Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **514,** 4026–4045 (2022) https://doi.org/10.1093/mnras/stac1609
Advance Access publication 2022 June 14

The scatter in the galaxy–halo connection: a machine learning analysis

Richard Stiskalek [1]★ Deaglan J. Bartlett [2] Harry Desmond [2,3,4] and Dhayaa Anbajagane [5,6]

[1] *Universitäts-Sternwarte, Ludwig-Maximilians-Universität München, Scheinerstr. 1, D-81679 München, Germany*
[2] *Astrophysics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK*
[3] *McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA*
[4] *Institute of Cosmology & Gravitation, University of Portsmouth, Dennis Sciama Building, Portsmouth PO1 3FX, UK*
[5] *Department of Astronomy and Astrophysics, University of Chicago, 5640 S. Ellis Ave, Chicago, IL 60637, USA*
[6] *Kavli Institute for Cosmological Physics, University of Chicago, 5640 S. Ellis Ave, Chicago, IL 60637, USA*

Accepted 2022 June 7. Received 2022 May 30; in original form 2022 February 28

ABSTRACT

We apply machine learning (ML), a powerful method for uncovering complex correlations in high-dimensional data, to the galaxy–halo connection of cosmological hydrodynamical simulations. The mapping between galaxy and halo variables is stochastic in the absence of perfect information, but conventional ML models are deterministic and hence cannot capture its intrinsic scatter. To overcome this limitation, we design an ensemble of neural networks with a Gaussian loss function that predict probability distributions, allowing us to model statistical uncertainties in the galaxy–halo connection as well as its best-fitting trends. We extract a number of galaxy and halo variables from the Horizon-AGN and IllustrisTNG100-1 simulations and quantify the extent to which knowledge of some subset of one enables prediction of the other. This allows us to identify the key features of the galaxy–halo connection and investigate the origin of its scatter in various projections. We find that while halo properties beyond mass account for up to 50 per cent of the scatter in the halo-to-stellar mass relation, the prediction of stellar half-mass radius or total gas mass is not substantially improved by adding further halo properties. We also use these results to investigate semi-analytic models for galaxy size in the two simulations, finding that assumptions relating galaxy size to halo size or spin are not successful.

Key words: methods: numerical – galaxies: fundamental parameters – galaxies: haloes – dark matter.

1 INTRODUCTION

The dynamics of the Universe are governed by dark matter (DM), yet our knowledge of this in galaxies comes entirely from the photometry and spectroscopy of baryons, with five times lower average density. Relating the behaviour of baryons and DM is a key task of astrophysics. DM may be usefully described by the halo model, in which DM clumps into gravitationally bound quasi-spherical structures called haloes, which in turn host cold baryons in the form of galaxies. In this regime, relating baryons and DM is tantamount to discovering the correlations between galaxies and haloes: the galaxy–halo connection. This would be the first step towards uncovering the physical processes responsible for the baryon–dark matter relation in a bottom-up approach to galaxy formation.

With DM being unobservable, the galaxy–halo connection cannot be inferred directly. Instead we are forced to do this either empirically – by studying the properties of galaxies such as their kinematics which are determined by DM – or by postulating scaling relations between galaxy and halo variables and calculating how they would affect observables: examples include halo abundance matching (Kravtsov et al. 2004; Behroozi, Conroy & Wechsler 2010), the halo occupation distribution (Berlind & Weinberg 2002; and embellishments thereof; Paranjape et al. 2015; Hearin et al. 2016; Hadzhiyska et al. 2021), the angular momentum (AM) partition model of Mo, Mao & White (1998), and the rich field of semi-analytic modelling (see Baugh 2006 for a review). A third approach is to infer the galaxy–halo connection *post hoc* from cosmological hydrodynamical simulations, which model numerically the coevolution of baryons and DM from their initial conditions with a mix of physical prescriptions and semi-empirical subgrid models and feedback schemes. While each simulation implicitly produces its own galaxy–halo connection, this is often far from clear from the raw output of the simulation, nor is it obvious how this comes about or what the implications are for other types of modelling.

We focus here on the third approach. As the correlations between galaxy and halo properties in a given simulation are non-linear and have unknown functional forms *a priori*, the best approach for reconstructing them is to leverage *machine learning* (ML) whereby a computer can rapidly test a huge range of possibilities. Indeed, ML the galaxy–halo connection is already becoming a cottage industry, particularly in supervised settings where algorithms are trained on data sets generated by empirical, semi-analytic, or simulation-based models. For example, von Marttens et al. (2021) explored a variety of supervised models, such as Support Vector Machine or Random Forests. Among the tree-based models, the Random Forest has been used extensively to study the galaxy–halo connection (Man et al. 2019; Delgado et al. 2021; Lovell et al. 2022) along with e.g. Gradient Boosted Decision Trees (Chen et al. 2021). Similarly, the use of neural network (NN)-based models is becoming increasingly popular, with fully connected networks (Moster et al. 2021; Shao et al. 2021), Graph Neural Networks (Villanueva-Domingo et al.

