Painting galaxies into dark matter halos using machine learning

Shankar Agarwal[1,⋆], Romeel Davé[1,2,3,4♭], Bruce A. Bassett[1,4,5,6†]

[1] *African Institute for Mathematical Sciences, 6 Melrose Road, Muizenberg, 7945, Cape Town, South Africa*
[2] *Institute for Astronomy, Royal Observatory, Edinburgh EH9 3HJ, United Kingdom*
[3] *University of the Western Cape, Bellville, Cape Town, 7535, South Africa*
[4] *South African Astronomical Observatory, Observatory, Cape Town, 7925, South Africa*
[5] *SKA South Africa, The Park, Park Road, Pinelands, Cape Town 7405, South Africa*
[6] *Department of Maths and Applied Maths, University of Cape Town, Cape Town, South Africa*
emails: ⋆ *agarwalshankar@aims.ac.za*, ♭ *romeeld@gmail.com*, † *bruce.a.bassett@gmail.com*

ABSTRACT

We develop a machine learning (ML) framework to populate large dark matter-only simulations with baryonic galaxies. Our ML framework takes input halo properties including halo mass, environment, spin, and recent growth history, and outputs central galaxy and halo baryonic properties including stellar mass (M_*), star formation rate (SFR), metallicity (Z), neutral (H I) and molecular (H_2) hydrogen mass. We apply this to the MUFASA cosmological hydrodynamic simulation, and show that it recovers the mean trends of output quantities with halo mass highly accurately, including following the sharp drop in SFR and gas in quenched massive galaxies. However, the scatter around the mean relations is under-predicted. Examining galaxies individually, at $z = 0$ the stellar mass and metallicity are accurately recovered ($\sigma \lesssim 0.2$ dex), but SFR and H I show larger scatter ($\sigma \gtrsim 0.3$ dex); these values improve somewhat at $z = 1, 2$. Remarkably, ML quantitatively recovers second parameter trends in galaxy properties, e.g. that galaxies with higher gas content and lower metallicity have higher SFR at a given M_*. Testing various ML algorithms, we find that none perform significantly better than the others, nor does ensembling improve performance, likely because none of the algorithms reproduce the large observed scatter around the mean properties. For the random forest algorithm, we find that halo mass and nearby ($\sim$ 200 kpc) environment are the most important predictive variables followed by growth history, while halo spin and ~Mpc scale environment are not important. Finally we study the impact of additionally inputting key baryonic properties M_*, SFR and Z, as would be available e.g. from an equilibrium model, and show that particularly providing the SFR enables H I to be recovered substantially more accurately.

Key words: cosmology: theory – galaxies: evolution – large-scale structure of Universe – methods: N-body simulations – machine learning

1 INTRODUCTION

Galaxies are the primary markers by which we trace out the structure and evolution of the Universe over most of cosmic time. Galaxies form within dark matter halos along the nodes and filaments of the Cosmic Web, thereby providing a luminous (albeit sparse) sampling of the underlying matter distribution. Owing to this, the spatial distribution of galaxies represents one of the observational pillars that allows us to constrain the parameters of the Lambda Cold Dark Matter (ΛCDM) concordance cosmology, and particularly search for deviations from the paradigm that may signify new physics such as the nature of dark energy.

To do so, however, requires understanding how the light emitted by galaxies in various wavelengths traces the underlying matter distribution. Thanks to impressive gravity-only (N-body) simulations, it is now straightforward to precisely relate the population of dark matter halos to the underlying matter distribution. However, relating the luminous components of galaxies to the halo population is less well understood. Galaxy formation involves a complex set of highly nonlinear and self-regulating physical processes spanning an enormous range of physical and temporal scales which cannot be solved ab initio at present (Somerville & Davé 2015; Naab & Ostriker 2017). As such, models of galaxy formation within a cosmological context typically invoke heuristic representations of the small-scale processes such as star formation and supernova feedback that govern galaxy growth.

Recent years have seen remarkable progress in modeling galaxy formation. The pioneering Millennium Simulation (Springel et al. 2005) exceeded the 10 billion particle barrier for dark matter-only runs, with modern N-body simulations now on the verge of exceeding a trillion particles (Potter et al. 2017). Traditionally, to model the galaxy population, the N-body simulations are used as an input to so-called semi-analytic models (SAMs) (White & Frenk 1991; Kauffmann et al. 1993; Cole et al. 1994) that employ highly parameterized descriptions of galaxy evolution within halos, thus enabling direct predictions for how galaxies trace large-scale structure. The technique of populating N-body simulations with SAMs has continued to grow in sophistication, but fundamentally SAMs rely on a simplified analytic description of galaxy formation and have a large number of free parameters. Hence while they have demonstrated numerous successes in reproducing the observed galaxy population, it is unclear whether they provide sufficient robustness in modeling galaxy formation to be relied upon for precision cosmology.

A separate rapidly-advancing approach to modeling galaxy formation is through the use of hydrodynamic simulations. Here, dark matter and gas are evolved concurrently, with the gas subject to additional hydrodynamic and radiative processes that enable the formation of baryonic condensations that are identifiable as galaxies. Typically, such models still rely on heuristic prescriptions for sub-resolution processes such as star formation, but at least the dynamical influence of the baryons is handled in an ab initio manner. The dynamical motion of inflows and outflows from galaxies, often called the baryon cycle, can be followed directly in such models, resulting in many emergent phenomena. Recent simulations have achieved impressive levels of realism in modeling the observed galaxy population, and therefore offer a competitive approach to connecting galaxies and large-scale structure to SAMs, with many fewer assumptions.

Unfortunately, given current computing capabilities, hydrodynamic simulations require heroic efforts to achieve the Gpc^3-scale volumes required to constrain cosmology at a precision level (Remus et al. 2017), and do not cover the dynamic range to robustly model the detailed buildup of baryonic galaxy components. Utilizing hydro simulations for cosmology optimally thus still requires connecting galaxies to dark matter halos in some way. A simple way might be to assume fitting functions for how galaxies populate halos as a function of halo mass or other properties, but such an approach requires choosing somewhat arbitrary functional forms and does not take advantage of all the correlations among galaxy properties as predicted in the hydro simulations. Ideally, we desire a non-parametric, holistic, and flexible approach to characterize how galaxies populate dark matter halos.

Machine learning provides a potentially powerful approach to address this problem in a way that can exploit hidden insights and correlations in order to accurately populate galaxies into halos. Machine learning is an approach to solving problems that exploits the ability of certain algorithms to cluster data (an example of so-called "unsupervised learning") or to successfully extrapolate from training data – where a target quantity is known – to test data where the target of interest is unknown (so-called "supervised learning"). In general the target of interest may be discrete, in which case one typically considers it a classification problem (such as predicting whether a galaxy will be spiral or elliptical), or may be continuous, in which case it is known as a regression problem. Recent progress in machine learning and artificial intelligence has been spectacular, particularly in areas related to computer vision. In astronomy and cosmology, machine learning has been applied in a myriad of ways[1] including the classification of transients. See `http://www.astroml.org/` for more examples and sample code. Our problem of populating galaxies into halos is a regression problem since the output properties are continuous, and can be solved with a supervised learning approach since a known training sample exists.

In this paper, we use machine learning (ML) to "teach" the computer how to populate galaxies into dark matter halos. For this exercise, we train the ML on the recent Mufasa simulations, which have been shown to provide an impressive match to a wide range of galaxy stellar, gas, and metal properties over cosmic time (Davé et al. 2016, 2017b,a), and then test its efficacy on the separate testing subsample. We show that using information about current halo properties and recent growth history, we can recover the stellar, neutral hydrogen, star formation rate, and metal properties of galaxies to within well-characterized uncertainties. These physical properties directly relate to the luminous output of galaxies in various bands, providing an avenue to connect large-scale halo catalogs with observable galaxy properties in optical/near-infrared bands, nebular emission lines, and 21 cm emission. This work thus sets the stage for machine learning to provide a competitive approach to populating large-scale N-body simulations with galaxies for cosmology applications.

The approach of using machine learning to populate dark matter halos by training on hydrodynamic simulations has previously been examined by Kamdar et al. (2016), who used the Illustris simulation (Vogelsberger et al. 2014) as the training set. Our approach will be generally similar to theirs, and in areas of overlap our results quantitatively corroborate theirs. We then go on to extend this approach in several ways: Besides utilizing Mufasa which agrees better with data than Illustris for key galaxy properties (particularly massive quenched systems), we specifically examine H I with an eye towards future H I intensity mapping surveys, we investigate the nature of the scatter by examining second parameter correlations as predicted by ML, and

[1] see http://arxiv.org/find/astro-ph/1/abs:+AND+machine+learning/0/1/0/all/0/1

we explore ways to make the ML more accurate through additional inputs. Our results lend further support to the idea that machine learning provides a valuable approach towards characterizing the baryonic and observable properties of galaxies within large-volume N-body simulations.

In §2 we briefly review the MUFASA simulation. In §3 we describe the input and output parameters for our ML framework, and describe the various ML methods that we employ here. In §4 we describe our results for how well machine learning can recover intrinsic galaxy baryonic properties from only the dark matter halo catalog and its recent growth history. In §4.4 we show how H I and H_2 can be recovered more accurately when M_*, SFR, and Z are used as input properties for ML. In §5 we compare results from the various ML algorithms explored in this paper. Finally, we summarize and conclude in §6.

2 COSMOLOGICAL SIMULATIONS

For this work we employ the MUFASA cosmological hydrodynamic simulation, described fully in Davé et al. (2016). MUFASA uses the GIZMO code (Hopkins 2015) to evolve dark matter using a tree-particle-mesh scheme, and gas using a meshless finite mass (MFM) algorithm that offers important numerical advantages over traditional smoothed particle hydrodynamics solvers in terms of accuracy and robustness.

