunner 2015) in the series where we apply ML to NBHS. We show there that when haloes of all masses are considered and we

[5] http://gavo.mpa-garching.mpg.de/Millennium/Help/mergertrees

see that there is more scatter at lower masses but a reasonably high amount of information is recovered from the DM halo properties using ML throughout.

2.4 Machine learning

In this subsection, we briefly talk about the basics of ML and outline the best performing algorithms.

2.4.1 Overview

ML is a bustling field in computer science, with a wide variety of applications in a number of other areas. The basic idea of ML algorithms is to 'learn' relationships between the input data and the output data without any explicit analytical prescription being used. Supervised learning techniques are provided some training data ($\boldsymbol{X}$, $\boldsymbol{y}$) and they try to learn the mapping $G(\boldsymbol{X} \rightarrow \boldsymbol{y})$ in order to apply this mapping to the test data.

ML has been applied to several subfields in astronomy with a lot of success; see, for example, Ball & Brunner (2010) and Ivezić et al. (2014) for a review of the applications of ML to astronomy. A decent majority of the applications of ML in astronomy have either been in classification problems such as star–galaxy classification (Ball et al. 2006; Kim et al. 2015), galaxy morphology classification (Banerji et al. 2010; Dieleman, Willett & Dambre 2015), or have been regression applications like photometric redshift estimation (Ball et al. 2007; Gerdes et al. 2010; Kind & Brunner 2013), and estimation of stellar atmospheric parameters (Fiorentin et al. 2007).

To the best of our knowledge, however, only a few have applied ML to the problem of galaxy formation and the galaxy–halo connection. Xu et al. (2013), who inspired this paper, predicted the number of galaxies in a DM halo to create mock galaxy catalogues. They used k-Nearest Neighbors (kNN) and Simple Vector Machines to obtain promising results. Furthermore, Ntampaka et al. (2015) used ML for dynamical mass measurements of galaxy clusters also showing promise. Given their non-parametric nature and incredibly powerful predictive capabilities, ML provides an attractive and intriguing method to study galaxy formation and evolution. For our study, we used a variety of ML algorithms: kNN, decision trees, random forests (RF), and extremely randomized trees (ERT). To quantify how well the algorithms are doing at learning relationships in the data, we use the mean-squared error (MSE) metric. The MSE is defined as follows:

$$\mathrm{MSE} = \frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{i=N_{\mathrm{test}}-1} \left(X^i_{\mathrm{test}} - X^i_{\mathrm{predicted}}\right)^2. \tag{21}$$

Here, X^i_{test} is the ith value of the actual test set and $X^i_{\mathrm{predicted}}$ is the ith value of the predicted set. Furthermore, to gauge the relative performance of the algorithms, we also introduce the base MSE, following in the footsteps of Xu et al. (2013), defined as follows:

$$\mathrm{MSE_b} = \frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{i=N_{\mathrm{test}}-1} \left(X^i_{\mathrm{test}} - X_{\mathrm{mean,train}}\right)^2. \tag{22}$$

Here, $X_{\mathrm{mean,train}}$ is, as the name suggests, the mean of the training data set. $\mathrm{MSE_b}$ is an extremely naive prediction of the error since each test point is simply predicted as the mean of the training data set. $\mathrm{MSE_b}$ will serve as an extremely useful metric when we want to measure the relative performance of our ML algorithms and the factor $\frac{\mathrm{MSE_b}}{\mathrm{MSE}}$ will quantitatively show how good our model is at minimizing error. The lower MSE, and consequently the higher factor $\frac{\mathrm{MSE_b}}{\mathrm{MSE}}$, implies a more robust prediction.

Furthermore, we will also be using the following two metrics to check for the robustness of the prediction: the Pearson correlation and the coefficient of determination ('regression score function'). The Pearson correlation is defined as

$$\rho = \frac{\mathrm{cov}(X_{\mathrm{predicted}} X_{\mathrm{test}})}{\sigma_{X_{\mathrm{predicted}}} \sigma_{X_{\mathrm{test}}}}, \tag{23}$$

and the coefficient of determination is defined as

$$R^2 = 1 - \frac{\sum_i (X^i_{\mathrm{test}} - X^i_{\mathrm{predicted}})^2}{\sum_i (X^i_{\mathrm{test}} - X_{\mathrm{mean,train}})^2}. \tag{24}$$

A higher ρ and R^2 imply a robust prediction. As shown in the results section, ERT and RF consistently outperform the other algorithms; consequently, we now briefly review the basics of these two techniques.

2.4.2 Extremely randomized trees

ERT is an ensemble learning technique that builds upon the widely used decision trees (for the purposes of regression, decision trees are called regression trees). Therefore, to understand how ERT works, we must first discuss the fundamentals of regression trees. What follows is only a brief overview of both techniques; for a more comprehensive account of the technique, the reader is referred to Breiman et al. (1984) and Geurts, Ernst & Wehenkel (2006). Regression trees follow a relatively simple procedure.

(i) *Step 1:* construct a node containing all the data points and compute m_c and S, where m_c and S are defined as

$$m_c = \frac{1}{n_c} \sum_{i \epsilon c} z_i \tag{25}$$

$$S(M) = \sum_{c \in \mathrm{leaves}(M)} \sum_{i \in c} (z_i - z_c)^2. \tag{26}$$

Here, c are the possible values of dimension M, z_i gives the target value on each branch c, and z_c gives the mean value on that branch c. We can, therefore, rewrite equation (26) as

$$S(M) = \sum_{c \in \mathrm{leaves}(M)} = n_c V_{\mathrm{c}}. \tag{27}$$

$S(M)$ signifies the sum of the squared errors for some node M,*where* n_c *is the number of samples in a leaf* c *and* $\mathrm{V_c}$ *is the variance in leaf* c.

(ii) *Step 2:* if all the points in the node have the same value for all the input variables, we stop the algorithm. Otherwise, we scan over all dimension splits of all variables to find the one that will reduce $S(M)$ as much as possible. If the largest decrease in $S(M)$ is less than some threshold ϵ, we stop the algorithm. Otherwise, we take that split, creating new nodes of the specified dimension.

(iii) *Step 3:* go to Step 1.

Regression trees are usually considered to be weak learners. A technique that is used to turn these weak learners into strong ones involves building an ensemble of weak learners. In the context of regression trees, there are two popular ensemble methods: boosting and randomization. Since we have a multidimensional output, we focus on randomized ensemble techniques. The essence of ERT is to build an ensemble of regression trees where both the attribute and split-point choice are randomized while splitting a tree node. We provide pseudo-code for the full algorithm in Table 1, which closely follows the algorithm outlined in Geurts et al. (2006). In the

Table 1. An outline of the extremely randomized trees regression algorithm.

Extremely randomized trees
Inputs: a training set S corresponding to (X, y) input–output vectors, where
$X = (X_1, X_2, \ldots, X_N)$ and $y = (y_1, y_2, \ldots, y_l)$, M (number of trees in the ensemble),
K (number of random splits screened at each node), and $n_{\rm min,samples}$ (number of samples required to split a node).
Outputs: an ensemble of M trees: $\mathcal{T} = (t_1, t_2, \ldots, t_M)$.
Step 1: randomly select K inputs $(X_1, X_2, \ldots, X_K)$ where $1 \leq K \leq N$.
Step 2: for each selected input variable X_i in $i = (1, 2, \ldots, K)$ –
(i) compute the minimal and maximal value of X in the set: $X_i^{\rm min}$ and $X_i^{\rm max}$;
(ii) randomly select a cut-point $X_{\rm c}$ in the interval $[X_i^{\rm min}, X_i^{\rm max}]$;
(iii) return the split in the interval $X_i \leq X_{\rm c}$.
Step 3: select the best split s_* such that $\mathrm{Score}(s_*, \mathrm{S}) = \max_{i=1,2,\ldots,K} \mathrm{Score}(s_i, \mathrm{S})$.
Step 4: using s_*, split S into $S^l(X_i)$ and $S^r(X_i)$.
Step 5: for $S^l(X_i)$ and $S^r(X_i)$, check the following conditions –
(i) $\lvert S^l(X_i)\rvert$ or $\lvert S^r(X_i)\rvert$ is lower than $n_{\rm min,samples}$;
(ii) all input attributes $(X_1, X_2, \ldots, X_N)$ are constant in $\lvert S^l(X_i)\rvert$ or $\lvert S^r(X_i)\rvert$;
(iii) the output vector $(y_1, y_2, \ldots, y_l)$ is constant in $\lvert S^l(X_i)\rvert$ or $\lvert S^r(X_i)\rvert$.
Step 6: if any of the conditions in Step 5 are satisfied, stop. We are at a leaf node.
If none of the conditions are satisfied, repeat Steps 1 through 5.

algorithm, the Score is the relative reduction in the variance. For the two subtrees S^l and S^r corresponding to the split s_*, the Score(s_*, S) (abbreviated to Sc(s_*, S)) is given by

$$\mathrm{Sc}(s_\star, S) = \frac{\mathrm{var}(\boldsymbol{y}, S) - \frac{|S^l|}{|S|}\mathrm{var}(\boldsymbol{y}, S^l) - \frac{|S^r|}{|S|}\mathrm{var}(\boldsymbol{y}, S^r)}{\mathrm{var}(\boldsymbol{y}, S)}. \quad (28)$$

The estimates produced by the M trees in the ERT ensemble are finally combined by averaging y over all trees in the ensemble. The use of the original training data set in place of a bootstrap sample (as is done for RF) is done to minimize bias in the prediction. Furthermore, the use of both randomization and averaging is aimed at reducing the variance of the prediction (Geurts et al. 2006).