★ E-mail: richard.stiskalek@protonmail.com

2021), and Convolutional Neural Networks (de los Rios et al. 2021) all having recently been brought to bear. The advantage of this approach is that – if successful – it enables a faithful reconstruction of the galaxy–halo connection of a simulation, as well as aiding its interpretability. The latter may be done either with a feature importance analysis, or by identifying analytic expressions that are successful at describing the connection, e.g. through symbolic regression (Delgado et al. 2021). Recent work has also demonstrated the power in utilizing dedicated ML interpretibility tools – such as Google DeepDream (Ntampaka et al. 2019) and the Shapley values (Machado Poletti Valle et al. 2021; Anbajagane, Evrard & Farahi 2022a) – in determining key galaxy or halo properties. From a practical point of view, ML approaches are also useful for painting galaxies on to the haloes produced in N-body, DM only simulations – or haloes on to observed galaxies – at low computational expense compared to full hydrodynamical simulation. The disadvantage is that the results are tied to a particular simulation with no guarantee of accuracy should that simulation fail to describe the real Universe. Indeed, recent results have indicated that ML algorithms trained on one simulation will not in general be successful at reproducing others (Villaescusa-Navarro et al. 2021a).

A key limitation of many previous ML analyses is that the scatter in predicted variables tends to be underestimated (Kamdar, Turk & Brunner 2016; Agarwal, Davé & Bassett 2018) because mock catalogues are generated using a deterministic ML algorithm such that a given input always implies the same output. Such algorithms do not reflect the nature of the galaxy–halo connection, which is inherently stochastic in the absence of complete information. The main aim of this work is to construct and showcase a ML algorithm which returns not only best-fitting predictions for galaxy properties given halo properties (or vice versa) but also the *statistical uncertainty* on those predictions. This latter goal has been pursued by only a minor subset of existing ML works (e.g. Kodi Ramanah et al. 2020; Kodi Ramanah, Wojtak & Arendse 2021; Ho et al. 2021; Eisert et al. 2022). It is crucial to capture the scatter in the galaxy–halo connection, and this will allow us to quantify the extent to which this scatter correlates with other galaxy/halo properties such that it can be reduced or removed when including those properties in the feature set. With this algorithm in hand, we ask which galaxy properties can be predicted from their hosts' (sub)halo properties and vice versa, and which features are the most important for making these predictions. To test the robustness of these results to modelling assumptions we use two different cosmological hydrodynamical simulations, Horizon-AGN (H-AGN) and IllustrisTNG, as training sets. Along the way we will determine the extent to which some popular semi-empirical galaxy formation models – halo abundance matching and DM-baryon AM partition – are successful at describing the simulations' outputs.

In Section 2 we introduce the simulations and describe the simulated galaxy and halo properties we use as features and targets for our ML algorithm. Our ML architecture is outlined in Section 3, where we also discuss how we reduce the impact of correlations in the input features and feature selection. We present the correlations of the simulated galaxy–halo connections in Section 4, and their implications for semi-empirical models in Section 5. Section 6 discusses the results and suggests further work, and Section 7 concludes.

2 SIMULATED DATA SETS

We describe the simulations we use in Section 2.1, and the definitions of the halo/galaxy properties and how we compute them in Section 2.2.

2.1 Cosmological simulations

2.1.1 *Horizon-AGN*

For one of our training sets, we consider the galaxies and their corresponding haloes in the $(100\,\mathrm{Mpc}/h)^3$ cosmological hydrodynamical simulation H-AGN[1] (Dubois et al. 2014). This adopts a standard ΛCDM cosmology consistent with WMAP-7 (Komatsu et al. 2011) and therefore has total matter density $\Omega_\mathrm{m} = 0.272$, dark energy density $\Omega_\Lambda = 0.728$, amplitude of the matter power spectrum $\sigma_8 = 0.81$, baryon density $\Omega_\mathrm{b} = 0.045$, Hubble constant $H_0 = 70.4\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$, and power spectrum slope $n_\mathrm{s} = 0.967$.

Using 1024^3 DM particles to give a DM mass resolution of $M_{\mathrm{DM,\,res}} = 8 \times 10^7\,\mathrm{M_\odot}$, the simulation is run with the Adaptive Mesh Refinement code RAMSES (Teyssier 2002), with cells refined whenever the mass in that cell surpasses eight-times the initial mass resolution, up to an effective physical resolution of 1 kpc. Stars can form when the hydrogen number density exceeds $n_\mathrm{H} \geq 0.1\,\mathrm{H\,cm^{-3}}$, yielding a stellar mass resolution of $M_{\star,\,\mathrm{res}} \approx 2 \times 10^6\,\mathrm{M_\odot}$, and this process follows a Schmidt law with a 1 per cent efficiency (Rasera & Teyssier 2006). Baryonic effects such as background UV heating, gas cooling (including a contribution from metals), feedback from stellar winds, and type Ia and type II supernovae assuming a Salpeter initial mass function (IMF; Dubois & Teyssier 2008; Kimm 2012) are also incorporated in the simulation. Following Dubois et al. (2013) we split the gas associated with each galaxy into the following four phases by its temperature T and hydrogen number density n_H

(i) Star forming: $n_\mathrm{H} \geq 0.1$ H cm^{-3}, which is the star formation threshold in H-AGN,
(ii) Hot diffuse: $n_\mathrm{H} < 0.1$ H cm^{-3} and $T \geq 10^5$ K,
(iii) Cold collimated: $100\,n_\mathrm{avg} \leq n_\mathrm{H} < 0.1$ H cm^{-3} and $T < 10^5$ K, where n_avg is the mean hydrogen number density of the Universe,
(iv) Cold diffuse: $n_\mathrm{H} < 100\,n_\mathrm{avg}$ and $T < 10^5$ K.