MUFASA includes a wide range of physical phenomena for producing realistic galaxies, including radiative cooling with metal lines, photoionisation heating, molecular gas-based star formation using a subgrid model for computing the molecular gas density (Krumholz et al. 2009), chemical enrichment from Type II and Type Ia supernovae along with stellar evolution, and prescriptions for feedback from star formation and for quenching massive galaxies. The star formation feedback is done kinetically using two-phase outflows, based on the results from high-resolution simulations of the Feedback in Realistic Environments project (Muratov et al. 2015). The quenching feedback is crudely modeled by keeping hot gas in massive halos hot as in "radio mode" feedback (Croton et al. 2006), which as shown by Gabor et al. (2016) is effective at producing quenched galaxies with a sharp turnover in the galaxy mass function and a red sequence that is in reasonable agreement with observations.

MUFASA does comparably well or better at reproducing galaxy survey observations over a wide range of epochs, as compared to other modern hydrodynamic simulations such as Illustris (Genel et al. 2014) or Eagle (Schaye et al. 2015). The stellar mass function, which is a crucial barometer of how galaxy growth tracks halo growth, is in very good agreement with observations from $z \sim 0 \to 6$ (Davé et al. 2016). Gas and metal properties are also in good agreement (Davé et al. 2017b), and MUFASA produces a bimodal galaxy population of quenched and star-forming systems broadly similar to that observed (Davé et al. 2017a). While non-trivial discrepancies versus observations remain, the overall success of MUFASA at matching observations suggests that it provides a state of the art platform to examine the connection between dark matter halos and key observable galaxy properties.

For this paper, we utilize the 50 h^{-1}Mpc MUFASA run, which has a minimum (adaptive) softening of 0.5 h^{-1}kpc (comoving), a gas element mass of $1.8 \times 10^7 M_\odot$, and a dark matter particle mass of $9.6 \times 10^7 M_\odot$, allowing us to resolve galaxies down to a stellar mass of $5.8 \times 10^8 M_\odot$; we will only use galaxies above this stellar mass in our analysis. We have shown that the stellar (Davé et al. 2016) and gas (Davé et al. 2017b) mass functions are well-converged above this limit; the SFR function less so but still reasonable. We employ a Planck-concordant cosmology with $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, $\Omega_b = 0.048$, $H_0 = 68$, $n = 0.97$, and $\sigma_8 = 0.82$.

We output 135 snapshots to $z = 0$ of this simulation, with a typical time separation between snapshots at later epochs of $200 - 300$ Myr. For every snapshot, we identify all galaxies using Spline Kernel Interpolative Denmax (SKID, Kereš et al. 2005), and we identify halos using a 3-D friends-of-friend halo finder that is native in the public YT simulation analysis package with a linking length of 0.2 times the mean interparticle spacing. The linking length b is roughly related to the spherical overdensity Δ w.r.t the background matter density through $\Delta/178 \approx (0.2/b)^3$ (Cole & Lacey 1996; Courtin et al. 2011). Quoted halo masses represent the total dynamical mass including baryons. We use CAESAR[2], a YT add-on package, to cross-match the galaxies and halos, and to compute a range of galaxy and halo properties for each snapshot. Finally, we identify galaxy progenitors by finding the galaxy at the previous snapshot that holds the largest fraction of baryonic mass (gas+stars) that is now in the current SKID galaxy; and similarly for halo progenitors. These CAESAR galaxy/halo catalogs, plus the progenitor information, represent the input data for the machine learning analysis presented here.

3 MACHINE LEARNING SETUP

Here we briefly describe the input and outputs parameters for our ML framework, and describe the various ML methods we will cross-compare to determine the most effective algorithm.

3.1 ML inputs and outputs

Fig. 1 shows a skeleton of the supervised machine learning network. Using a suitable training set (input parameters for which data is available), the machine learning algorithm is trained to learn a mapping. With this mapping the network is capable of reproducing (as closely as possible) the output, when queried with input parameter settings that are part of the training set. The trained network can now be presented with new settings of the input parameters (for which one does not have any prior data) and by using the same mapping learnt during the training process, the network makes predictions.

Adjusting the ML network parameters to get the desired mapping is called *training* the network. For estimating the

[2] `http://caesar.readthedocs.io/en/latest/`

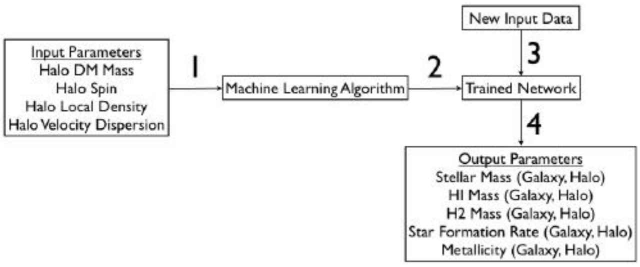

Figure 1. Steps 1 and 2: a machine learning network learns to parametrize the output, for the input patterns that form the training set. Steps 3 and 4: the trained network is capable of making predictions when presented with input parameter settings. The queried input settings must lie within the parameter ranges of the patterns in the training set.

galaxy and halo quantities, we use a training set of dark matter halos, each with the following known parameters:

$$\mathbf{I} \equiv (\{\mathrm{M_{halo}}\}_\mathrm{i}, \{\rho\}_\mathrm{j}, \lambda, \sigma_\mathrm{vel}), \qquad (1)$$

where $\{\mathrm{M_{halo}}\}_\mathrm{i}$ are the halo masses at the present snapshot ($i = 0$) and the 5 immediately preceding snapshots (thus characterizing the recent merger history), $\{\rho\}_\mathrm{j}$ are the nearby halo mass densities within radii r $= 200, 500, 1000$ kpc centered on the halo center-of-mass, λ is the halo spin, and σ_vel is the halo velocity dispersion.

The galaxy and halo quantities that ML predicts are:

$$\mathbf{O} \equiv (\mathrm{M_*}, \mathrm{M_{HI}}, \mathrm{M_{H2}}, \mathrm{SFR}, Z), \qquad (2)$$

where $\mathrm{M_{HI}}$ is the H I (neutral hydrogen) mass, $\mathrm{M_{H2}}$ is the molecular hydrogen mass, M_* is the stellar mass, SFR is the star formation rate, and Z is the SFR-weighted metallicity. We predict these 5 quantities both for entire halos and their central galaxies, thereby making **O** a 10-dimensional output space. Likewise, **I** is an 11-dimensional input space. In this paper, we will only present central galaxy properties, for brevity; the results for the entire halos show very similar trends and do not provide significant additional insights, so we omit them and simply note that the ML algorithm is able to also predict these if required, e.g. for intensity mapping purposes.

Since the quantities in **I** and **O** span many orders of magnitude, in our fitting procedure we work with their logarithm, giving us our cost function χ^2, which is presented to the ML setup for minimization with respect to the network weights **w**,

$$\chi^2(\mathbf{w}) = \sum_{t=1}^{T} \sum_{k=1}^{N_{out}} \log^2 \left[\frac{\mathbf{O}^{\mathrm{ML}}(k|\mathbf{w}, \mathbf{I}_t)}{\mathbf{O}(k|\mathbf{I}_t)} \right], \qquad (3)$$

where the index t runs over all dark matter halos in the training set, and the index k runs over the N_{out} quantities in **O** (for our case, $N_{out} = 10$). $\mathbf{O}(k|\mathbf{I}_t)$ is the k^{th} component (i.e., output quantity) of **O**, given the input quantities $\mathbf{I}_t$ corresponding to the t^{th} halo in the training set. The numerator is the ML prediction given the weights **w**, while the denominator is the true value from MUFASA.

3.2 ML algorithms

We constructed the mapping between dark matter halo and galaxy properties using the ML routines implemented within the publicly available PYTHON ML library, SCIKIT-LEARN[3] (Pedregosa et al. 2012) . We explored the following ML routines: Support Vector Regression (SVR), k-Nearest Neighbors (kNN), Multilayer Perceptron (MLP), Decision Trees (DT), Random Forests (RF), Extremely Randomized Trees (ET) and AdaBoostRegressor (AdR). We briefly describe each below; for more information on each of the above ML algorithms, refer to the SCIKIT-LEARN library.

3.2.1 Support Vector Regression (SVR)

SVR builds on the Support Vector Machine classification algorithm which finds the hyperplane that separates the features optimally. SVR finds the best surface to explain the data. As with SVM, one can either fit a flat hyperplane (the linear case), or via the kernel trick, fit a nonlinear function efficiently to the data. We treat the kernel as a hyperparameter and optimize over polynomial, sigmoid and radial basis function (RBF) kernels with C, the error penalty parameter, varied on a grid between 0.1 and 10.

3.2.2 k-Nearest Neighbors (kNN)

kNN is one of the simplest non-parametric classification and regression algorithms. It proceeds via a type of ”voting” system by simply finding the k nearest neighbors of the test point in feature space (where k is a user-chosen integer which is a hyper-parameter that can be optimized) and using the mean of these targets to perform the regression.

3.2.3 Multilayer Perceptron (MLP)

MLP (a form of an artificial neural network) is a collection of nodes arranged in a series of layers, with each node in a layer connected to all other nodes in adjacent layers. For example, a network with an architecture $5 : 15 : 10 : 4$ has 5 input features, two hidden layers with 15 and 10 nodes respectively, and finally 4 output features. An extra node (called the bias node) is added to the input layer as well as to each of the hidden layers to compensate for the difference between the network's mean prediction and the mean of the outputs of training set patterns (for a review, refer Bishop 1995). The connections are referred to as MLP weights. During network training the weights are adjusted so as to minimize the difference between network predictions and the target values of the validation set patterns.

3.2.4 Decision Trees (DT)

DT are graphs implementing non-parametric learning. They consist of trees where each node poses a test question (e.g. ”Is the metallicity greater than some threshold, Z_*?”) to

[3] http://scikit-learn.org

the incoming data that needs to be classified or regressed on. The answers (called leafs) either lead down the tree to new nodes or to a final prediction. The overall complexity of a DT depends on the depth of the tree which can be controlled by various parameters such as (i) maximum tree depth, (ii) minimum number of data samples to use when deciding whether to implement an internal node split, (iii) minimum error reduction (over the training set) that a daughter node must achieve over its parent node when deciding whether to implement a node split. The metric we use to select the features and thresholds at the nodes is Mean Square Error. Note that DT are susceptible to over-fitting the data, a problem mostly solved by RF.