2.4.3 *Random forests*

The methodology of RF (Breiman 2001) is very similar to that of ERT. There are two central algorithmic differences between the two methods. First, RF use a bootstrap replica (Breiman 1996). Bootstrap replica consists of selecting a random sample without replacement from the training data $(\boldsymbol{X}, \boldsymbol{y})$. ERT, on the other hand, uses the original training data set. Secondly, ERT chooses the split randomly from the range of values in the sample at each split, whereas RF tries to determine the best split at each internal node. We briefly outline the basic steps of the algorithm in Table 2.

For our analyses, we used the implementation of ERT and RF provided in the PYTHON library SCIKIT-LEARN (Pedregosa et al. 2011). The parameters we used and the runtime of the techniques for the problem are discussed in the results section. ERT tends to be faster than RF because of the randomization in finding the split, reducing the training time. The reduced training time lets us build a bigger ensemble of trees and explains why our ERT results are, generally, marginally better than RF's.

Table 2. An outline of the random forests algorithm.

Random forests
Inputs: a training set S corresponding to (X, y) input–output vectors, where $X = (X_1, X_2, \ldots, X_N)$ and $y = (y_1, y_2, \ldots, y_l)$,
M (number of trees in the ensemble),
K (number of features to consider when looking for best split),
$n_{\rm min}$ (minimum number of samples required to split a node).
Outputs: an ensemble of M trees: $\mathcal{T} = (t_1, t_2, \ldots, t_M)$
Step 1: select a new bootstrap sample from the training data set.
Step 2: grow an unpruned tree on this bootstrap.
Step 3: for each node in the tree, randomly select K features, look for the best split using only these features.
Step 4: save the tree and do not perform pruning.
Step 5: perform steps 1 through 4 M times.
Step 6: the overall prediction is the average output from each tree in the ensemble.

3 RESULTS

In this section, we present and discuss the results that were obtained when we applied the algorithms to the Millennium data. Using the DM internal halo properties and a partial merger history as our inputs, the following components of the final mass of the galaxy are predicted: stellar mass in the bulge, total stellar mass, cold gas mass, central BH mass, and hot gas mass. In nature, these attributes are the result of billions of years of dissipative, non-linear baryonic processes. As discussed earlier, the basic ingredient for large-scale structure formation is the ΛCDM model; but, on a smaller scale, the story is incredibly different and vastly more complicated. In this section, we try to draw a link between the two regimes using ML algorithms to explore the halo–galaxy connection. We first discuss

Table 3. The performance of different ML techniques in predicting the different mass components of the central galaxy in each DM halo at $z = 0$.

	Technique	MSE_b	MSE	Factor ($\frac{MSE_b}{MSE}$)	Pearson correlation	R^2
	kNN		6.661	3.624	0.852	0.724
	Decision trees		7.448	3.241	0.832	0.691
	Random forests		5.763	4.301	*0.876*	*0.768*
$M_{\star,\mathrm{total}}$	**Extremely randomized trees**	24.788	*5.755*	*4.307*	*0.876*	0.766
	kNN		7.066	3.906	0.864	0.744
	Decision trees		8.012	3.444	0.843	0.710
	Random forests		*6.305*	*4.467*	*0.881*	*0.775*
$M_{\star,\mathrm{bulge}}$	Extremely randomized trees	28.165	6.306	4.466	*0.881*	*0.776*
	kNN		1243.182	47.786	0.991	0.979
	Decision trees		144.537	411.014	*0.999*	0.998
	Random forests		70.398	929.358	*0.999*	*0.999*
M_{hot}	**Extremely randomized trees**	65 424.910	*57.536*	*1137.113*	*0.999*	*0.999*
	kNN		0.401	1.311	0.487	0.237
	Decision trees		0.445	1.182	0.393	0.155
	Random forests		*0.319*	1.652	*0.631*	*0.395*
M_{cold}	**Extremely randomized trees**	0.527	*0.319*	*1.654*	*0.632*	*0.395*
	kNN		0.000 063	6.958	0.926	0.856
	Decision trees		0.000 081	5.432	0.903	0.815
	Random forests		0.000 068	6.456	0.925	0.856
M_{BH}	**Extremely randomized trees**	0.000 439	*0.000 066*	*6.652*	*0.927*	*0.859*

the performance of ML in reproducing the simulated properties of the galaxies in G11 and the implications of our results for the halo–galaxy connection. Then, we address some discrepancies in our results (particularly the cold gas mass), discuss why the cold gas mass prediction is not robust and provide an alternative, significantly more accurate model that includes two baryonic inputs over two snapshots.

Table 3 lists the results we obtained with the different ML algorithms for each component of the mass of the central galaxy. MSE_b and the MSE are listed for each technique. The factor reduction of the MSE is also listed to test the relative performance of the algorithms to quantify how much they are learning. Finally, the Pearson correlation between the predicted and the true data set and the coefficient of determination (R^2) are also listed. 17 different plots (Figs 2–17) are included to show the best results we obtained by using both ERT and RF. A hexbin plot and a violin plot are shown for each component of the mass to compare the predictions to the G11 test data. A hexbin plot is a 2D histogram and provides information about the goodness of fit. A kernel density estimate (KDE) is plotted on each axis to overlay information about the distribution as well. A violin plot is a box plot with a KDE on the side, providing more information about the distribution of a particular set. Furthermore, a stellar mass–halo mass (SMHM) relation plot is shown to compare the physical reasonability of our stellar mass results with G11 results. A plot showing the cold gas mass fraction as a function of stellar mass is also included. Lastly, the G11 and ML BH mass–bulge mass relation for the predicted galaxies and G11 galaxies is shown in Fig. 10. All plots were created using SEABORN[6] and MATPLOTLIB. For the hexbin plot, a grid size of 30 was used and the colour map was logarithmically scaled. For the KDE in the violin plot, the bandwidth was chosen using the Silverman (1986) method and the density is evaluated on 1000 grid points. The violin plots serve two purposes: providing an insightful look into how good ML is at reproducing a similar population of galaxies as found in G11 and providing a zoomed in alternative of what the mass distribution looks like.

The algorithms that performed the best, ERT and RF, were outlined in Section 2. Using SCIKIT-LEARN's implementation, we used the following parameters for ERT: $n_{\mathrm{trees}} = 750$, and minimum sample split (n_{min}) = 5. For RF, we used the following parameters: $n_{\mathrm{trees}} = 325$, and minimum sample split (n_{min}) = 5. We used 35 per cent of the G11 and Millennium data for training and the rest were used for testing. The entire pipeline for ERT (includes data pre-processing, training, testing, and generating all the plots) using the listed parameters ran on 2.7-GHz Intel Dual-Core Processor in 73 min. Likewise, the entire pipeline for RF using the parameters above ran on the same system in 122 min. Note that in both cases, these times are orders of magnitude smaller than SAM computation times.

3.1 Stellar mass

The first thing to notice in our results is the total SMHM relation. In Behroozi, Conroy & Wechsler (2010) and Moster et al. (2010), the SMHM is extensively studied and compared to a variety of observational data and prevalent empirical halo–galaxy models. The SMHM provides a very powerful tool to check whether our results seem physically meaningful and not just numerically reasonable. We can see in Fig. 1 that the SMHM is reconstructed almost perfectly. The curves for the predicted set and the G11 results line up almost perfectly. One thing to notice here is that our prediction is slightly off for the higher halo masses. Moreover, there is more noticeable scatter in the Millennium SMHM than the reconstructed SMHM. These discrepancies are probably present because the ML algorithms are unable to model extreme cases, a hypothesis which is supported by the hexbin plot in Fig. 2 as well; the galaxies with higher masses ($M_g > 60 \times 10^{10}\,\mathrm{M}_\odot$) are being underpredicted. The SMHM being reproduced strongly implies that ML is able to approximate the mapping between the stellar mass and the halo mass that is prescribed in G11 very well. A subtlety to note here is that ML does not a priori assume that a direct mapping exists between the stellar mass and halo mass like other SMHM studies;

[6] http://stanford.edu/mwaskom/software/seaborn/

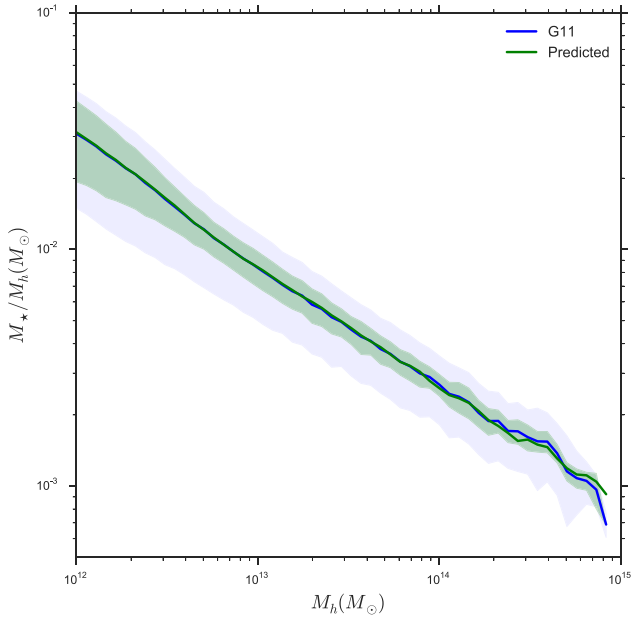

Figure 1. The SMHM relation for the predicted total stellar mass using ML and the total stellar mass in G11 are compared for central galaxies. Both quantities are binned using the halo mass from Millennium. The two different shadings (blue for G11 and green for ML) represent the standard deviation at each binned point for the respective technique.