Using the ADAPTAHOP structure finder (Aubert, Pichon & Colombi 2004; Tweed et al. 2009), galaxies (haloes) consist of a minimum of 50 (100) particles, and are identified when the stellar (DM) smoothed density field calculated using the 20 nearest neighbours exceeds 178 times the mean total matter density (Gunn & Gott 1972). A shrinking sphere approach (Power et al. 2003) finds the position of the densest particle, which is taken to be the centre of the structure. Galaxies are then paired to haloes by considering each halo in turn from the most to least massive, and choosing the most massive unassigned galaxy which lies within 10 per cent of the virial radius, R_vir, of the halo as its partner (Chisari et al. 2017; Bartlett et al. 2021). All galaxy and subhalo properties for H-AGN are computed using all available particles within R_vir, and this includes a minor fraction of gravitationally unbound material as well.

We enforce the following threshold on our resulting catalogue: halo mass $M_\mathrm{h} > 10^{10}\,\mathrm{M_\odot}$, corresponding to 125 particles, stellar mass $M_\star > 10^{8.3}\,\mathrm{M_\odot}$, corresponding to 100 particles, and the galaxy average star forming rate over the last 300 Myrs be greater than 0 since some star-forming galaxies may be assigned a star formation rate (SFR) of 0 due to the true SFR being below the resolution limit.

2.1.2 *TNG*

Our second training set consists of the galaxies and haloes from the TNG100-1 simulation[2] (Marinacci et al. 2018; Naiman et al.

[1] http://www.horizon-simulation.org/about.html
[2] https://www.tng-project.org/

2018; Nelson et al. 2018; Pillepich et al. 2018b; Springel et al. 2018; henceforth TNG). Like H-AGN, TNG adopts a ΛCDM cosmology, however the parameters are those from the 2015 analysis of *Planck* (Planck Collaboration 2016): $\Omega_{\rm m} = 0.3089$, $\Omega_\Lambda = 0.6911$, $\Omega_{\rm b} = 0.0486$, $\sigma_8 = 0.8159$, $H_0 = 67.74\ {\rm km\,s^{-1}\,Mpc^{-1}}$, and $n_{\rm s} = 0.9667$.

This hydrodynamical simulation is run using the moving-mesh code AREPO (Springel 2010) with a volume of $(75\ {\rm Mpc}/h)^3$, which is ∼50 per cent of H-AGN's volume. The Springel & Hernquist (2003) model of star formation and pressurization of the multi-phase interstellar medium is used, such that stars form stochastically according to the Kennicutt–Schmidt relation with a Chabrier IMF (Chabrier 2003) when the gas density exceeds $n_{\rm H} \approx 0.1\ {\rm cm^{-3}}$. Numerous baryonic effects are included such as feedback from active galactic nuclei (AGNs), galactic winds, metal cooling, and ideal magneto-hydrodynamics; for full details see Weinberger et al. (2017) and Pillepich et al. (2018a).

The Friends-of-Friends (FoF) algorithm is run on the DM particles, with a linking length $b = 0.2$, to determine the main haloes, and other particle species are associated with the halo of their closest DM particle. The SUBFIND algorithm (Springel et al. 2001; Dolag et al. 2009) is then used on all particle species to identify substructures, and hence galaxies. The most massive subhalo within a main halo is defined to be the central subhalo, and the others are classified as satellite subhaloes.

The SUBFIND algorithm, when identifying subhaloes within FoF haloes, also performs an unbinding step to remove any material that is not gravitationally bound to the subhalo. All subhalo properties for TNG are computed using only the bound particles. Matching H-AGN, we define our sample so that halo mass $M_{\rm h} > 10^{10}\ {\rm M_\odot}$, corresponding to 1300 particles, stellar mass $M_\star > 10^{8.3}\ {\rm M_\odot}$, corresponding to 140 particles, and the instantaneous SFR in the subhalo is larger than 0.

2.2 Halo and galaxy property definitions

We describe here the full set of halo and galaxy properties used in this work, including differences in definitions used by each simulation and the public catalogues utilized for some of the properties.

In general, a key difference between simulation properties is that the H-AGN properties are computed using all particles (bound or unbound) within a spherical aperture cut, whereas the TNG properties use only the gravitationally bound particles without any such aperture cut. This arises from differences in the subhalo finders used by each simulation.

2.2.1 *Subhalo/galaxy mass and radius*

The subhalo radius is utilized to define the subhalo mass, in addition to being an input property in our method, so we begin by describing it. For H-AGN, this is obtained by binning the DM particles in concentric ellipsoidal shells, and determining the outermost shell for which the virial theorem is satisfied to within 20 per cent. The virial radius, $R_{\rm vir}$, is defined to be the geometric mean of the three axis radii of this ellipsoid. If no such shell exists, then the virial radius is defined to be the largest radius within which the mean density exceeds that expected from a spherical top-hat model. For parent haloes this procedure yields a near one-to-one relation between the radius and virial mass. This matches the typical virial radius definition,

$$\bar{\rho}(< R_{\rm vir}) \equiv \frac{M(< R_{\rm vir})}{\frac{4}{3}\pi R_{\rm vir}^3} = \Delta_{\rm vir}\rho_c, \tag{1}$$

where $\bar{\rho}$ is the average enclosed density, $M(< r)$ is the enclosed mass, $\Delta_{\rm vir}$ is the overdensity as defined in Bryan & Norman (1998), and ρ_c is the critical density of the Universe at the present epoch. For subhaloes the radius specified in equation (1) acts as an upper limit, with a significant fraction having smaller sizes. Therefore, H-AGN subhaloes' sizes may be an informative quantity even when their masses are already included in the feature set.