3.2.5 *Random Forests (RF) and Extremely Randomized Trees (ET)*

RF and ET are ensemble learning methods that aggregate the predictions of many decision trees and output a mean prediction to improve accuracy and control over-fitting. The major difference between these two ensemble methods is that ET has many more leaf nodes, because the samples are drawn from the entire training set instead of a bootstrap sample and the splits are chosen at random from the range of values in the sample. One of the advantages of ensemble tree methods is that they automatically provide a measure of feature importance. A key feature will appear in many of the trees in the ensemble while an irrelevant feature will show up only rarely. We will use this characteristic to rank the importance of our features, as shown in Figs (6) and (9).

3.2.6 *AdaBoostRegressor (AdR)*

AdR fits a base regressor (any ML routine; for example SVR) on the original dataset. A boosted ensemble is then built by fitting additional copies of the regressor on the same dataset with weights of instances adjusted according to the error of the current prediction. Dataset samples with wrong predictions get their weights boosted, thereby making subsequent regressors focus more on difficult cases. In the figures below we will show AdR's performance with the base regressors SVR and RF (which we call SVR+ and RF+), with similar results achieved for the other regressors.

It turns out that RF and SVR provide the best results on average, i.e. the tightest relationship between the training set and predictions, as we demonstrate in §5. We note here that an exhaustive search through the hyper-parameter space for each of the ML methods could improve upon our ML results, as well as change the ranking of the various ML methods. For example, using the EDEN (Dufourq & Bassett 2017) framework to genetically evolve the optimal MLP, we were able to get MLP to perform at par with RF. However, the marginal improvement in the results suggests that we have reached the limit on the amount of information that is possible to extract from our input data.

Here we particularly focus on the RF results because RF also provides a straightforward way to assess the importance level of each input parameter in the final predictions, which is an interesting diagnostic.

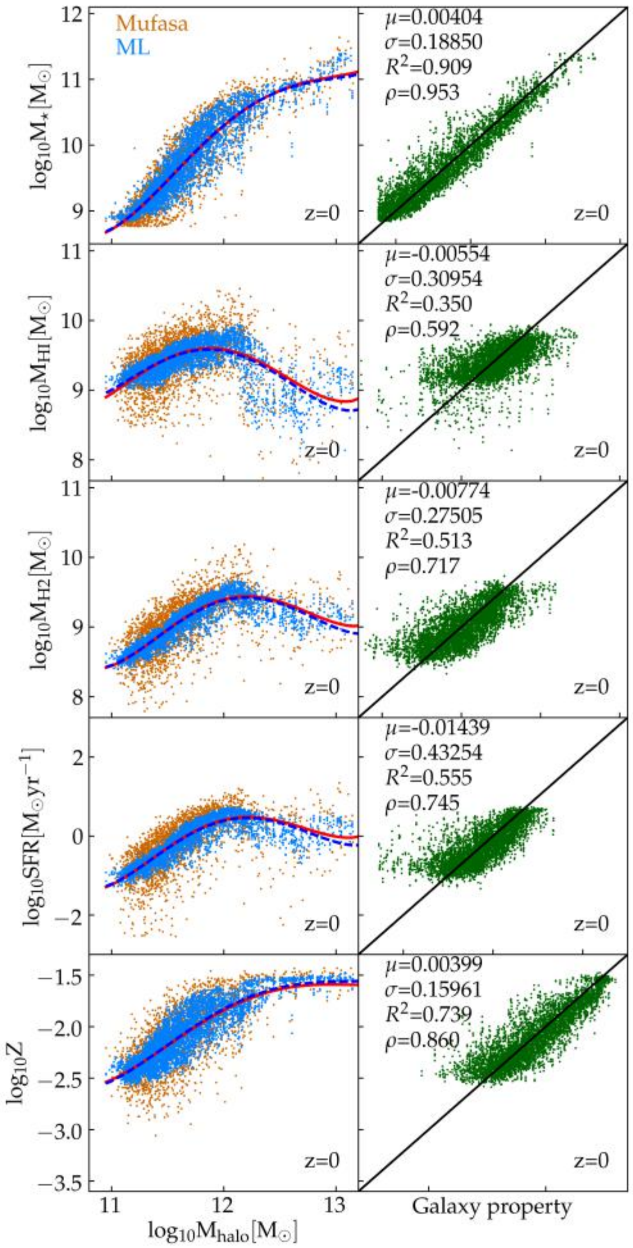

Figure 2. *Left column:* Central galaxy properties $[M_*, M_{HI}, M_{H2}, SFR, Z]$ vs. host halo mass, at $z = 0$. The brown and blue points correspond to values from the MUFASA simulation and ML using random forests, respectively. The mean scaling relations are shown by red solid (MUFASA) and blue dashed (ML) lines, which mostly overlap showing that the mean trends are well recovered. The scatter in the true (MUFASA) values is however larger than predicted from ML. *Right column:* Comparison of individual central galaxy properties' true values from MUFASA (x-axis) vs. ML predictions (y-axis). A perfect prediction corresponds to the diagonal solid line. The mean μ and r.m.s width σ, in dex, along with the regression statistic R^2 and the Pearson correlation coefficient ρ (see equations 4-7) are listed in the upper left.

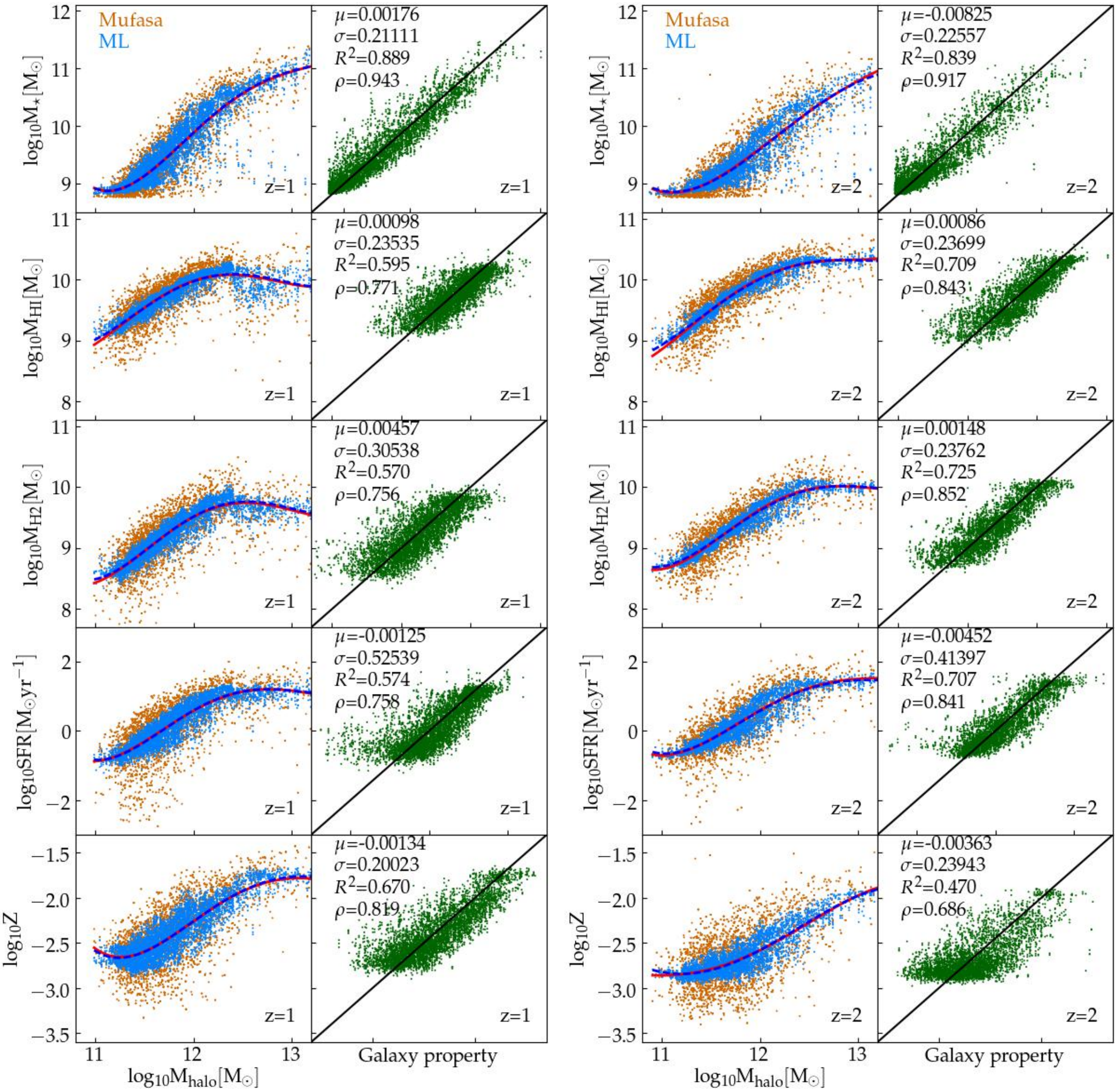

Figure 3. As in Fig. 2 except at $z = 1$ (left) and $z = 2$ (right). In each, the left panels show central galaxy properties as a function of halo mass for the original Mufasa simulation (brown points) and the ML prediction (blue points), and the right panels compare the ML predicted value (y-axis) vs. the true Mufasa values (x-axis). At higher redshifts, the ML does a better job recovering the true values for gas and SFR than at $z = 0$, but worse for M_* and Z.