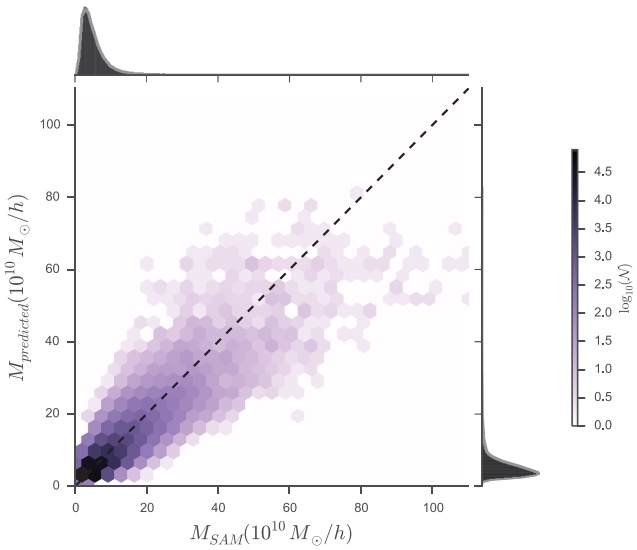

Figure 2. A hexbin plot of $M_{\mathrm{SAM},\star}$ and $M_{\mathrm{predicted},\star}$. The black dashed line corresponds to a perfect prediction. The MSE for the prediction is 5.755, the Pearson correlation between the predicted galaxy mass and the G11 galaxy mass is 0.876 and the regression score is 0.768.

instead, ML is trying to learn the relationship prescribed in G11 for how the galaxies are populated with stellar mass. This point will be important later in the paper when we compare our model with subhalo abundance matching.

Furthermore, we can clearly see in Fig. 2 that the stellar mass is being predicted fairly well. The regression score (R^2) is 0.77 and the correlation between the predicted set and the test G11 set is 0.876. We can see clearly that, while there is some noticeable scatter, a fair majority of the predictions lie on the dashed black line implying that

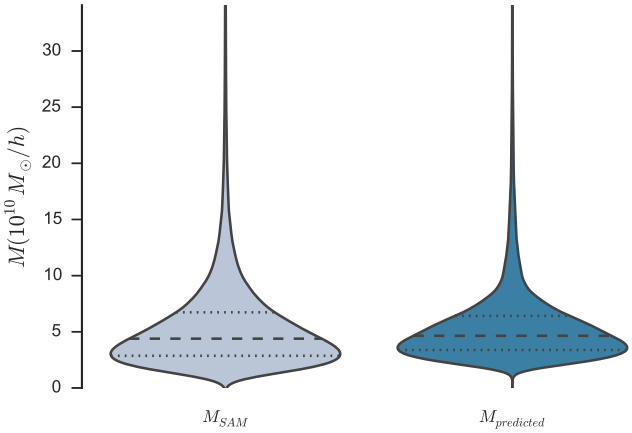

Figure 3. A violin plot is plotted that shows the distributions of the $M_{\mathrm{SAM},\star}$ and $M_{\mathrm{predicted},\star}$. The median and the interquantile range are shown for both sets of galaxy masses.

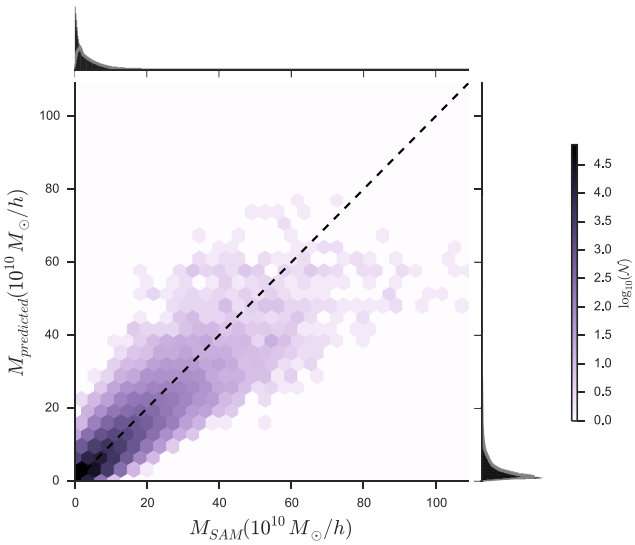

Figure 4. A hexbin plot of $M_{\mathrm{SAM,bulge}}$ and $M_{\mathrm{predicted,bulge}}$. The black dashed line corresponds to a perfect prediction. The MSE for the prediction is 6.305, the Pearson correlation between the predicted galaxy mass and the G11 galaxy mass is 0.881 and the regression score is 0.775.

our predictions line up with that of G11's fairly well. Moreover, and perhaps more surprisingly, we can see in Fig. 3 that the distribution of the stellar mass is reproduced perfectly using ML. Our model is able to pick up on the physical prescriptions that are used in G11 to populate galaxies with stars very well.

3.1.1 Bulge mass

We also predict the stellar mass that is in the bulge of each central galaxy at $z = 0$. In Figs 4 and 5, we can see that the bulge mass is also being accurately reproduced. The regression score is 0.775 and the correlation between the predicted set is 0.881. The hexbin plot shows that the bulge mass prediction is appreciably robust and very similar to the total stellar mass prediction, while the violin plot shows that the distribution of the bulge mass is also reproduced very well. The bulge mass is accumulated in G11 through two primary

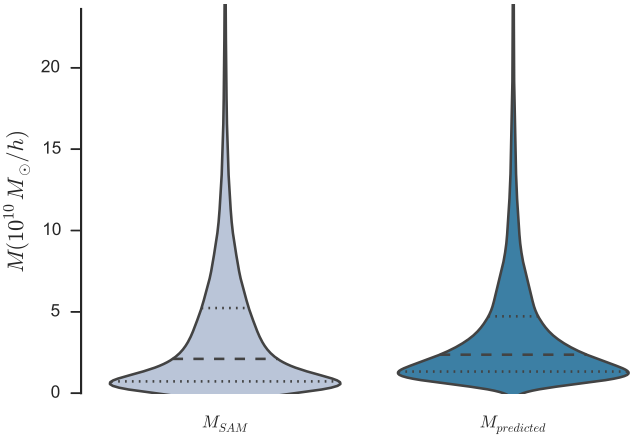

Figure 5. A violin plot showing the distributions of the $M_{\rm SAM,bulge}$ and $M_{\rm predicted,bulge}$. The median and the interquartile range for both sets of galaxy masses are shown.

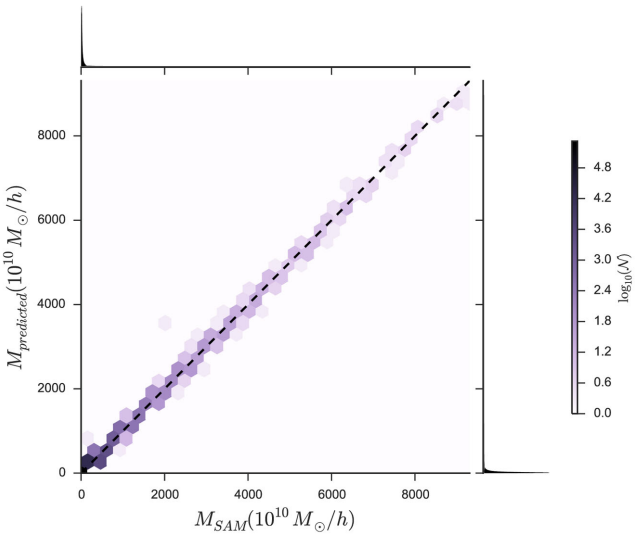

Figure 6. A hexbin plot of $M_{\rm SAM,hot}$ and $M_{\rm predicted,hot}$. The black dashed line corresponds to a perfect prediction. The MSE for the prediction is 57.536, the Pearson correlation between the predicted galaxy mass and the G11 galaxy mass is 0.999 and the regression score is 0.999.

mechanisms: mergers and disc instabilities. The reproduction of the distribution of the bulge mass implies that both these physical prescriptions in G11 are being approximately well modelled by ML.

However, there is a small discrepancy between the distribution of the true and the predicted set (Fig. 5); the predicted distribution shows that ML is overpredicting at lower masses. A possible explanation for this discrepancy is an overprediction of the stellar mass accumulated in the bulge in the event of mergers. We input the merger history of each DM halo into our model and not the galaxy–galaxy merger history. As outlined in Hopkins et al. (2010), there are two ways to link the two regimes: defining a delay time-scale, as discussed in Boylan-Kolchin, Ma & Quataert (2008), or tracking the subhaloes directly (only feasible for higher resolution simulations). Our robust predictions imply that ML is able to extrapolate a reasonable amount of information about galaxy–galaxy mergers through a halo-only merger history. However, the overprediction could easily be explained by ML's inability to fully pick up on the galaxy–galaxy merger time-scale by using only a halo merger history. This idea is further supported by our prediction for the central BH mass discussed in Section 3.3. Overall, our model is able to pick up on the physical prescriptions that are used in G11 to model bulge formation with a minor discrepancy that has no trivial solution in the framework of our model.