In TNG parent haloes the virial radius is defined to obey equation (1). For TNG, we define the subhalo radius as the radial distance to the furthest gravitationally bound DM particle in the subhalo particle set determined by the SUBFIND algorithm. Henceforth, we refer to all subhalo radii as R_h. The subhalo masses for both H-AGN and TNG are then simply the total mass of all available DM particles within $r < R_{\rm h}$. In TNG subhaloes we define the virial mass of the host halo as $M_{\rm vir}^{\rm P} \equiv M(< R_{\rm vir})$. On the other hand, for H-AGN subhaloes we define $M_{\rm vir}^{\rm P}$ to be the virial DM mass of their parent haloes.

The stellar/gas mass of a subhalo ($M_\star$, $M_{\rm gas}$) in each simulation is defined in a way analogous to how their total subhalo mass is defined above. For H-AGN, it is the total mass of all stars/gas within $R_{\rm vir}$, whereas for TNG it is the total mass of all stars/gas bound to the subhalo. In TNG, the stellar half-mass radius, $R_{1/2}$ is then defined as the radius containing half of the total stellar mass. In H-AGN we project the stellar particles on to three orthogonal planes, and compute the two-dimensional radius for each plane which contains half the stellar mass. $R_{1/2}$ is then the geometric mean of these radii.

2.2.2 *Dark matter subhalo concentration*

The density profiles of subhaloes can be parameterized according to the Navarro–Frenk–White (NFW) model (Navarro, Frenk & White 1996),

$$\rho\,(r) = \frac{\rho_0}{(r/r_{\rm s})\,(1 + r/r_{\rm s})^2}, \tag{2}$$

where ρ_0 and $r_{\rm s}$ are model parameters denoting the characteristic density and the scale radius, respectively. Once we fit the model to the profile of a given subhalo, we define the concentration as

$$c_{\rm h} \equiv \frac{R_{\rm h}}{r_{\rm s}}. \tag{3}$$

Note that this is analogous to the standard definition but using $R_{\rm h}$ rather than $R_{\rm vir}$. For TNG, r_s is obtained by fitting the *enclosed mass* NFW profile rather than the density profile (as is done for H-AGN) where the two profiles are related by a simple radial, volume integral.

For central subhaloes, we compared our TNG catalogue with the TNG mass–concentration relation from Anbajagane et al. (2022a), who performed a traditional density profile fit to determine r_s, and verified that the difference between the two r_s–mass relations is less than five per cent. For central subhaloes, we also confirmed that defining $c_{\rm h}$ using $R_{\rm h}$ instead of a traditional spherical overdensity radius results in a $c_{\rm h}$–$M_{\rm h}$ relation that still captures the relevant mass-dependent features induced by baryonic effects (see 'wiggle' features in Anbajagane et al. 2022a, fig. 3).

2.2.3 *Dark matter subhalo and galaxy shapes*

We parameterize the shape of a subhalo/galaxy by first calculating the inertia tensor

$$I_{ij}^a = \frac{1}{M_{\rm a}} \sum_n m_{\rm a}^{(n)} x_i^{(n)} x_j^{(n)}, \tag{4}$$

where the sum runs over the particles of type 'a' (where 'a' is DM when computing DM subhalo shapes, and stars when computing galaxy shapes), each of mass $m_a^{(n)}$ and at position $\boldsymbol{x}^{(n)} = \left(x_1^{(n)}, x_2^{(n)}, x_3^{(n)}\right)^{\mathsf{T}}$. The total mass of the particle set is denoted M_a. We diagonalize I_{ij}^a to find the eigenvalues $\mathcal{A}_a^2 \geq \mathcal{B}_a^2 \geq \mathcal{C}_a^2$, where $\mathcal{A}_a, \mathcal{B}_a, \mathcal{C}_a > 0$ (Chisari et al. 2015). We use the ratios of the eigenvalues (i.e. the principal axes) to obtain two shapes,

$$s_a \equiv \mathcal{C}_a/\mathcal{A}_a, \quad q_a \equiv \mathcal{B}_a/\mathcal{A}_a, \tag{5}$$

conforming to the notation of denoting halo properties with subscript 'h' and galaxy properties with subscript 'g'.

As with DM concentration, we compare the TNG central subhalo shapes with the TNG catalogues of Anbajagane et al. (2022a), who computed DM shapes by selecting particles within a chosen spherical aperture. Given we instead use the full DM subhalo (as determined by SUBFIND) without any spherical cut, the shapes we find are 5−10 per cent more elliptical, i.e. the ratios $s_{\rm h}$ and $q_{\rm h}$ are lower. This is an expected feature due to the difference in particle selection (Zemp et al. 2011). However, the functional form of the shape-mass relation is the same regardless of whether such an aperture cut is used; in particular, the mass-dependent slopes decrease linearly (in log-mass) towards higher masses (see Anbajagane et al. 2022a, fig. 5).