3.3 ML setup

We divide our dataset of almost 3400 dark matter halos into training (2500), validation (500) and testing sets (400). For each halo, the 11 quantities of **I** are fed to the network (see Fig. 1), while the output is held fixed at the 10 quantities of **O**. The space of the ML hyper-parameters was randomly sampled 2000 times. The hyper-parameter configuration that minimized the cost function with respect to the halos in the validation set was chosen as the optimized network architecture. The hyper-parameters we searched for the various ML routines are the error penalty term (for SVR), the number of neighbors (for kNN), the maximum tree depth, the number of samples required to split an internal node, and the subset of input features $\in$ **I** to consider when looking for the best split (for RF, DT and ET), the number of hidden layers and the number of nodes per hidden layer (for MLP), the number of boosts (for AdaBoostRegressor). For RF and ET, we trained a committee of 100 decision trees.

Furthermore, in order to get an estimate of the variability in our predictions, we shuffled our deck of 3400 halos and randomly split it into training, validation and testing sets with 2500, 500 and 400 halos respectively. We repeated this 10 times, thereby giving us a set of 10 trained ML solutions. We use these to estimate the variance in output quantities.

4 TESTING ML VS. MUFASA

Galaxies have a variety of properties that display both overall individual trends and complex interrelationships. Ideally, an ML algorithm would be able to learn the trends and interrelationships precisely, i.e. given **I**, ML should ideally predict **O** exactly (to within intrinsic uncertainties) as would be obtained from running a full hydrodynamic simulation. In practice, however, some aspects will be easier to recover than others. As such, we can test ML on a variety of levels.

Here we choose to focus on testing the trends predicted by ML versus halo mass M_h, as this is the most fundamental quantity describing halos in N-body simulations. In this context, we can envision testing ML in four different stages, in roughly increasing order of stringency: (i) that ML reproduces the *mean relationship* between galaxy properties and M_h; (ii) that ML reproduces the amplitude of the *scatter* around the mean relation, as a function of M_h; (iii) that ML produces the correct galaxy property for each *individual* galaxy; and (iv) that ML recovers intrinsic *higher-order correlations* between various galaxy properties. In this section we describe how ML fares in these four stages of testing using MUFASA.

Our goal is to use DM-only information to see how well the baryonic galaxy properties can be predicted. We will show in §5 that RF and SVR give the best performance, thus we choose to only show RF predictions of our five predicted quantities: neutral gas mass $\mathrm{M_{HI}}$, molecular gas mass $\mathrm{M_{H2}}$, stellar mass M_*, star formation rate SFR, and SFR-weighted metallicity Z for halos and their central galaxies, at three redshifts $z = 0, 1$ and 2. All plots correspond to halos in the testing set, unless stated otherwise. We augment the information presented to ML by providing the DM accretion history. For making predictions at $z = 0$, ML is given the four halo inputs **I** (see equation 1) computed at $z = 0$, as well as the halo mass, $\mathrm{M_{halo}}$, from redshifts $z = 0.25, 0.20, 0.10$ and 0.05. Likewise, for making predictions at $z = 1$, ML is given **I** computed at $z = 1$, plus $\mathrm{M_{halo}}$ from redshifts $z = 1.5, 1.4, 1.3, 1.2$ and 1.1. For $z = 2$, the information provided is **I** computed at $z = 2$, and $\mathrm{M_{halo}}$ from redshifts $z = 2.5, 2.4, 2.3, 2.2$ and 2.1. These choices correspond to the frequency of snapshot outputs in MUFASA.

4.1 Mean relations and scatter

Fig. 2 shows ML predictions using random forests for the five galaxy quantities at $z = 0$. The panels are M_*, $\mathrm{M_{HI}}$, $\mathrm{M_{H2}}$, SFR and Z (top to bottom row). The brown and blue points correspond to values from the MUFASA simulation and RF respectively. The mean scaling relations are shown by red solid (MUFASA) and blue dashed (ML) lines. Note that we use a quartic polynomial fit to get the mean scaling relations. One could instead compute the means over the halo mass bins. We have checked that the two methods give similar scaling relations (except for the lowest and highest halo mass bins) and does not change our overall results.

The green panels on the right compare the RF predictions of the galaxy quantities (on y-axis) vs. their true values (on x-axis) from MUFASA. A perfect prediction corresponds to all points lying along the diagonal solid line. To quantify deviations from perfect prediction, we compute four metrics: mean deviation μ, r.m.s width σ, regression score R^2 and the Pearson correlation coefficient ρ (all listed in the right panels, with μ and σ in dex units). If x_i and $\hat{x}_i$ are the logarithms true and predicted values of the i^{th} data point, then the four metrics are estimated over a sample size N as

$$\mu(x, \hat{x}) = \frac{1}{N}\sum_i^N (x_i - \hat{x}_i), \tag{4}$$

$$\sigma^2(x, \hat{x}) = \frac{1}{N-1}\sum_i^N \left[(x_i - \hat{x}_i) - \mu\right]^2, \tag{5}$$

$$R^2(x, \hat{x}) = 1 - \frac{\sum_i^N \left[x_i - \hat{x}_i\right]^2}{\sum_i^N \left[x_i - <x>\right]^2}, \tag{6}$$

where $<x>= \sum_i^N x_i/N$ is the mean of true values x.

The Pearson coefficient is a measure of the linear relationship between two quantities. In our case, for x_i and $\hat{x}_i$, ρ estimated over a sample size N is

$$\rho(x, \hat{x}) = \frac{\sum_i^N (x_i - <x>)(\hat{x}_i - <\hat{x}>)}{\sqrt{\sum_i^N (x_i - <x>)^2}\sqrt{\sum_i^N (\hat{x}_i - <\hat{x}>)^2}}, \tag{7}$$

where $<\hat{x}>= \sum_i^N \hat{x}_i/N$ is the mean of predictions $\hat{x}$.

Comparing the ML predictions and true MUFASA values for the testing set at $z = 0$, we see that they agree extremely well in the mean. The red and blue mean lines in the left panels are almost completely overlapping. This is reflected in the μ values in the right panels, which show extremely small systematic deviations typically well below 1%. This shows that ML provides an excellent recovery of the mean trends, which is already non-trivial given that there are interesting physical features in the simulated galaxies such as the break in properties around $M_{\mathrm{halo}} \approx 10^{12} M_\odot$ above which galaxies become quenched. We note that, owing to our imposed stellar mass resolution limit, there is a bend at the low-M_{halo} end of the M_* curve, as well as for other quantities that correlate strongly with M_*; this is an artifact of the limited dynamic range of our simulation, but does not affect our conclusions.

Although the mean is well-predicted, it is clear that the scatter appears to be under-predicted, in the sense that the original MUFASA (brown) points span a larger range of values than the ML predicted (blue) points. This indicates that ML prefers to predict objects closer to the mean relation than they should be. The typical difference in the scatter around the mean relation is $\sim 10-30\%$ between the original Mufasa sample and the ML predictions.

This inability to fully follow the scatter manifests in deviations when comparing galaxies individually, as done in the right panels that show the original (x-axis) vs. predicted (y-axis) values for each galaxy. Ideally, the predictions would exactly follow the 1-to-1 diagonal solid line. While on average they do so quite well, as noted by the low values of μ, there is substantial scatter around these relations that indicate ML does not provide a perfect recovery. The scatter σ around these 1-to-1 relations, plus the other measures shown such as the regression coefficient R^2 and the Pearson correlation coefficient ρ, thus represents a measure of the accuracy of ML.

The scatter σ is lowest for M_* and Z, both less than 0.2 dex, while the gas masses $\mathrm{M_{HI}}$ and $\mathrm{M_{H2}}$ have scatter around 0.3 dex. SFR shows the largest scatter of over 0.4 dex. This likely owes to the fact that star formation rate is driven by processes that are fairly rapidly fluctuating, such as stochastic accretion and mergers, which is not well correlated with large-scale halo properties. Similarly, gas content, which drives star formation, is also dependent on more rapidly varying baryonic physics. In contrast, stellar mass and metallicity are primarily accumulated quantities, hence short-timescale variations tend to average out to yield a better correlation.

R^2 and ρ yield qualitatively the same impression as σ: Stellar mass and metallicity show good correlation between true and predicted values, but the other quantities are worse. In particular, H I shows the poorest correlation among all the quantities. This is particularly troubling for predicting H I intensity maps, since even though the low resolution intensity mapping will Poisson-average over small-scale variations, a large predicted scatter will translate to a larger required mapping area in order to achieve the same level of statistical accuracy.

Fig. 3 shows the corresponding plots to Fig. 2 except at $z = 1, 2$. The basic story is the same, in that M_* and Z are best-predicted, while gas content and SFR are less well predicted. Examining the values of σ, R^2, and ρ, we see that the correlations are better for $\mathrm{M_{HI}}$, $\mathrm{M_{H2}}$, and SFR than at $z = 0$, albeit slightly worse for M_* and Z. It appears that ML does an increasingly better job at recovering these "short-term" galaxy properties to higher redshifts, likely because quenching is less prevalent at high-z and it is quenching that drives galaxies to deviate systematically from the gas and SFR values they typically have while on the galaxy main sequence. However, ML does poorer at the "cumulative" quantities partly owing to the smaller training sets available at these higher redshifts.

Comparing to the Illustris-based ML study of Kamdar et al. (2016), we find that our values of R^2 and ρ for M_* and SFR agree quite well with their predictions, showing that this level of agreement is generally independent of the details of ML algorithm and training simulation. However, for gas mass and metallicity, our correlation measures are somewhat worse than their quoted values. The reason for the discrepancy is not clear, although it could stem from the way these quantities are defined and calculated in Illustris vs. Mufasa. It could also be because Illustris, having a larger dynamic range than Mufasa, contains many more galaxies, and thus is able to obtain a tighter fit. Judging directly from their figures, it appears that the level of scatter between the true and predicted values for these quantities are broadly similar as ours.

In summary, our ML framework is able to accurately recover the mean stellar, gas, and star formation contents of galaxies versus halo mass from Mufasa at all redshifts. However, individual galaxies can deviate substantially, in a manner that depends on the quantity being examined; M_* and Z are most tightly predicted, while gas content and SFR are less well predicted at $z = 0$. The latter predictions improve significantly when going to higher redshifts.