3.2 Hot gas mass

The hot gas mass prediction, as shown in Figs 6 and 7 and Table 3, is outstanding. The Pearson correlation is 0.999 and the regression score is 0.999. ML is able to pick up on the way that hot gas is modelled in G11 incredibly well. As discussed in Section 2.2.5, the amount of hot gas available at each snapshot is directly dependent upon the total virial mass in the DM halo. Even though supernovae feedback plays an important role in reheating some of the gas found in the halo, ML is still able to pick up on how the prescriptions for the hot gas mass in G11 are set. The distribution of the hot gas mass is reproduced perfectly and the MSE is reduced by a factor of 1137.

As discussed in Section 2.2.5, the main contributors to the hot gas mass in the central galaxy are gas stripping and supernovae feedback. Our almost perfect results show that ML is able to model these two physical prescriptions very well. The former involves hot gas being stripped from a satellite galaxy and being added to the

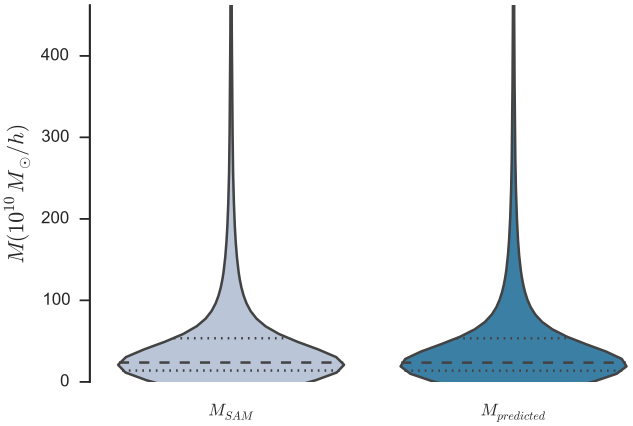

Figure 7. A violin plot showing the distributions of the $M_{\rm SAM,hot}$ and $M_{\rm predicted,hot}$. The median and the interquartile range for both sets of galaxy masses are shown.

central galaxy for cooling. The latter, which plays a larger role in the determination of the hot gas mass, involves the reheating of cold gas due to supernova feedback. As outlined in equations (19) and (20), the amount of mass that is reheated has a partial dependence, both directly (in $\epsilon_{\rm disc}$) and indirectly (in δM_*), on halo properties and ML is able to model both quite well. The amount of hot gas mass plays an important role in the cooling (equation 5). The almost perfect predictions for the hot gas mass are promising and show ML's strength in modelling a mapping that is dominated by a direct analytical relationship (equation 18).

3.3 Central black hole mass

As discussed in Section 2.2.2, the central BH mass is mostly accreted through the 'quasar' mode (equation 9) during major mergers or gas-rich mergers. Our results for the central BH mass are shown in Table 3 and Figs 8 and 9. The regression score is 0.86 and the

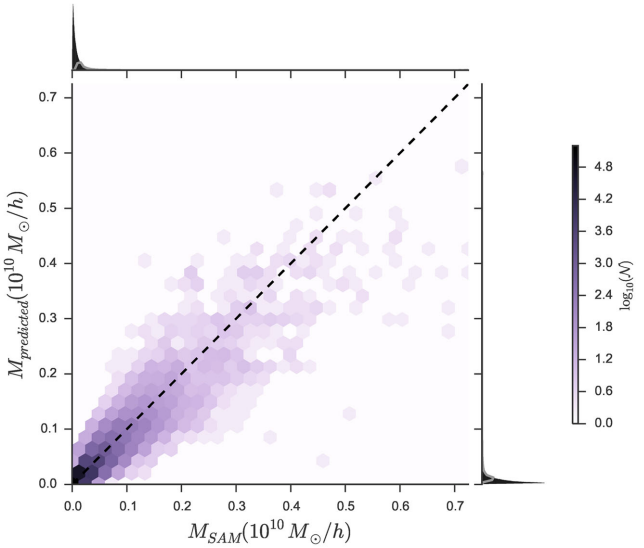

Figure 8. A hexbin plot of $M_{\rm SAM,BH}$ and $M_{\rm predicted,BH}$. The black dashed line corresponds to a perfect prediction. The MSE for the prediction is 0.000 066, the Pearson correlation between the predicted galaxy mass and the G11 galaxy mass is 0.927, and the regression score is 0.86.

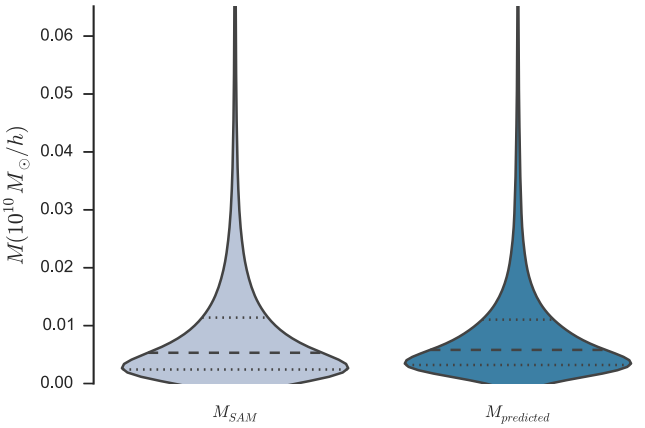

Figure 9. A violin plot showing the distributions of $M_{\rm SAM,BH}$ and $M_{\rm predicted,BH}$. The median and the interquartile range for both sets of galaxy masses are shown.

correlation is 0.929, implying that our predictions match up well with the G11 data. An interesting point here is the overprediction of the BH mass at lower masses shown in Fig. 9. This overprediction places confidence in the theory that the ML may only be partially able to pick up on the galaxy–galaxy merger time-scale that was outlined in the bulge mass discussion.

In Fig. 10, we show the predicted and the G11 BH mass–bulge mass relation. At higher bulge masses, the predicted and the G11 curves agree very well. For lower masses, there is a discrepancy between the bulge masses at which the curve starts. The predicted curve starts at a noticeably higher mass because of the overprediction of the bulge mass that was discussed earlier (i.e. the curve starts at the bin that corresponds to the lowest predicted bulge mass). Overall, the prediction for the central BH mass and the reproduction of the BH–bulge relation reinforce the solidity of ML's predictive power in the context of galaxy evolution. Furthermore, the robust BH mass predictions imply that the internal DM halo properties

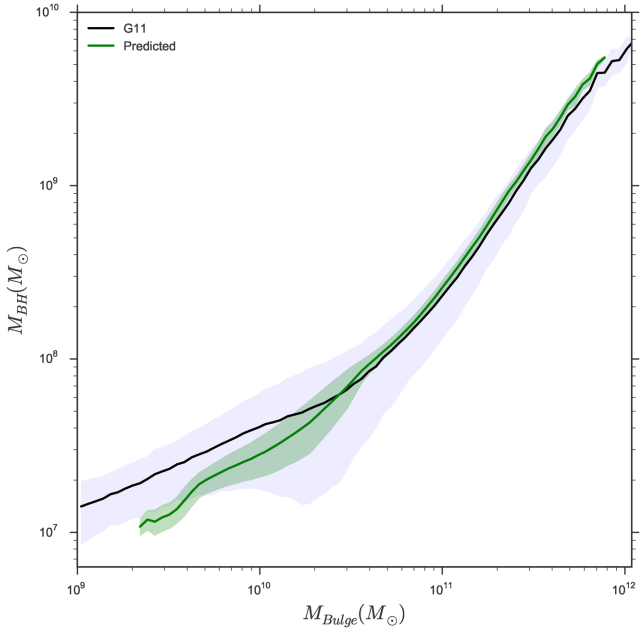

Figure 10. The BH mass–bulge mass relation is plotted on a log scale for the predicted population of galaxies and G11's population of galaxies. For the predicted curve, $M_{\rm BH,predicted}$ points are binned by the predicted bulge mass and for G11, $M_{\rm BH,SAM}$ is binned by the corresponding G11 bulge mass. The shaded areas correspond to the standard deviation at each binned point (blue for G11 and green for ML).

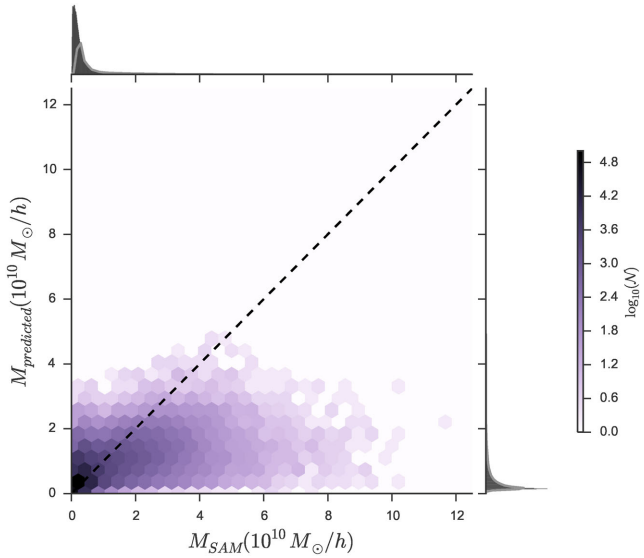

Figure 11. A hexbin plot of $M_{\rm SAM,cold}$ and $M_{\rm predicted,cold}$. The black dashed line corresponds to a perfect prediction. The MSE for the prediction is 0.319, the Pearson correlation between the predicted galaxy mass and the G11 galaxy mass is 0.632, and the regression score is 0.40.

and the merger history contain sufficient information to make robust predictions of the central BH mass.