2.2.4 *Spins*

The spins of the DM, $\lambda_{\rm h}$, and the stars, λ_g, are computed as,

$$\lambda_{\{h,\,g\}} = \frac{|\vec{J}_{\{h,\,g\}}|}{\sqrt{2} M_{\rm h} R_{\rm h} V_{\rm h}} \tag{6}$$

where $\vec{J}_{\{h,\,g\}}$ is the sum of the angular momenta of the particles, either DM or stars, within the subhalo radius as defined above. The normalization is the self-similar expectation for the total AM in a halo, defined using its mass and radius. The velocity scale is $V_{\rm h} = \sqrt{GM_{\rm h}/R_{\rm h}}$, with the mass and radius definitions described above. This definition of spin is analogous to that of Bullock et al. (2001), but with different mass/radius conventions.

2.2.5 *Other galaxy properties*

We use the time-averaged SFR, defined as

$$\text{SFR} = \frac{M_\star(\Delta t)}{\Delta t}, \tag{7}$$

where $M_\star(\Delta t)$ is the total stellar mass formed over the last Δt years. For the time-scale, we use 300 Myr for H-AGN and 200 Myr in TNG, with the difference purely set by the available catalogues. For TNG, we obtain this quantity from the publicly released catalogues (Donnari et al. 2019; Pillepich et al. 2019).

The mass-weighted mean stellar metallicity, Z, is computed using stellar particles within twice the stellar half-mass radius of the subhalo centre. We also consider the galaxy rest-frame photometry, namely the absolute magnitudes M_u, M_g, M_r, M_i computed using the *u, g, r, i* band filters of the Sloan Digital Sky Survey (Kaviraj et al. 2017; Laigle et al. 2019). The magnitude of a galaxy is obtained by summing up the luminosities of individual stellar particles.

For TNG, we also use $M_{\rm H\,I}$, which is the total neutral hydrogen mass in the subhalo, and this is once again obtained from public catalogues (Diemer et al. 2018, 2019).

We show the correlations of M_h with $M_\star$, $R_{1/2}$, and $M_{\rm gas}$ for central and satellite galaxies in H-AGN and TNG in Fig. 1. The full set of galaxy and halo properties is summarized in Table 1.

3 METHODOLOGY

In Section 3.1 we describe how we calculate bivariate scatters in the simulations to set the stage for full multivariate scatter calculated by our NNs, then in Section 3.2 discuss the data preprocessing, in Section 3.3 the architecture of our NN model and in Section 3.4 the interpretation of its predictions. Lastly in Section 3.5 and Section 3.6 we describe the Extra-Trees model and feature selection, respectively.

3.1 Quantifying the bivariate scatter and its uncertainty

A key property of the galaxy–halo connection is its intrinsic scatter, i.e. the variation across the population of galaxy/halo quantities conditional on one or more others. Developing a framework to model this – along with its uncertainty – is a key goal of our work. A simple benchmark for the scatter may be obtained by specializing to the correlation between just one galaxy property and one halo property. In this case the scatter provides an upper limit on how well one can determine one property from the other (i.e. in the absence of any further information) and can be calculated straightforwardly from the simulated quantities. If the residuals are correlated with a third property then including that in the relation will reduce the scatter and hence allow a more precise prediction of the target variable.

Consider the correlation between two quantities X and Y, where we wish to predict Y given X. To quantify the scatter in the X−Y relation and its uncertainty, we first sort the data by X into n_X bins between the minimum and maximum values of X (typically we choose bins of width 0.15 dex when X is the halo mass). We assume that in each bin B_i, Y is Gaussian distributed about a mean μ_i with standard deviation σ_i. Assuming each galaxy is independent, the likelihood for the observed Y in bins $\{B_i\}$ given these parameters is

$$\mathcal{L}\left(\{Y_g\}_i|\mu_i, \sigma_i\right) = \prod_{g \in B_i} \mathcal{N}\left(Y_g|\mu_i, \sigma_i^2\right), \tag{8}$$

where Y_g is the value of Y for galaxy g, $\{Y_g\}_i \equiv \{Y_g\colon g \in B_i\}$, and $\mathcal{N}(Y_g|\mu_i, \sigma_i^2)$ is a Gaussian of mean μ_i and variance σ_i^2. Given a prior, $p(\mu_i, \sigma_i)$, on μ_i and σ_i, we use Bayes' theorem to obtain the posterior distributions of μ_i and σ_i,

$$p\left(\mu_i, \sigma_i^2|\{Y_g\}_i\right) \propto \mathcal{L}\left(\{Y_g\}_i|\mu_i, \sigma_i\right) p\left(\mu_i, \sigma_i\right), \tag{9}$$

which we sample using the Markov chain Monte Carlo sampler EPSIE[3] (Capano & Stiskalek 2021). Within each bin B_i we assume a uniform prior on μ_i between the minimum and maximum values of $\{Y_g\}_i$ and a uniform prior on σ_i between 0 and the difference between the maximum and minimum values of $\{Y_g\}_i$.