4.2 Second-parameter correlations

In the real Universe, it is known that the observed scatter in metallicity, gas content, and SFR are correlated. For instance, it is observed that at a given stellar mass, a galaxy with high SFR will tend to have higher gas content (Bothwell et al. 2013) and lower metallicity (Lara-López et al. 2010; Mannucci et al. 2010). This can be understood theoretically in a model in which fluctuating gas inflow drives galaxies off the mean relation with respect to stellar mass, but subsequent processes tend to recover the galaxies back onto their equilibrium relations (Davé et al. 2011, 2017b). Galaxies with temporarily elevated inflows (such as a merger) of lower-metallicity gas achieve higher H I and $\mathrm{H_2}$ content, which stimulates SFR; conversely a lull in accretion results in more of a closed-box situation where the gas gets consumed and more metals are produced. In the previous section we saw that ML can well reproduce mean trends, but the individual scatter can be substantial, particularly for gas content and SFR. To investigate this further, we explore whether ML recovers the intrinsic second-parameter correlations in the scatter of these quantities.

Fig. 4 shows the second-parameter dependences with specific SFR ($\mathrm{sSFR} \equiv \mathrm{SFR}/M_*$) for the RF predictions at $z = 0$ and $z = 1$. The panels (from top to bottom) show the H I fraction ($\mathrm{f_{HI}}$), $\mathrm{H_2}$ fraction ($\mathrm{f_{H2}}$) and metallicity (Z) as a function of galaxy stellar mass M_*. Galaxies are color-coded by their distance from the mean M_*-sSFR relation obtained from ML (blue dashed and dotted lines), such that galaxies with above average sSFR are bluer, and galaxies with lower sSFR are redder. These can be compared e.g. to plots such as Fig. 3 and Fig. 5 in Davé et al. (2017b). Like in the case versus halo mass, the mean relations versus galaxy stellar mass agree extremely well between Mufasa and ML at both $z = 0$ and $z = 1$.

More remarkable is the fact that ML recovers the second-parameter correlations in the scatter, despite not being provided any baryonic information. That is, for a given M_*, galaxies with higher sSFR than typical are also metal-poor with an above-average H I and $\mathrm{H_2}$ fractions. The correlations also exist at redshifts $z = 2$ (not shown), although are somewhat less evident. These trends qualitatively mimic those seen in observations (Salim et al. 2014, 2015) and hydrodynamic simulations (Davé et al. 2011).

Property		$z=0$			$z=1$			$z=2$	
	σ	R^2	ρ	σ	R^2	ρ	σ	R^2	ρ
M_*	0.19	0.91	0.95	0.21	0.89	0.94	0.23	0.84	0.92
$Z_{\rm SFR}$	0.16	0.74	0.86	0.20	0.67	0.82	0.24	0.47	0.69
SFR	0.43	0.56	0.75	0.53	0.57	0.76	0.41	0.71	0.84
$\rm M_{HI}$	0.31	0.36	0.59	0.24	0.60	0.77	0.24	0.71	0.84
$\rm M_{H2}$	0.28	0.51	0.72	0.31	0.57	0.76	0.24	0.73	0.85
$\rm M_{HI}$|Baryon	0.25	0.59	0.77	0.21	0.69	0.83	0.21	0.78	0.88
$\rm M_{H2}$|Baryon	0.13	0.88	0.94	0.17	0.87	0.93	0.17	0.85	0.92

Table 1. The scatter σ, the regression score R^2 and the Pearson correlation coefficient ρ for the MUFASA vs. ML scatter plots from Figs. 2-3. The last two rows correspond to Fig. 8 where we additionally supply as inputs the baryonic properties M_*, SFR, and Z.

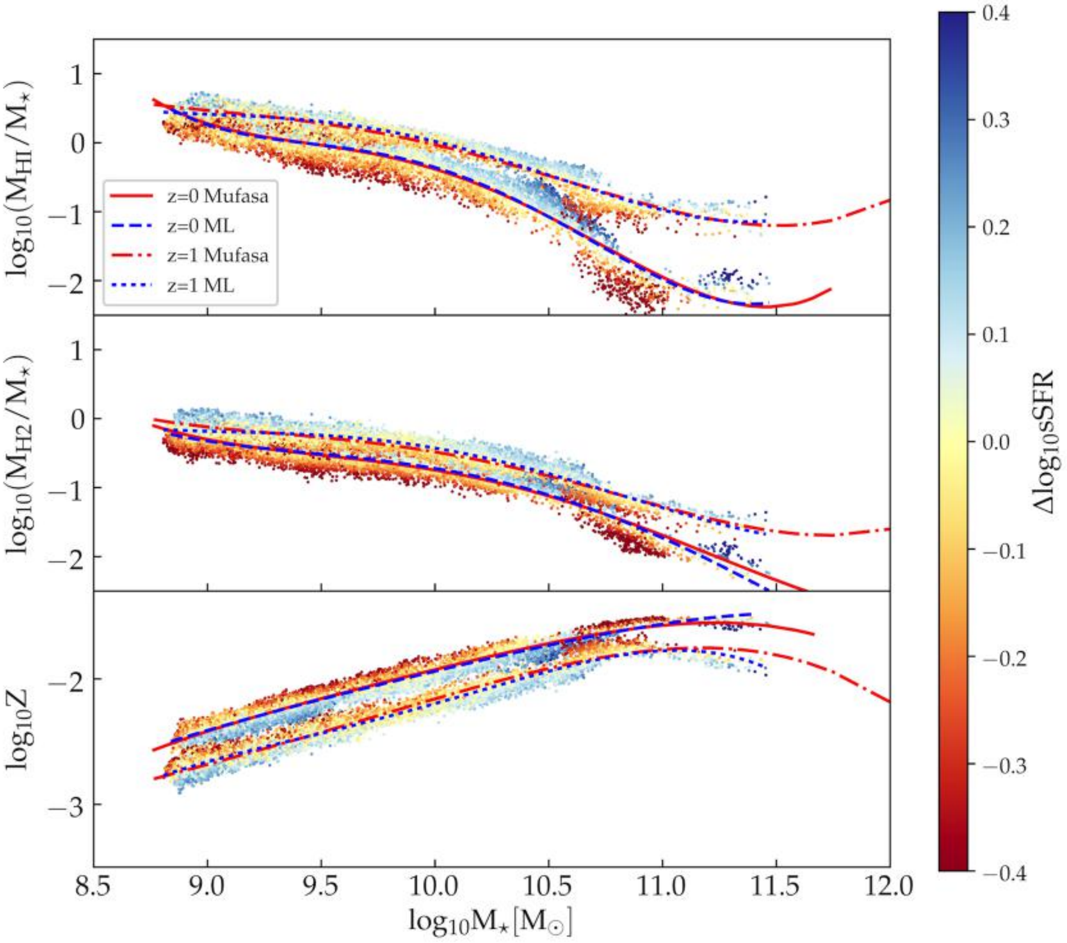

Figure 4. Second-parameter correlations at $z=0$ and $z=1$. The panels (from top to bottom) show the H I fraction ($\rm f_{HI}$), $\rm H_2$ fraction ($\rm f_{H2}$) and metallicity (Z) of galaxies as a function of their stellar mass M_*. The mean scaling relations are shown by the red (solid and dash-dotted, for MUFASA) and blue (dashed and dotted for ML predictions) lines, and are in very good agreement with each other. Galaxies are color-coded by their distance from the mean M_*-sSFR relation ($\Delta\log_{10}$sSFR) obtained from ML (blue dashed and dotted lines). At each redshift, for a given M_*, galaxies with higher sSFR than the mean relation tend to be metal-poor with a higher than average H I and $\rm H_2$ content.