3.4 Cold gas mass

As shown in Fig. 11, the cold gas mass is being visibly underpredicted. This underprediction is unfortunately not surprising. The recipe used in G11, outlined in Section 2.2.1, has a partial halo

Table 4. Cold gas mass prediction with the inclusion of cooling radius and hot gas mass as inputs.

	Technique	MSE_b	MSE	Factor ($\frac{MSE_b}{MSE}$)	Pearson correlation	R^2
	Random forests		*0.319*	1.652	0.631	*0.395*
$M_{cold,w/o}$	**Extremely randomized trees**	0.527	*0.319*	*1.654*	*0.632*	*0.395*
	Random forests		*0.0972*	*5.422*	*0.905*	*0.816*
$M_{cold,with}$	Extremely randomized trees	0.527	0.113	4.664	0.892	0.786

dependence but the baryonic processes play a far more important role in determining the cold gas mass. Indeed, NW found no easy way to parametrize the efficiency of cooling rate in terms of the host halo mass and time and had to empirically estimate this value by running DLB07 on the mini-Millennium Simulation. To investigate possible reasons for the underprediction, recall that the cooling rate is heavily dependent upon the cooling radius. By using the same algorithms and the same inputs, we predicted the cooling radius at $z = 0$ for the central galaxy. Our results are shown in Figs 14 and 15. We obtained a regression score of 0.86 and a correlation of 0.931; thus, the prediction is fairly robust. The cooling radius, as shown in equation (4), also has only a partial halo dependence. It is remarkable that our prediction is so accurate, then, using solely DM halo information; one would expect that baryonic recipes, for instance, the cooling function prescribed in Sutherland & Dopita (1993), play a far larger role in the determination of the cooling radius.

The reproduction of the cooling radius and the robust hot gas mass prediction discussed earlier raise the following question: why is the cold gas mass evolution not being captured? ML is able to predict the two basic ingredients of gas cooling well, but the cooled mass itself is not being robustly predicted. We hypothesize that this discrepancy is a result of the variability in the cooling radius. Without some form of explicit inclusion of the time evolution of hot gas mass and cooling radius over snapshots, ML is unable to capture the evolution of the cooling radius and, consequently, the accumulated cold gas mass. We tested this hypothesis by including the cooling radius and the hot gas mass over only two snapshots ($z = 0$ and 0.012) in our inputs for ERT and RF and repeating the cold gas mass prediction with the same parameters.

Our results are presented in Figs 16 and 17 and Table 4. By just including these four additional inputs, our predictions became significantly more robust. Our regression score more than doubled and we can see in Fig. 17 that the distribution from G11 is being fairly reproduced. The comparison of the two distributions (Figs 12 and 17) shows that the prescription used for evolving the cooling Ordinary Differential Equation (ODE) is being fairly well picked up by the additional baryonic inputs included. This result is particularly striking because the cooling ODE was evolved 20 times between each snapshot in the Munich SAMs (De Lucia et al. 2010; Knebe et al. 2015) and we included results only at the two extreme points. The improved results imply that ML is strong at extrapolating the underlying physical process in G11 for cooling only if baryonic ingredients are partially, explicitly included in the inputs to guide the algorithms. Furthermore, our results also instil confidence in our hypothesis that ML is unable to predict the accumulated cold gas mass because of its inability to pick up on the cooling evolution using solely DM inputs.

Moreover, from both Table 3 and Figs 11 and 12, we can see that the ML algorithms are definitely learning partial information about how cold gas is being accumulated in G11. The distribution, while not a perfect fit by any means, does reproduce a narrower peak at about the same mass and the median and interquartile range

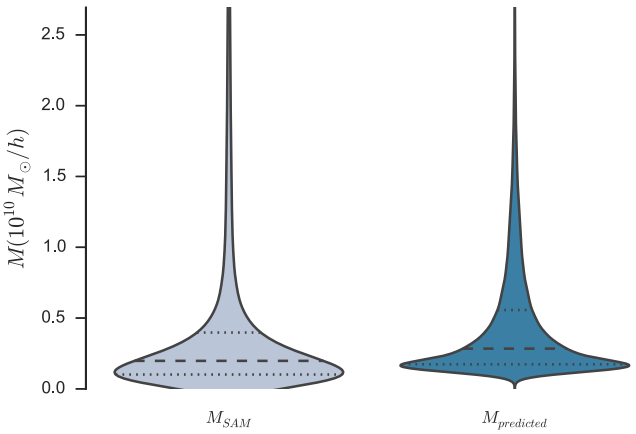

Figure 12. A violin plot showing the distributions of $M_{SAM,cold}$ and $M_{predicted,cold}$. The median and the interquartile range for both sets of galaxy masses are shown.

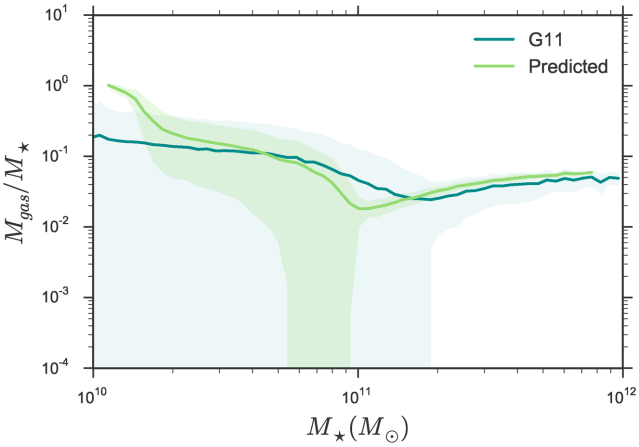

Figure 13. The average cold gas mass fraction as a function of stellar mass is shown for G11 galaxies and ML galaxies. For G11, $\frac{M_{cold,SAM}}{M_{\star,SAM}}$ points are binned by $M_{\star,SAM}$ and for ML, $\frac{M_{cold,predicted}}{M_{\star,predicted}}$ are binned by $M_{\star,predicted}$.

are appreciably similar. Moreover Table 3 shows that the Pearson correlation between the predicted and the test set is 0.63 and R^2 is 0.40. We also see in the hexbin plot that the densest hexbins lie on the straight line, but there is noticeable scatter for higher cold gas masses. Furthermore, in Fig. 13 , we show the average cold gas fraction as a function of stellar mass for both G11 galaxies and ML galaxies. There are some discrepancies between our results and G11's at the beginning, but the general progression of the two curves matches up quite well. Our relatively poor results for the cold gas mass confirm previous results in literature by demonstrably and quantitatively showing the absence of a relatively simple mapping between cold gas mass and DM halo properties.

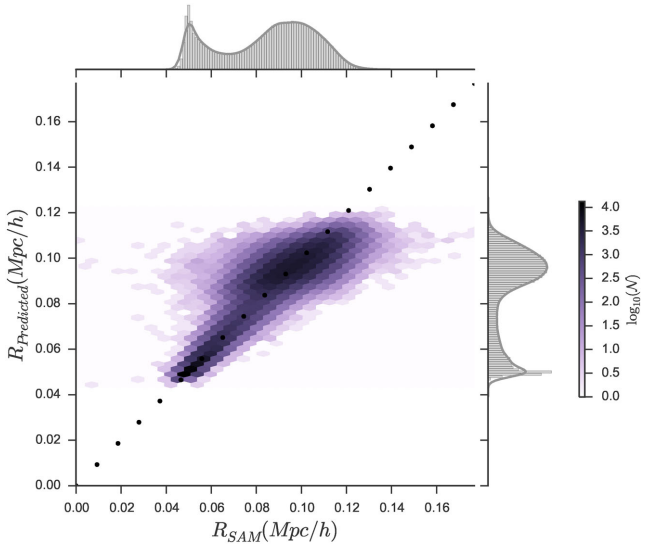

Figure 14. A hexbin plot of the cooling radii $R_{\rm SAM,cool}$ and $R_{\rm predicted,cool}$. The black dashed line corresponds to a perfect prediction. The MSE for the prediction is 0.000 059, the Pearson correlation between the predicted galaxy mass and the G11 galaxy mass is 0.930, and the regression score is 0.87.

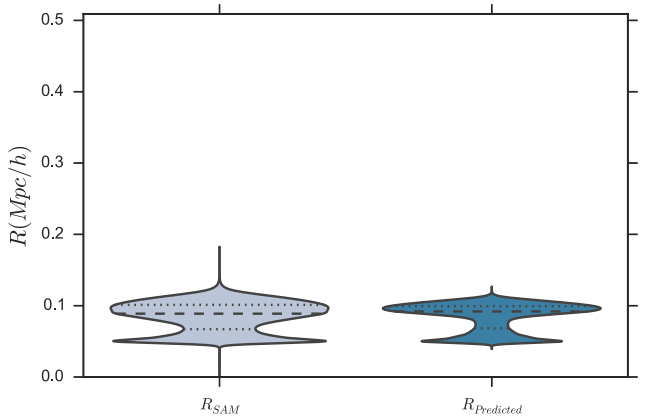

Figure 15. A violin plot showing the distributions of the $R_{\rm SAM,cooling}$ and $R_{\rm predicted,cooling}$. The median and the interquartile range for both sets of cooling radii are shown.