We run this procedure separately for each B_i and obtain samples from the two-dimensional posterior distributions. We find the one-dimensional marginalized posterior for each σ_i,

$$p\left(\sigma_i|\{Y_g\}_i\right) = \int p\left(\mu_i, \sigma_i|\{Y_g\}_i\right) \mathrm{d}\mu_i, \tag{10}$$

by projecting the posterior samples on to the σ_i axis. We quote the scatter in the X–Y relation for bin B_i as the median of $p(\sigma_i|\{Y_g\}_i)$ and its upper and lower uncertainties as the absolute difference between the median and the 84th and 16th percentiles of $p(\sigma_i|\{Y_g\}_i)$, respectively. This procedure is designed to penalise bins with few samples, since although the quoted scatter for those bins may be small, it will have a large uncertainty.

[3] www.cdcapano.github.io/epsie

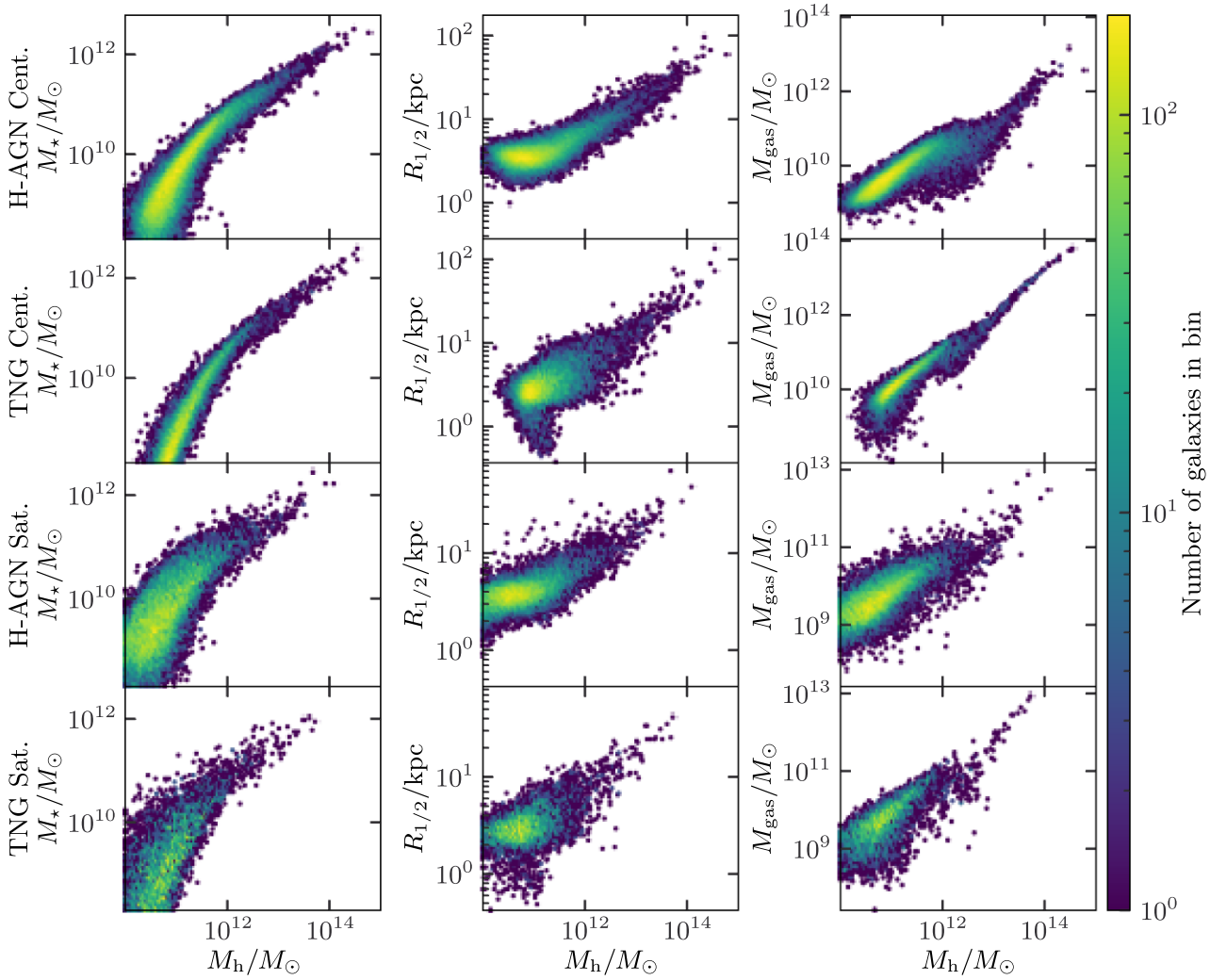

Figure 1. H-AGN and TNG correlations of $M_{\rm h}$ with $M_\star$ (left column), $R_{1/2}$ (central column), and $M_{\rm gas}$ (right column) for central and satellite galaxies separately. Rows are ordered as H-AGN centrals, TNG centrals, H-AGN satellites, and TNG satellites. The colour scale indicates logarithmic number counts.

Table 1. Summary of the different (sub)halo/galaxy properties of each simulation used in this analysis, including key differences. The definitions are discussed in detail in Section 2.2.