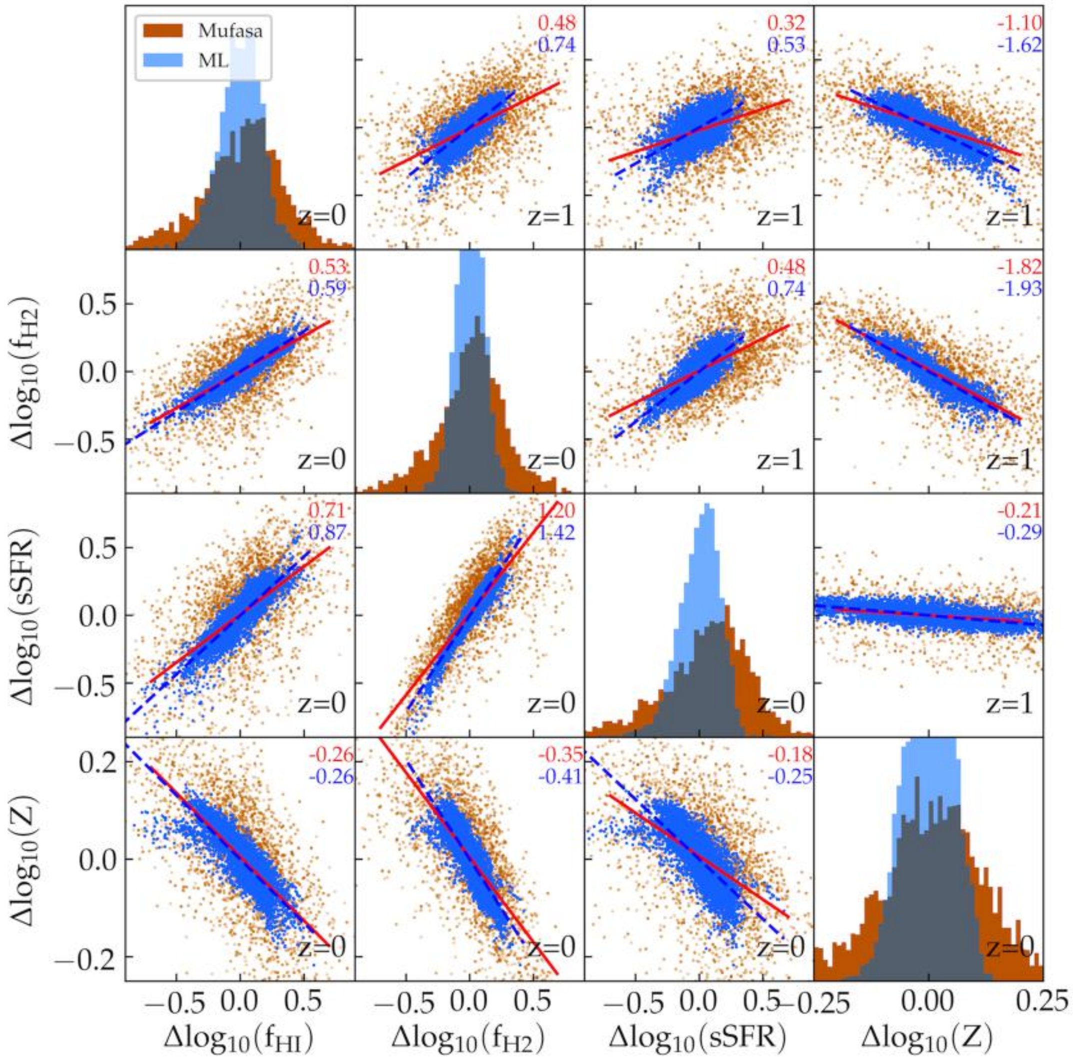

Figure 5. Deviation of H I fraction (f_{HI}), H_2 fraction (f_{H2}) and metallicity (Z) of galaxies from their mean scaling relations versus M_* (i.e. the red solid and blue dashed lines in Fig. 4), at $z = 0$ (lower left triangle) and $z = 1$ (upper right triangle). The brown and blue points correspond to galaxies from the MUFASA simulation and ML respectively. Histograms of each quantity at $z = 0$ are shown along the diagonal. The best-fit power laws (red solid line for MUFASA and blue dashed line for ML) are overplotted, and their slopes are listed in the upper right of each panel. Generally, the ML recovers the trend and approximate slope of the MUFASA deviation plots, showing that it is capable of learning higher-order correlations.

We can conduct a more quantitative comparison to MUFASA using deviation plots (Davé et al. 2017b), in which we show the deviation of the quantities from their mean scaling relations vs M_*, plotted against each other. Fig. 5, shows the same data in Fig. 4, re-plotted in terms of deviations of f_{HI}, f_{H2}, sSFR and Z from their respective mean scaling relations vs. M_*. Brown points show the second-parameter dependences directly from the MUFASA simulation, while the blue points show the ML predictions. Lines (red solid for MUFASA and blue dashed for ML) show best-fit power laws, with the slopes of the trends listed in the upper right of each panel.

The predictions from ML reasonably well reproduce MUFASA power law slopes of all these deviations, though not perfectly so. The broad agreement is nonetheless a remarkably stringent test of ML's ability to mimic simulated galaxy properties in terms of higher-order correlations. The scatter in the deviations is also reasonably reproduced, but as with the overall scatter in the quantities themselves, the blue ML points do not quite span the full range of MUFASA's red points, reflecting the the smaller scatter predicted by ML than seen in MUFASA.

More broadly, the second-parameter dependence shows the power of machine learning in being able to be trained on all the available information to predict galaxy properties in greater detail. A simple parameterization of the mean scaling relations of individual properties versus halo mass would not contain such second-parameter information; this

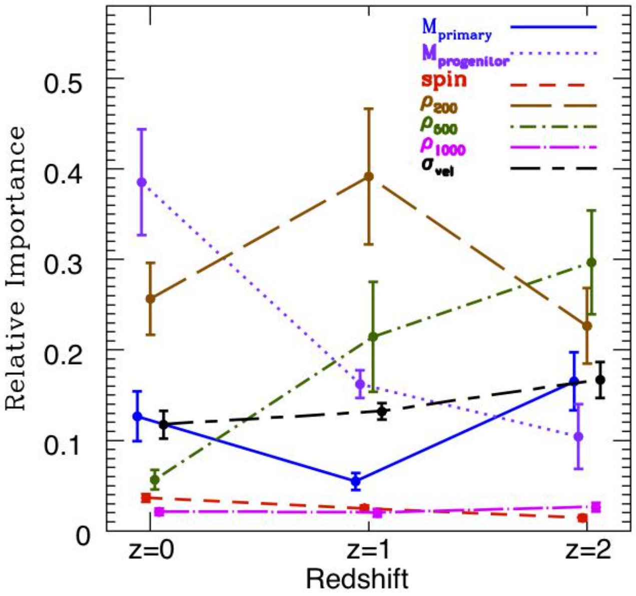

Figure 6. Relative importance of input features as determined in RF from $z = 0$ to $z = 2$. The error bars correspond to the standard deviation of each quantity from 10-fold cross validation. See §4.3 for exact definitions of these quantities.

is obtained purely because the ML has been trained on the full set of galaxy properties in the simulation, which contain these correlations. This hints that providing some baryonic information, particularly related to star formation rate, may substantially improve the ML predictions. We examine this later on.

4.3 Relative importance of inputs

The random forest algorithm uses a decision tree to generate the mapping from input to output quantities. For RF (as well as DT and ET) an input feature is ranked in importance according to the depth of the decision tree to which the feature in question is used in decision making. This thus provides a way to estimate, at least for these algorithms, which input variables are most responsible for governing the mapping from inputs to outputs. We examine these relative importances for our RF results in this section.

Fig. 6 shows the relative importance of the input features at $z = 0 - 2$. The input features are as described in §3.1, with two new aggregate definitions for the primary and progenitor halo masses: $\mathrm{M_{primary}}$ corresponds to the sum of importances for the current redshift and its immediate predecessor, while $\mathrm{M_{progenitor}}$ is the combined importance of the remaining earlier progenitors. We find these aggregate quantities are more stable and useful than considering each individual progenitor mass separately. The error bars reflect the variance over the 10 ensemble ML trials.

At $z = 0$, the progenitor mass and nearby (200 kpc) environment are the driving inputs, and combine to produce two-thirds of the total importance. Primary mass and velocity dispersion are also significant drivers at the ∼ 10% level each. At face value, this suggests that including the immediate mass growth (i.e. merger) history of dark matter halos is crucial for obtaining a good ML mapping.

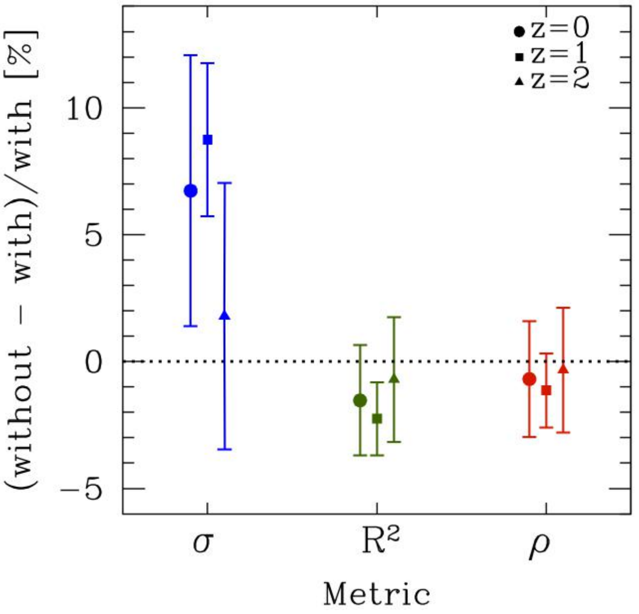

Figure 7. Percentage change in the three metrics σ, R^2 and ρ for M_* when merger history of DM halos is neglected during the ML training phase. The error bars correspond to the standard deviation from 10-fold cross validation. We deduce that neglecting halo merger history only slightly impacts the overall fits. This plot employs to RF; similar results are obtained for other ML methods.

Going to higher redshifts, the intermediate-scale (500 kpc) environment becomes much more important, while the progenitor mass commensurately reduces in importance. Note that these environmental measures are in comoving kpc, so in fact at $z \sim 1-2$, this corresponds to ∼ 200 physical kpc. At face value, this suggests that galaxies properties at all epochs are mostly driven by environmental processes on a ∼ 200 kpc physical scale, although one must be careful not to read too much physics into these importances. Also, the halo velocity dispersion is reasonably important, although in simulations this is very tightly correlated with the halo mass, so it is unclear whether this is providing distinct constraints.

Another interesting result is that at all epochs, the large-scale (Mpc) environment and spin are unimportant. The spin in particular may be surprising, given that simple galaxy models would associate spin as a determining factor for whether a galaxy collapses early and forms a massive elliptical (low spin) or it is a disky star-forming object until today (high spin). At least within this ML framework, it appears the halo spin is not an important predictor of galaxy quantities such as mass and SFR.

Given the importance of the progenitor mass in RF, we explored more directly the results if we removed the progenitor information from the inputs. For this, we retrain and rerun the RF network with the only halo mass information provided being the current mass, without any progenitor masses. This is expected to give a worse fit. We quantify how much worse in Fig. 7, which shows the percentage change in the scatter and correlation properties of the M_* distribution when halo merger history is neglected. A worse fit results in a larger σ, and a smaller R^2 and ρ. Despite the RF importance level, the change is relatively minor, with an increase in σ at the few to 10% level, and only a percent-level reduction in R^2 or ρ. This suggests that the progenitor information is actually sufficiently encoded in the remaining quantities that removing this quantity does not greatly impact the overall fits. This cautions against reading too much into the details of importance levels from the RF algorithm.