While the main point of this study was to explore the halo–galaxy connection in a DM-only context, we are able to show the power of ML in modelling the complicated cold gas mass recipe reasonably well when very crude, partial baryonic ingredients are included as inputs. The analysis of cold gas mass accumulation using both DM-only inputs and with baryonic inputs raises two interesting points; first, ML, by itself, is unable to pick up on the baryonic evolution involved in gas cooling using solely DM inputs. Secondly, with just the addition of four baryonic inputs, the prediction for the cold gas mass is vastly improved. The results above show that there is more room for exploration and that the relatively poor cold gas mass prediction does not undermine the overall usefulness of ML as a solid tool in exploring the problem of galaxy formation.

4 GENERAL DISCUSSION

The results above show that ML is able to recreate a population of galaxies that is strikingly similar to that of G11's in our DM-only

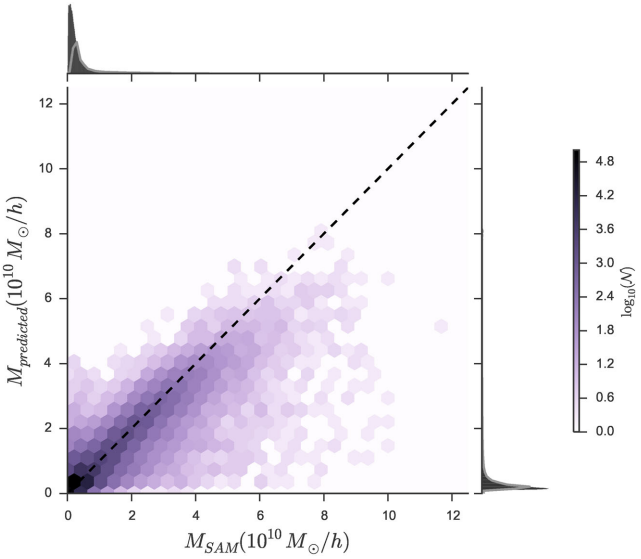

Figure 16. A hexbin plot of $M_{\rm SAM,cold}$ and $M_{\rm predicted,cold}$. The black dashed line corresponds to a perfect prediction. The MSE for the prediction is 0.097, the Pearson correlation between the predicted galaxy mass and the G11 galaxy mass is 0.905, and the regression score is 0.82. The black dotted line corresponds to a perfect prediction. The main difference between this figure and Fig. 11 is that the cooling radius and the hot gas mass for two snapshots are explicitly included in the inputs for the ML algorithms.

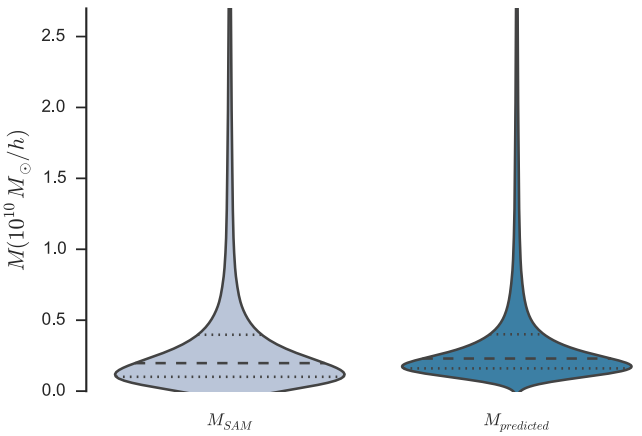

Figure 17. A violin plot showing the distributions of $M_{\rm SAM,cold}$ and $M_{\rm predicted,cold}$. The median and the interquartile range for both sets of galaxy masses are shown. The main difference here is that the cooling radius and the hot gas mass for two snapshots are explicitly included in the inputs for the ML algorithms.

framework. While the reduced MSEs we found for the different components of the mass are surprisingly low, they are still high enough to merit a discussion of the sources of error. First, and most important, a reason for the relatively high MSE is the absence of any baryonic processes or results being input into our ML algorithms. In NW, only the efficiency of the physical processes were modelled by using the host halo mass and redshift; NW did include simplified, but still physically motivated, baryonic processes in the model, with hand-tuned free parameters. Our model, on the other hand, is not grounded on physical motivations related to baryons at all; it is instead an effort to explore the halo–galaxy connection in the framework of SAMs by using solely halo properties and a partial merger history. Another possible reason for the relatively high MSE

may be a result of the fact that we are only looking at the mass of the central galaxy of a halo and ignoring the satellite galaxies, while using the inputs for the entire halo. This may also explain some of the scatter we see in our results for the stellar mass and the cold gas mass.

There are certain deficiencies in our model. As mentioned earlier, the model is not motivated by baryonic physics like most SAMs. G11 and other SAMs offer simple, but incredibly illuminating treatments of interactions at the galactic scale. ML, on the other, by its very nature, does not leave room for exploration into the subtleties of baryonic physics and is a purely phenomenological model. Moreover, the predictions shown above do not imply that ML is a replacement to SAMs. Our results imply instead that ML offers an interesting and promising avenue for exploration in the domain of galaxy formation, primarily because of its simplicity, efficiency, and its ability to provide a unique platform that allows us to probe how much information can be extracted from just DM haloes. In the case of the stellar mass, BH mass, and hot gas mass, ML is able to pick up on the physical prescriptions used by G11 very well by using solely a partial merger history and information about the halo environment. In the case of a cold gas mass, ML is not able to pick up on the evolution of gas cooling by itself; with a partial inclusion of a couple of baryonic inputs over two snapshots, however, we are able to double our regression score and make predictions that are vastly more robust. The improved results place confidence in the predictive power of ML and imply that ML may be a useful tool for future studies of NBHS and other problems in theoretical astrophysics.

An interesting point here is the superficial similarity between our model and suhalo abundance matching (SHAM; Conroy & Wechsler 2009). Both models use halo information to glean physical information about the galaxies residing in the halo. Our study differs from SHAM in one very key aspect: SHAM involves populating haloes with galaxies assuming that there already exists a monotonic relationship between halo mass and galaxy stellar mass (or luminosity). On the other hand, our model predicts properties of galaxies that have already been populated using a SAM (G11) with no relationship being fed to the algorithms. The results obtained, then, imply that the key assumption in most SHAM that observable properties of galaxies are monotonically related to the dynamical properties of DM substructures is partially valid. The discrepancy in our cold gas mass result implies that the baryonic physics plays a vastly more important role in the cooling rate than the halo environment itself. But the reproduction of the total stellar mass implies that the SHAM assumptions in the context of stellar mass hold true and instil further confidence into the general methodology of SHAM.

The cold gas mass prediction raises several interesting points. First, and most important, similar to Contreras et al. (2015), only a weak mapping between the cold gas mass and the internal halo properties was found. However, the robust prediction of the cooling radius and the hot gas mass does leave the door open for further exploration into modelling cold gas mass using ML. Our results also quantitatively verify what NW found; they were unable to parametrize the cooling efficiency in terms of host halo mass and the redshift. By the inclusion of just the cooling radius and hot gas mass over two snapshots, ML is able to vastly improve upon the DM-only predictions. The improved predictions naturally raise the question: can ML be applied to other, more complicated evolutionary models and still reproduce physically and numerically reasonable results?

Our results give a unique and deeper look into the galaxy–halo connection that is grounded upon SAMs. We are able to quantitatively estimate the amount of information that DM haloes and merger trees hold about the baryonic processes that drive galaxy formation and evolution, in the context of SAMs and how SAMs populate haloes with galaxies. We have quantitatively shown that the environmental dependence of galaxy evolution on the surrounding DM halo is surprisingly strong. As mentioned earlier, our robust predictions for the total stellar mass, stellar mass in the bulge, BH mass, and the hot gas mass strongly imply that it is possible to learn the physical processes used in cutting edge SAMs to evolve these components by using solely DM properties, a merger history, and ML. The relatively weaker results for the cold gas mass imply that our phenomenological, DM-only model fails in reproducing the cold gas mass evolution in galaxies. However, our improved cooling model with baryonic inputs solidifies the viability of ML in future galaxy formation studies.

Overall, we get somewhat surprising results since one would expect that gaseous interactions play a significantly more important role in predicting the final components of mass of a single galaxy than just the basic DM halo model; but, we have shown that ML provides a unique and fairly robust avenue to quantitatively analyse the role that just the DM haloes play in galaxy formation in the context of SAMs. We showed that the SMHM relation is reproduced almost perfectly, the shapes of the predicted and true distributions of the different mass components are very similar, the BH mass–bulge mass relation is reproduced, and the cold gas mass fraction as a function of stellar mass is reasonably reproduced. ML is able to learn an appreciable portion of the physical prescriptions used in G11 for galaxy formation using solely DM inputs. Moreover, the amount of time it took to run the whole pipeline took about three hours, considerably less than the hundreds or thousands of hours a typical SAM would require.

5 CONCLUSIONS

We have performed an extensive study of the halo–galaxy connection by using novel ML techniques in the backdrop of a state-of-the-art SAM. Using G11 to train our ML algorithms, the total stellar mass, stellar mass in the bulge, cold gas mass, and hot gas mass in the Millennium Simulation are predicted. ML provides a powerful framework to explore the problem of galaxy formation in part due to its relative simplicity, computational efficiency, and its ability to model complex physical relationships. The discrepancies in and weaknesses of our phenomenological model were discussed and the reasons for some of our relatively less robust predictions were also discussed. An improved cooling model with four additional baryonic inputs was implemented, which made the cold gas mass predictions significantly better and solidified ML's position as a model that can be used to probe the halo–galaxy connection in more detail, perhaps with sophisticated NBHS.