Property	H-AGN	TNG100
$R_{\rm h}$	Radius within which DM (sub)halo is both bound and virialised	Distance from halo centre to furthest bound DM particle
$M_{\rm h}$	Total mass of DM within $R_{\rm h}$	Total mass of DM bound to (sub)halo
$c_{\rm h}$	NFW subhalo concentration ($R_{\rm h}/r_{\rm s}$)	Same as H-AGN†
$M_{\rm vir}^P$	Virialised DM mass of subhalo's parent halo	Spherical overdensity total virial mass of subhalo's parent halo
$\lambda_{\rm h}$	Bullock spin of DM particles	Same as H-AGN†
$s_{\rm h}$, $q_{\rm h}$	Ratios of principal axes of DM mass ellipsoid	Same as H-AGN†
$M_\star$	Total mass of stars within $R_{\rm h}$	Same as H-AGN†
$M_{\rm gas}$	Total mass of gas within $R_{\rm h}$	Same as H-AGN†
$R_{1/2}$	Radius containing half the total stellar mass	Same as H-AGN†
SFR	Time-averaged star formation rate over the last 300 Myr	Time-averaged star formation rate over the last 200 Myr
Z	Mass-weighted metallicity of stars	Same as H-AGN†
$M_{\rm u,g,r,i}$	Absolute magnitude of galaxy in different bands	Same as H-AGN†
$M_{\rm HI}$	N/A	Total mass of neutral hydrogen bound to (sub)halo
$\lambda_{\rm g}$	Bullock spin of stellar particles	Same as H-AGN†
$s_{\rm g}$, $q_{\rm g}$	Ratios of principal axes of stellar mass ellipsoid	Same as H-AGN†

†Note that while H-AGN and TNG calculate properties using similar methods/definitions, the particle set that is operated on differs noticeably between the two. H-AGN uses all available particles (both gravitationally bound and unbound) within a spherical aperture cut, whereas TNG only uses bound particles and without any spherical aperture cut.

Note that the above models the scatter as being in the Y direction, which is appropriate for predicting Y given X. This is affected by the slope of the Y–X relation, so a more suitable measure of intrinsic scatter in general may be that orthogonal to the mean relation. However, since we are interested in the prediction of Y given X, we do not consider this further. We will use these results to quantify the extent to which including additional variables reduces the scatter in the galaxy–halo connection.

3.2 ML preprocessing

We split the data into into training (80 per cent) and test (20 per cent) sets using a stratified shuffle split, ensuring that the split preserves the same fraction of samples as in the original set when binned by the target property. We reserve 20 per cent of the training set for validation, which is used for either hyperparameter selection or deciding whether to continue training. The test set remains unseen during training and is only used to evaluate the performance of a trained model. All results are evaluated on the test set.

To account for the fact that most galaxy and halo properties span large dynamic ranges we work with their logarithms, except in case of magnitudes, principle axes ratios, and star metallicity. We apply the 'standard' scaling to transform the features, enforcing a mean of 0 and variance of 1. We also transform the features into an orthogonal basis (as outlined below) because most ML algorithms display marginally superior performance on uncorrelated features. All transformations are fitted on the training set.

The properties of galaxies and haloes quoted in the simulation catalogues form a mutually correlated basis. We map them to a linearly uncorrelated 'orthogonal' basis using principal component analysis (F.R.S. 1901; Hotelling 1936) such that the components maximize the variance of the data. Dimensionality reduction could then be done by projecting into a subspace containing only components with high variance, although we do not do show here because it is not necessary due to a small number of variables and because it discards information.

3.3 NN architecture

To model the galaxy–halo connection we use an ensemble of feed-forward NNs with a Gaussian loss function. A feed-forward NN defines a mapping $\mathcal{F} : \mathbb{R}^m \rightarrow \mathbb{R}^n$ with m input nodes (input layer) and n output nodes (output layer), possibly with hidden layers in between. The simplest fully connected NN with no hidden layers is a single-layer perceptron which connects the input layer m-dimensional vector x_i directly to the output n-dimensional vector y_i such that

$$y_i = f\left(\sum_{j=1}^{n} W_{ij} x_j + b_i\right), \tag{11}$$

where W_{ij} is a $n \times m$-dimensional matrix of weights and b_i is a bias vector, both of which are trainable parameters, and $f(\cdot)$ is an activation function. For a network with matrix multiplication layers (as outlined above) to learn a non-linear mapping, the activation function $f(\cdot)$ must be non-linear. Common choices of $f(\cdot)$ are the sigmoid function or hyperbolic tangent. If a network contains hidden layers then the vector y_i is used as input to the next layer. It has been shown that a NN is capable of approximating any Borel-measurable function to any desired accuracy, thus acting as a universal approximator, provided sufficient hidden nodes are used (Hornik, Stinchcombe & White 1989).

We choose a scaled exponential linear unit (SELU) activation function (Klambauer et al. 2017), defined as

$$f(x) = \lambda \begin{cases} x & \text{if } x \geq 0 \\ \alpha\left(e^x - 1\right) & \text{if } x < 0, \end{cases} \tag{12}$$

where $\lambda = 1.05$ and $\alpha = 1.67$ are chosen to induce self-normalization of the network, which preserves zero mean and unit variance between consecutive layers if the initial weights are drawn to satisfy these constraints on average. Networks with the SELU activation are robust to perturbations, can be efficiently used to train deep feed-forward networks, and avoid the problem of vanishing or diverging gradients (Klambauer et al. 2017).