4.4 Improvement with baryonic inputs

N-body simulation only contain information about dark matter halos. However, it is possible to populate N-body simulations with galaxies using simplified prescriptions. A traditional approach is to use SAMs, but these typically contain a large number of free parameters. A more recent approach is to use so-called equilibrium (Davé et al. 2012) or bathtub-type models (Bouché et al. 2010; Lilly et al. 2013). In such models, galaxy evolution is fully captured via a relatively small number of physical parameters that quantify the baryon cycling process, which can be constrained to observations with well-characterized Bayesian posteriors. For instance, Mitra et al. (2015) found that with 8 free parameters that describe ejective feedback, preventive feedback, and wind recycling, it was possible to well predict the evolution of the stellar mass–halo mass, star formation rate–stellar mass, and mass-metallicity relations from $z = 0 - 2$. The outputs of such a model are, for a given halo with a mass assembly history, the resulting stellar mass, star formation rate, and metallicity of the halo's central galaxy. Hence by running such a model on a halo merger tree, it becomes possible to predict M_*, SFR, and Z for each halo (along with their histories), which can then be used as an input to ML rather than as an output. In the present case, the remaining outputs are then the H I and H_2 masses.

In this section, we examine how much improvement one can expect in M_{HI} and M_{H2} by providing the equilibrium model output quantities as *inputs* to ML, in addition to the full complement of dark matter properties. Fig. 8 shows the result, at $z = 0$ and $z = 1$. Comparing to the original determinations in Figs. 2 and 3, it is clear that the scatter σ is significantly reduced, and R^2 and ρ are higher. It is still not ideal, but clearly provides a step forward towards accurate predictions of the gas content of galaxies.

To understand which baryonic quantity provided this improvement, in Fig. 9, we show the relative importance of the input features when galaxy quantities M_*, SFR and Z are included on the input side of the network. It appears that the key new quantity that is most responsible for the improvement is the SFR, at least at $z = 0, 1$; at $z = 2$, the primary halo mass remains the most important. This is not surprising since the gas content is most closely related to the SFR; the metallicity and halo mass are also helpful, and in fact at $z = 0$ those quantities are the top 3 in importance. This experiment thus demonstrates that including some baryonic information as might be available from well-constrained analytic models can help to improve the gas predictions from ML particularly for H I.

5 COMPARING MACHINE LEARNING METHODS

So far we have focused on the RF algorithm for ML. In this section we show the tests that led to this decision, by comparing various ML methods with regards to our correlation scores.

Fig. 10 shows the fractional difference in the scatter (σ, blue), regression (R^2, green), and Pearson correlation coefficient (ρ, red), for our tested ML methods, normalized to the RF case. The leftmost points correspond to RF, and are by definition unity. Compared to RF, ET and SVR perform approximately as well, SVR perhaps even slightly better. MLP and kNN are significantly worse, with higher scatter and lower correlation coefficients. All three metrics degrade slightly when RF and SVR are combined with AdaBoost-Regressor (denoted as RF+ and SVR+), possibly due to over-fitting. Owing to this analysis, we chose to use Random Forests to quote our ML results.

To further explore ways to improve the method, we tried combining the results from our full ensemble of ML methods, either by taking the mean result among all the methods for each galaxy, or taking the median. Unfortunately, in neither case was there any noticeable improvement in the predictions. The issue likely traces back to the scatter. As shown in Fig. 2 and related figures, the scatter in ML predictions around the mean relations is too small compared to the true range. All other ML methods behave comparably – none are able to properly recover the full scatter in the relations. It appears all ML methods are like RF in their tendency to push the predictions towards the mean relations, and avoid outliers. It remains to be seen if there is an ML algorithm that is better suited to reproducing the full distribution of the training set rather than biasing the predictions towards the mean.

6 SUMMARY AND DISCUSSION

We have investigated machine learning as a framework to populate baryonic galaxies inside dark matter halos. We train on the MUFASA cosmological hydrodynamic simulations, and test the efficacy of various machine learning regression algorithms against a distinct subset of galaxies from the same simulation. Dark matter inputs to the machine learning are the halo masses over a range of cosmic ages, the halo densities in radii of 200, 500 and 1000 kpc, the halo spin and the halo velocity dispersion. The predicted outputs are stellar mass, H I and H_2 mass, star formation rate (SFR) and SFR-weighted metallicity both for the entire halo and the central galaxy.

Our main conclusion is that machine learning is able to correctly predict the mean baryonic properties well at

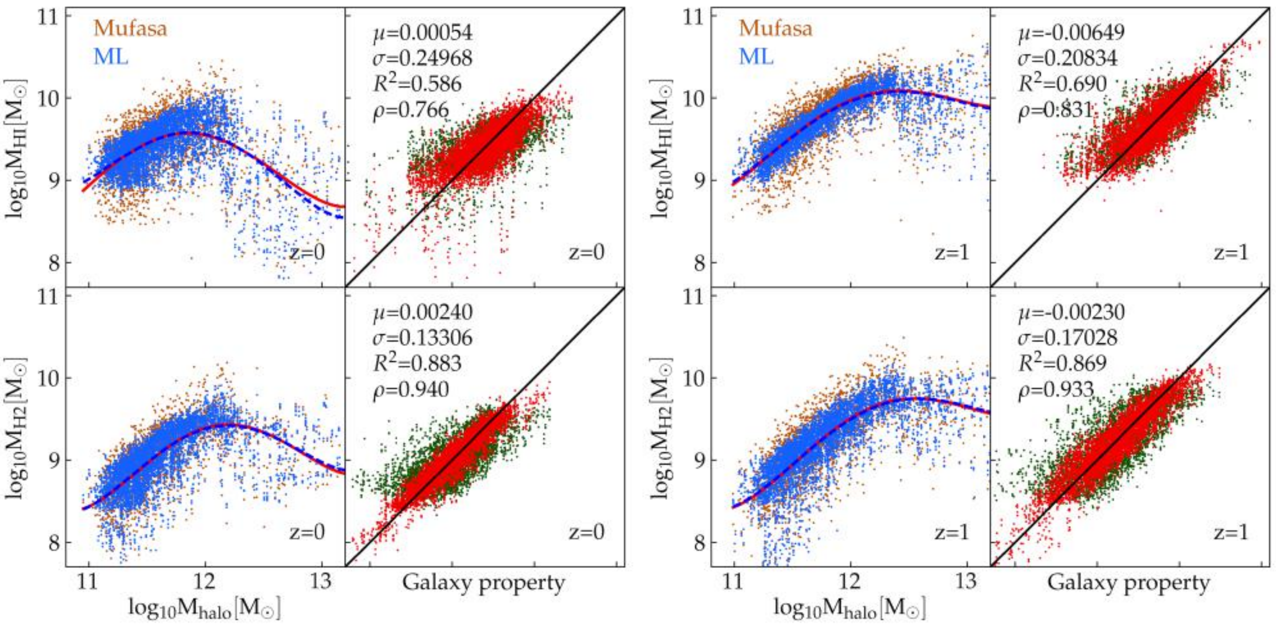

Figure 8. Galaxy H I (top panels) and H_2 (bottom panels) mass vs. host halo mass, at $z = 0$ (left 4 panels) and $z = 1$ (right 4 panels). In making the predictions, RF was trained using the usual halo quantities, as well as galaxy quantities M_*, SFR and Z. Inclusion of baryonic information of galaxies improves the H I and H_2 predictions considerably (red points). For comparison, see the H I and H_2 predictions in Fig. 2 and Fig. 3 (reproduced here by green points).

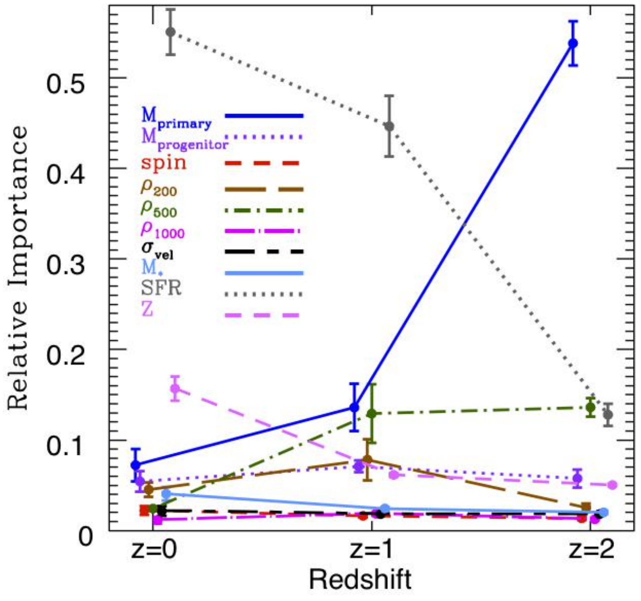

Figure 9. Relative importance of input features as a function of redshifts as in Fig. 6, now including the baryonic inputs. The dramatic improvement in predictions is driven primarily by including information about the current SFR (gray dotted line) of the galaxy; stellar mass and metallicity are important as well but clearly sub-dominant to SFR.

all halo masses, except at the low mass end where they are over-estimated which is likely an artifact of limited dynamic range. Somewhat of a surprise was the result that including halo merger history leads to only marginally improved results.

We find that none of the algorithms capture the full range of observed scatter in baryonic properties. As a result, ensembling the outputs from the various algorithms does not lead to improved results. The machine learning predictions do however capture the second-parameter correlations implicit in the MUFASA simulations which encode important physics, e.g. high SFR is linked to higher gas content and lower metallicity.

We show that if some baryonic information is available through additional pre-modeling such as provided by the equilibrium model, the machine learning algorithms can exploit this and can match the observed scatter much better, giving improvements of up to 100%, especially in H_2 predictions. When including the baryonic information the most important feature in the machine learning fits is the star formation rate.

Finally, we compare seven common regression algorithms (and EDEN – a bespoke evolutionary deep learning algorithm, but not shown in this paper). We also explore an ensemble approach looking at the mean and median of the outputs of the various algorithms. Performance is broadly similar, with the random forests and support vector regression marginally better than others. This suggests that the limitations are intrinsic to the amount of information contained in the dark matter distribution, rather than in the performance of the ML algorithm.