Our primary conclusions are as follows.

(1) Exploring the extent of the influence of DM haloes and its past environment on galaxy formation and evolution is a non-trivial problem with poorly defined inputs and mappings. SAM is the prevalent galaxy formation modelling technique that uses simple, yet physically powerful, recipes to populate DM haloes with galaxies. However, there is no clear way to explore the extent of the influence of DM haloes on the halo–galaxy solely by using just SAMs. ML, on the other hand, provides an interesting alternative to standard techniques for three main reasons: powerful predictive capabilities, simplicity, and efficiency.

(2) By using the Millennium Simulation and G11, we set up a model that used internal halo properties ($\mathcal{N}$, spin, $M_{\rm crit200}$, $v_{\rm max}$, and

σ_v) and a partial merger history to predict different mass components of the central galaxy in each DM halo at $z = 0$. No baryonic processes were incorporated in our initial analysis. We applied several sophisticated algorithms (kNN, regression trees, RF, and ERT) to the Millennium data and we were able to reproduce a similar galaxy population.

(3) The total stellar mass and the stellar mass in the bulge are predicted very well. The predicted and true distributions for both are almost identical. ML is able to model the physical prescriptions laid out in G11 for galactic stellar mass evolution. The SMHM relation that G11 found is recreated almost perfectly with some very minor discrepancies for $M_{\rm h} \approx 10^{15}\,{\rm M}_{\odot}\,h^{-1}$. The bulge mass prediction is also fairly robust, with the distributions being remarkably consistent. However, the bulge mass is slightly overpredicted for lower masses. We hypothesize that this may be because our inputs include a partial merger history of the haloes and not galaxies. Consequently, ML is possibly overpredicting as a result of its inability to fully model the galaxy–galaxy merger time-scale using only a halo merger history. The central BH mass prediction is also very robust and the distribution is recreated almost perfectly. The BH–bulge mass relation for the ML simulated galaxies and G11 galaxies is very consistent. There is a slight overprediction for lower masses for the central BH mass prediction, which further places confidence in the hypothesis that ML is unable to fully pick up on the galaxy–galaxy merger time-scale using only a halo merger history.

(4) The hot gas mass is predicted outstandingly well. ML is demonstrably able to model G11's prescriptions for gas stripping and supernovae feedback. The cold gas mass prediction, on the other hand, is relatively weak with a correlation of only 0.63. However, the robust cooling radius and the hot gas prediction imply that the ingredients for cooling are sufficiently modelled by ML. We hypothesized that our poor prediction was a result of the inability of ML to model the cooling radius evolution without any baryonic guidance (i.e. by using solely DM inputs and merger history). We tested this hypothesis by including the cooling radius and the hot gas mass for only the last two snapshots and found significantly better predictions with a correlation of 0.91 and R^2 of 0.82. The improved cooling predictions place confidence in the predictive power of ML and imply that ML will be a useful tool in future studies in galaxy formation and evolution. The average cold gas mass fraction as a function of stellar mass was also plotted for G11 galaxies and the predicted galaxies. The shape of the two curves is reasonably similar with a minor discrepancy at the lowest masses.

(5) Our results provide a unique framework to explore the galaxy–halo connection that is built by using SAMs. We are able to quantitatively estimate the amount of information that DM haloes and merger trees hold about the baryonic processes that drive galaxy formation and evolution, in the context of G11. Our robust predictions for the total stellar mass, stellar mass in the bulge, central BH mass, and the hot gas mass strongly imply that it is possible to successfully model the physical prescriptions used in SAMs to evolve these mass components using solely DM properties, a merger history, and ML. However, ML is unable to find a robust approximate mapping between the internal DM halo properties and the cold gas mass, like NW, Faucher-Giguère, Kereš & Ma (2011), and Contreras et al. (2015).

(6) ML is a phenomenological model and not a physical one, and, consequently, is not a replacement for SAMs. However, ML offers a solid and intriguing framework to explore the halo–galaxy connection with solid results comparable to G11, which conventional modelling techniques do not provide.

The results presented in this paper show the usefulness of ML in providing a solid framework to probe the halo–galaxy connection in the backdrop of SAMs. Future work includes exploring more sophisticated ML techniques to probe galaxy formation and evolution in NBHS.

ACKNOWLEDGEMENTS

The authors thank Christopher Chan, Rishabh Jain, and Dingcheng Yue for help in gathering data and exploring preliminary ML approaches. HMK and RJB acknowledge support from the National Science Foundation Grant no. AST-1313415. HMK has been supported in part by funding from the LAS Honors Council at the University of Illinois and by the Office of Student Financial Aid at the University of Illinois. RJB has been supported in part by the Center for Advanced Studies at the University of Illinois. MJT is supported by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through Grant GBMF4561. We would like to thank the reviewer for their helpful comments that made this paper better.

The Millennium Simulation data bases used in this paper and the web application providing online access to them were constructed as part of the activities of the German Astrophysical Virtual Observatory (GAVO).

REFERENCES

Angulo R., Springel V., White S., Jenkins A., Baugh C., Frenk C., 2012, MNRAS, 426, 2046
Baldry I. K., Balogh M. L., Bower R., Glazebrook K., Nichol R. C., Bamford S. P., Budavari T., 2006, MNRAS, 373, 469
Ball N. M., Brunner R. J., 2010, Int. J. Mod. Phys. D, 19, 1049
Ball N. M., Brunner R. J., Myers A. D., Tcheng D., 2006, ApJ, 650, 497
Ball N. M., Brunner R. J., Myers A. D., Strand N. E., Alberts S. L., Tcheng D., Llorà X., 2007, ApJ, 663, 774
Banerji M. et al., 2010, MNRAS, 406, 342
Baugh C. M., 2006, Rep. Prog. Phys., 69, 3101
Behroozi P. S., Conroy C., Wechsler R. H., 2010, ApJ, 717, 379
Benson A. J., 2012, New Astron., 17, 175
Benson A., Pearce F., Frenk C., Baugh C., Jenkins A., 2001, MNRAS, 320, 261
Blumenthal G. R., Faber S., Primack J. R., Rees M. J., 1984, Nature, 311, 517
Bower R., Benson A., Malbon R., Helly J., Frenk C., Baugh C., Cole S., Lacey C. G., 2006, MNRAS, 370, 645
Bower R., Vernon I., Goldstein M., Benson A., Lacey C. G., Baugh C., Cole S., Frenk C., 2010, MNRAS, 407, 2017
Boylan-Kolchin M., Ma C.-P., Quataert E., 2008, MNRAS, 383, 93
Breiman L., 1996, Mach. Learn., 24, 123
Breiman L., 2001, Mach. Learn., 45, 5
Breiman L., Friedman J., Stone C. J., Olshen R. A., 1984, Classification and Regression Trees. CRC Press, Boca Raton, FL
Chabrier G., 2003, PASP, 115, 763
Cole S., Aragon-Salamanca A., Frenk C. S., Navarro J. F., Zepf S. E., 1994, MNRAS, 271, 781
Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, MNRAS, 319, 168
Conroy C., Wechsler R. H., 2009, ApJ, 696, 620
Contreras S., Baugh C., Norberg P., Padilla N., 2015, MNRAS, 452, 1861
Croton D. J. et al., 2006, MNRAS, 365, 11
Cucciati O. et al., 2012, A&A, 548, A108
Davis M., Efstathiou G., Frenk C. S., White S. D., 1985, ApJ, 292, 371
De La Torre S. et al., 2011, A&A, 525, A125
De Lucia G., Blaizot J., 2007, MNRAS, 375, 2 (DLB07)
De Lucia G., Kauffmann G., White S. D., 2004, MNRAS, 349, 1101