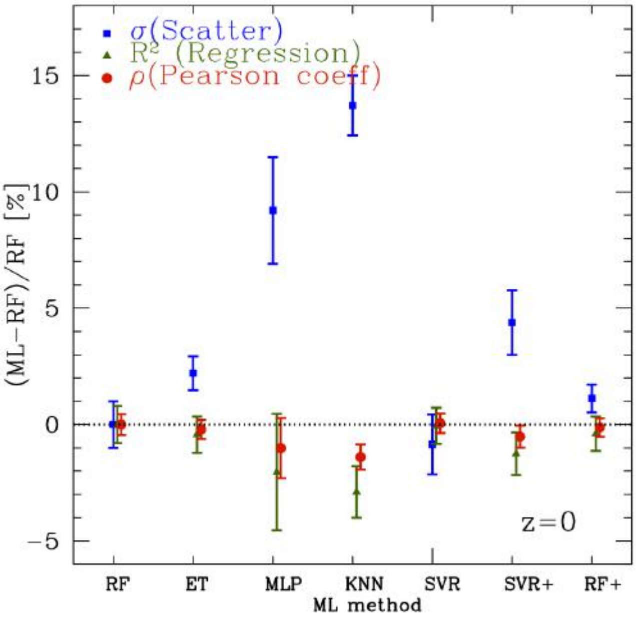

Figure 10. The performance metrics (scatter σ, regression R^2, and Pearson correlation coefficient ρ), quoted in Fig. 2 at $z = 0$ for galaxy stellar mass M_*, are plotted for the ML algorithms explored in this paper. The results are shown relative to the metric values for RF. For each algorithm, the error bars correspond to the standard deviation of each quantity from 10-fold cross validation. SVR+ and RF+ correspond to the boosted versions of SVR and RF using AdaBoostRegressor. All ML routines have comparable regression (R^2, green) and Pearson (ρ, red) coefficients. Compared to RF, the scatter (σ, blue) is $\sim 10 - 15\%$ larger for MLP and kNN. The corresponding plots for the other galaxy quantities ($\mathrm{M_{HI}}, \mathrm{M_{H2}}, \mathrm{SFR}, Z$), and for quantities predicted at $z = 1$ and $z = 2$, are similar. We deduce that for the most part the different algorithms all deliver similar performance, at least for R^2 and ρ.

The ML approach to populating galaxies into halos has intrinsic limitations, some of which can be at least partially mitigated. For one, any ML framework will only be as good as the training data used. In our case, we have used the MUFASA simulation, which provides quite good agreement with a wide range of (though not all) galaxy properties (Davé et al. 2016, 2017b,a; Rafieferantsoa & Davé 2018); other modern simulations can achieve similar agreement albeit with varying input physics (Somerville & Davé 2015). As simulations continue to improve in realism, applying this ML framework to the latest galaxy formation models will provide increasingly robust predictive power.

A second limitation is that the dynamic range in the hydrodynamic simulation will naturally be smaller than that of the target N-body simulation (else there would be no reason to do the ML). As such, populating large-volume simulations will require an extrapolation beyond the training set. One way to mitigate this would be to run zoom simulations targeting dwarfs and galaxy clusters (Cui et al. 2016) using the exact same physical model, thereby providing anchor points at large and small halo masses for training.

Another limitation is that we are training on halos and galaxies taken from a hydro simulations, but the gas dynamics could potentially have a non-negligible impact on halo properties. A way to mitigate this would be to run a pure dark matter version of the same hydro simulation, and then match up halos individually, and use the halo properties from the N-body run rather than the hydro run. This would more accurately mimic the situation in a pure dark matter run. Indeed, the ML framework could even calibrate how halos are modified owing to baryonic processes, which is interesting in of itself.

Finally, we have concentrated here on central galaxies, whereas fully populating an N-body run will require including the satellites as well. This is perhaps the most challenging aspect, because it will require identifying subhalos in N-body simulations as sites to populate with baryonic satellites, but sub-halo identification and numerical resolution issues in N-body simulations can be severe (van den Bosch & Ogiya 2018). An alternative approach would be to instead use the merger tree from the N-body simulation, and follow the satellites using a calibrated analytic model that does not suffer from numerical convergence issues (Simha & Cole 2017).

Modulo the above caveats that we aim to tackle in future work, our results suggest that machine learning is a promising approach to populate galaxies into halos in large N-body simulations, which is useful for a wide variety of galaxy formation and cosmology applications. One such application is H I intensity mapping as will be done with upcoming surveys such as CHIME, HIRAX, and MeerKLASS on MeerKAT. For this, the ML would need to output the halo H I content which is straightforward, and this would be mapped into halos from a dark matter simulation. Another application would be weak lensing galaxy shape correlations, for which the ML would need to be trained on galaxy ellipticities which would again be straightforward within our framework although we have not done it here. Future work includes identifying dark matter features that can act as suitable proxies for baryonic information, tuning more sophisticated ML algorithms such as deep learning, using recurrent neural networks to learn the temporal evolution of halos, applying to larger simulations with more galaxies to avoid being limited by the size of training samples, and marrying it with simple analytic models to additionally input an estimate of the instantaneous SFR. The flexibility of ML and the ability to assess its accuracy based on testing sets suggests that it is a worthwhile alternative approach to connecting galaxy baryonic properties to dark matter halos within large-scale structure.

ACKNOWLEDGEMENTS

The authors thank the referee for providing constructive comments and help in improving the contents of this paper. SA acknowledges Laboratoire Univers et Théories (LUTh) at the Observatoire de Paris, for allowing its computing resources be used for running SCIKIT-LEARN machine learning library. RD acknowledges support from the South African Research Chairs Initiative and the South

African National Research Foundation. RD further acknowledges long-term visitor support provided by the Simons Foundation's Center for Computational Astrophysics and the Distinguished Visitor Program at Space Telescope Science Institute. The Mufasa simulations were run on the Pumbaa astrophysics computing cluster hosted at the University of the Western Cape, which was generously funded by UWC's Office of the Deputy Vice Chancellor, and were run with revision e77f814 of Gizmo hosted at `https://bitbucket.org/rthompson/gizmo`.

REFERENCES

Bishop C. M., 1995, Neural Networks for Pattern Recognition. Oxford Univ. Press, New York

Bothwell M. S., Maiolino R., Kennicutt R., Cresci G., Mannucci F., Marconi A., Cicone C., 2013, MNRAS, 433, 1425

Bouché N., et al., 2010, ApJ, 718, 1001

Cole S., Lacey C., 1996, MNRAS, 281, 716

Cole S., Aragon-Salamanca A., Frenk C. S., Navarro J. F., Zepf S. E., 1994, MNRAS, 271, 781

Courtin J., Rasera Y., Alimi J.-M., Corasaniti P.-S., Boucher V., Füzfa A., 2011, MNRAS, 410, 1911

Croton D. J., et al., 2006, MNRAS, 365, 11

Cui W., et al., 2016, MNRAS, 458, 4052

Davé R., Finlator K., Oppenheimer B. D., 2011, MNRAS, 416, 1354

Davé R., Finlator K., Oppenheimer B. D., 2012, MNRAS, 421, 98

Davé R., Thompson R., Hopkins P. F., 2016, MNRAS, 462, 3265

Davé R., Rafieferantsoa M. H., Thompson R. J., 2017a, preprint, (arXiv:1704.01135)

Davé R., Rafieferantsoa M. H., Thompson R. J., Hopkins P. F., 2017b, MNRAS, 467, 115

Dufourq E., Bassett B. A., 2017, preprint, (arXiv:1709.09161)

Gabor J. M., Capelo P. R., Volonteri M., Bournaud F., Bellovary J., Governato F., Quinn T., 2016, A&A, 592, A62

Genel S., et al., 2014, MNRAS, 445, 175

Hopkins P. F., 2015, MNRAS, 450, 53

Kamdar H. M., Turk M. J., Brunner R. J., 2016, MNRAS, 457, 1162

Kauffmann G., White S. D. M., Guiderdoni B., 1993, MNRAS, 264, 201

Kereš D., Katz N., Weinberg D. H., Davé R., 2005, MNRAS, 363, 2

Krumholz M. R., McKee C. F., Tumlinson J., 2009, ApJ, 699, 850

Lara-López M. A., Bongiovanni A., Cepa J., Pérez García A. M., Sánchez-Portal M., Castañeda H. O., Fernández Lorenzo M., Pović M., 2010, A&A, 519, A31

Lilly S. J., Carollo C. M., Pipino A., Renzini A., Peng Y., 2013, ApJ, 772, 119

Mannucci F., Cresci G., Maiolino R., Marconi A., Gnerucci A., 2010, MNRAS, 408, 2115

Mitra S., Davé R., Finlator K., 2015, MNRAS, 452, 1184

Muratov A. L., Kereš D., Faucher-Giguère C.-A., Hopkins P. F., Quataert E., Murray N., 2015, MNRAS, 454, 2691

Naab T., Ostriker J. P., 2017, ARA&A, 55, 59

Pedregosa F., et al., 2012, preprint, (arXiv:1201.0490)

Potter D., Stadel J., Teyssier R., 2017, Computational Astrophysics and Cosmology, 4, 2

Rafieferantsoa M., Davé R., 2018, MNRAS, 475, 955

Remus R.-S., Dolag K., Hoffmann T., 2017, Galaxies, 5, 49

Salim S., Lee J. C., Ly C., Brinchmann J., Davé R., Dickinson M., Salzer J. J., Charlot S., 2014, ApJ, 797, 126

Salim S., Lee J. C., Davé R., Dickinson M., 2015, ApJ, 808, 25

Schaye J., et al., 2015, MNRAS, 446, 521

Simha V., Cole S., 2017, MNRAS, 472, 1392

Somerville R. S., Davé R., 2015, ARA&A, 53, 51

Springel V., et al., 2005, Nature, 435, 629

Vogelsberger M., et al., 2014, MNRAS, 444, 1518

White S. D. M., Frenk C. S., 1991, ApJ, 379, 52

van den Bosch F. C., Ogiya G., 2018, MNRAS, 475, 4066