De Lucia G., Springel V., White S. D., Croton D., Kauffmann G., 2006, MNRAS, 366, 499
De Lucia G., Boylan-Kolchin M., Benson A. J., Fontanot F., Monaco P., 2010, MNRAS, 406, 1533
Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441
Faucher-Giguère C.-A., Kereš D., Ma C.-P., 2011, MNRAS, 417, 2982
Fiorentin P. R., Bailer-Jones C., Lee Y., Beers T., Sivarani T., Wilhelm R., Prieto C. A., Norris J., 2007, A&A, 467, 1373
Gerdes D. W., Sypniewski A. J., McKay T. A., Hao J., Weis M. R., Wechsler R. H., Busha M. T., 2010, ApJ, 715, 823
Geurts P., Ernst D., Wehenkel L., 2006, Mach. Learn., 63, 3
Graff P., Feroz F., Hobson M. P., Lasenby A., 2014, MNRAS, 441, 1741
Guo Q. et al., 2011, MNRAS, 413, 101 (G11)
Henriques B. M., Thomas P. A., Oliver S., Roseboom I., 2009, MNRAS, 396, 535
Henriques B. M., White S. D., Thomas P. A., Angulo R. E., Guo Q., Lemson G., Springel V., 2013, MNRAS, 431, 3373
Hopkins P. F. et al., 2010, ApJ, 715, 202
Ivezić Ž., Connolly A. J., VanderPlas J. T., Gray A., 2014, Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data. Princeton Univ. Press, Princeton, NJ
Johnson R., Zhang T., 2011, preprint (arXiv:1109.0887)
Kamdar H., Turk M., Brunner R., 2015, MNRAS, submitted
Kang X., Jing Y., Mo H., Börner G., 2005, ApJ, 631, 21
Kauffmann G., 1996, MNRAS, 281, 475
Kauffmann G., White S. D., Guiderdoni B., 1993, MNRAS, 264, 201
Kennicutt R. C., Jr, 1998, ApJ, 498, 541
Kim E. J., Brunner R. J., Kind M. C., 2015, MNRAS, 453, 507
Kind M. C., Brunner R. J., 2013, MNRAS, 432, 1483
Klypin A. A., Trujillo-Gomez S., Primack J., 2011, ApJ, 740, 102
Knebe A. et al., 2015, MNRAS, 451, 4029
Lagos C. d. P., Cora S. A., Padilla N. D., 2008, MNRAS, 388, 587
Lemson G. et al., 2006, preprint (astro-ph/0608019)
Liu Y., Weisberg R. H., 2011, Self Organizing Maps – Applications and Novel Algorithm Design. INTECH Open Access Publisher, Croatia, Chapter 13, Available at: http://www.intechopen.com/books/self-organizing-maps-applications-and-novel-algorithm-design/a-review-of-self-organizing-map-applications-in-meteorology-and-oceanography
Liu L., Yang X., Mo H., Van den Bosch F. C., Springel V., 2010, ApJ, 712, 734
Martin C. L., 1999, ApJ, 513, 156
Mo H., Mao S., White S. D., 1998, MNRAS, 295, 319
Monaco P., Fontanot F., Taffoni G., 2007, MNRAS, 375, 1189
Monaco P., Benson A. J., De Lucia G., Fontanot F., Borgani S., Boylan-Kolchin M., 2014, MNRAS, 441, 2058
Moster B. P., Somerville R. S., Maulbetsch C., Van den Bosch F. C., Macciò A. V., Naab T., Oser L., 2010, ApJ, 710, 903
Neistein E., Weinmann S. M., 2010, MNRAS, 405, 2717 (NW)
Ntampaka M., Trac H., Sutherland D. J., Battaglia N., Poczos B., Schneider J., 2015, ApJ, 803, 50
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Peebles P., 1982, ApJ, 263, L1
Planck Collaboration XIII 2015, preprint (arXiv:1502.01589)
Roe B. P., Yang H.-J., Zhu J., Liu Y., Stancu I., McGregor G., 2005, Nucl. Instrum. Methods Phys. Res. A, 543, 577
Schaye J. et al., 2015, MNRAS, 446, 521
Silverman B. W., 1986, Density Estimation for Statistics and Data Analysis, Vol. 26. CRC Press, Boca Raton, FL
Skillman S. W., Warren M. S., Turk M. J., Wechsler R. H., Holz D. E., Sutter P., 2014, preprint (arXiv:1407.2600)
Somerville R. S., Davé R., 2015, ARA&A, 53, 51
Somerville R. S., Primack J. R., 1999, MNRAS, 310, 1087
Somerville R. S., Hopkins P. F., Cox T. J., Robertson B. E., Hernquist L., 2008, MNRAS, 391, 481
Springel V., 2005, MNRAS, 364, 1105
Springel V., White S. D., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726
Springel V. et al., 2005, Nature, 435, 629
Sutherland R. S., Dopita M. A., 1993, ApJS, 88, 253
Vogelsberger M. et al., 2014, MNRAS, 444, 1518
Wang L., Li C., Kauffmann G., De Lucia G., 2007, MNRAS, 377, 1419
Weinmann S. M., Kauffmann G., Von Der Linden A., De Lucia G., 2010, MNRAS, 406, 2249
White S. D., Frenk C. S., 1991, ApJ, 379, 52
Witten I. H., Frank E., 2005, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Mateo, CA
Xu G., 1995, ApJS, 98, 355
Xu X., Ho S., Trac H., Schneider J., Poczos B., Ntampaka M., 2013, ApJ, 772, 147
Yoshida N., Stoehr F., Springel V., White S. D., 2002, MNRAS, 335, 762

APPENDIX A: FEATURE IMPORTANCE

In the discussion of the halo properties chosen for our analysis, an evaluation of which attributes play a role in determining the galaxy properties was not performed. Here, we provide a feature importance plot that shows the relative importance of the halo properties (at $z = 0$) in predicting galaxy properties.

For tree-based ML techniques, the depth of a feature (i.e. relative rank) used as a decision node can be used to evaluate how important that particular feature is in the learning process. The

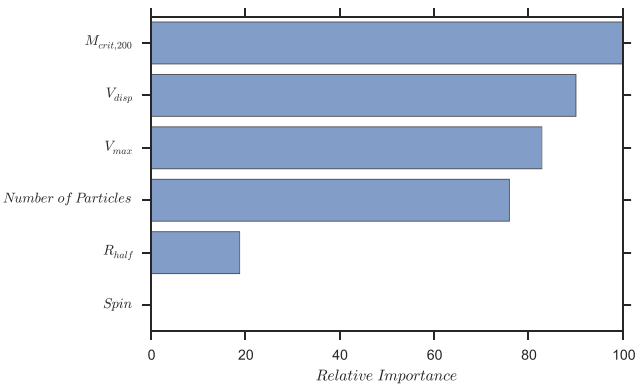

Figure A1. The relative importance of different halo properties in predicting different properties of the galaxy.

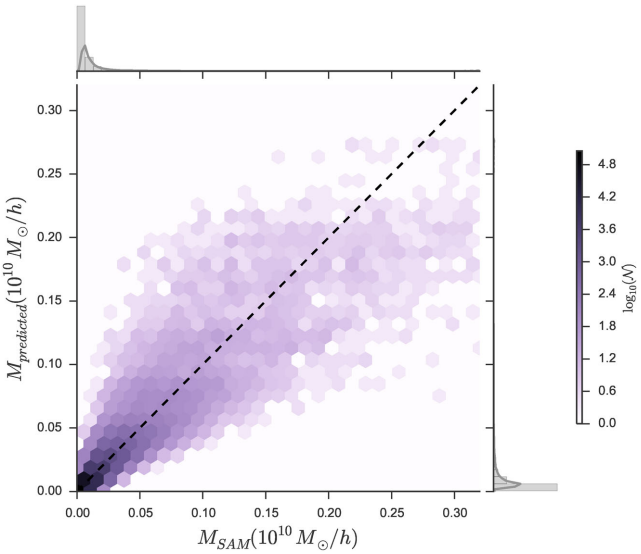

Figure B1. A hexbin plot showing the BH mass prediction for Bower et al. (2006) with a KDE on top.

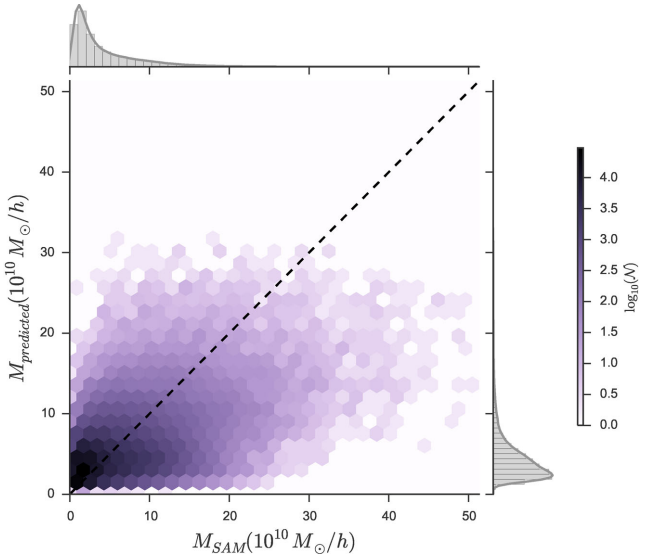

Figure B2. A hexbin plot showing the stellar mass prediction for Bower et al. (2006) with a KDE on top.

expected fraction of the samples a feature contributes to can be used as an estimate of the relative importance of the features. We then average this quantity over all trees in the ensemble to get a less biased estimate for the importance of a particular feature.

As one would expect, the mass of the halo plays an integral role in determining the galaxy properties. Perhaps surprisingly, the spin of the DM halo plays a minimal role in the learning process. This analysis of feature importances will guide future work that uses ML to extract information from DM haloes about the galaxies residing in the halo.

APPENDIX B: USING A DIFFERENT SEMI-ANALYTICAL MODEL

An interesting question is whether ML techniques perform similarly well using a different SAM. The reason we used G11 for this work in place of, or along with, a Durham SAM (Bower et al. 2006) was simply because more halo parameters were available in the merger trees that were constructed for DLB07 and G11. Using G11 offered the opportunity to explore a bigger parameter space.

Here, we explore the effect of using another SAM with fewer halo parameters. For Bower et al. (2006), only the halo mass is provided through the merger tree. We repeat our analysis using just the halo mass over four snapshots and predict only the stellar mass and the BH mass. The point of this analysis is to examine whether ML is able to model the same relationships when a different SAM is used with fewer inputs.

As we can see in the two figures attached, the predictions are noticeably more scattered (particularly the stellar mass) but the general trend is still recovered, even when only one feature is used in our prediction (out of necessity) in a SAM where some of the physics is treated differently. In Kamdar et al. (2015), we explore the feasibility of ML in making predictions from an NBHS, where the physics is vastly more complicated.

This paper has been typeset from a TEX/LATEX file prepared by the author.