The network parameters are trained by minimizing a loss function, L, which assesses the closeness of the prediction $\mathcal{F}(\boldsymbol{x})$ to $\boldsymbol{y}$. We train the network on known tuples of features and targets $\{\boldsymbol{x}, \boldsymbol{y}\}$ (*supervised learning*). They are fitted using backpropagation which computes the gradient of L with respect to the trainable parameters (Rumelhart, Hinton & Williams 1986). With the gradient available the optimal set of parameters can be found via e.g. a gradient descent algorithm.

During training, the gradient descent optimizer searches the parameter space of network weights and bias vectors until a global minimum is found. If the loss function is interpreted as the negative log-likelihood, then the minimum corresponds to the maximum-likelihood estimate. However, in practice the parameter space has a complex structure and therefore the optimizer can converge to a local minimum instead.

A common loss function used to infer the optimal networks parameters is the mean square error (MSE). MSE can be viewed as optimizing the mean network prediction, without assessing the prediction uncertainty. To account for the uncertainty Nix & Weigend (1994) suggest to use as a loss function the negative log-likelihood of a Gaussian distribution,

$$L\left(\{y\} \mid \{\boldsymbol{x}\}, \boldsymbol{\theta}\right) = \sum_k \left[\ln \sigma_{\boldsymbol{\theta}}^2(\boldsymbol{x}^k) + \frac{\left(y^k - \mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{x}^k)\right)^2}{2\sigma_{\boldsymbol{\theta}}^2(\boldsymbol{x}^k)}\right] + \text{const.}, \tag{13}$$

where $\boldsymbol{\theta}$ denotes the network's parameters, $\mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{x}^k)$ and $\sigma_{\boldsymbol{\theta}}(\boldsymbol{x}^k)$ are the network's mean prediction and standard deviation, respectively, and k indexes the tuples of features and targets. In practice, this is achieved by having the network's final layer output two values: both the mean and standard deviation of the Gaussian. We apply a so-called soft-plus transformation to the standard deviation to enforce positivity,

$$\sigma_{\boldsymbol{\theta}} \rightarrow \ln\left(1 + e^{\sigma_{\boldsymbol{\theta}}}\right). \tag{14}$$

We use an ensemble of deep NNs with the aforementioned Gaussian loss function trained with adversarial examples to obtain robust uncertainty estimates (Lakshminarayanan, Pritzel & Blundell 2016). We obtain the adversarial samples $\boldsymbol{x}'$ following the fast gradient sign method by perturbing the original samples $\boldsymbol{x}$ in the direction of the highest loss L,

$$\boldsymbol{x}' = \boldsymbol{x} + \epsilon \, \text{sign}\left(\nabla_{\boldsymbol{x}} L\left(\{y\} \mid \{\boldsymbol{x}\}, \boldsymbol{\theta}\right)\right), \tag{15}$$

where ϵ is the constant adversarial step size (Goodfellow, Shlens & Szegedy 2014). The target y of each sample remains unchanged under the adversarial attack. Ensembles of base models typically improve the predictive power (Dietterich 2000). Building an ensemble typically requires a randomization scheme to decrease the correlation between the individual base learners. However, the training of a NN is a stochastic process which already provides the necessary

randomization to decorrelate the base learners, thus avoiding the need for e.g. bootstrap aggregating as is done with Random Forests (described in Section 3.5).

For our NN architecture, we use a wide and deep neural network (WDNN) suggested in Cheng et al. (2016) and used to model the galaxy–halo connection in Moster et al. (2021). A WDNN consists of a wide part which connects the input layer directly to the output layer and a multilayer perceptron part with several hidden layers. The outputs of both parts are added. An advantage of WDNN is that it learns the linear relations through the wide part, while non-linear relation can be learnt through the deep part of the network. This improves the network's generalization to unseen samples without loss of accuracy and also improves the network's convergence.

We implement the ensemble of WDNN networks trained with adversarial examples in TENSORFLOW[4] (Abadi et al. 2015) and TENSORFLOW PROBABILITY[5] (Dillon et al. 2017). The NN hyperparameters include the number, type and size of hidden layers, activation function, choice of the gradient descent algorithm and its own hyperparameters, and training procedure. Our procedure for selecting these is as follows.

The network's input layer is normalizing: it removes the training samples' mean and rescales them to unit variance. The deep part of the base WDNN network contains $16 \times 16 \times 16 \times 16 \times 8$ hidden nodes whose weights we initialize with the LeCun normal initializer (LeCun et al. 2012) with the SELU activation function. We choose the ensemble size to be 250, the adversarial loss function's weight to be 0.5, the adversarial step size to be $\epsilon = 0.001$, and use the Adam stochastic optimizer (Kingma & Ba 2014) to find the optimal network parameters with a decaying learning schedule. We set the initial rate to 0.004 and enforce it to decay by a factor of 1.3 every 3000 iterations. During training the training set is split into batches of size 1000. These batches are used to update the networks' weights. A single iteration through all batches defines one epoch and we train the networks until the loss function evaluated on the validation set has not lowered for at least 100 epochs. The samples in the validation set are *not* used to update the network's weights. We make our NN implementation publicly available at https://github.com/Richard-Sti/scatternn.

3.4 Quantifying the neural networks' predictions

The q^{th} base network of the ensemble returns a Gaussian distribution on a target property y centred at μ_q with standard deviation σ_q. We average the individual Gaussian distributions, yielding a distribution that can be summarized with mean and variance, respectively,

$$\mu = \frac{1}{N}\sum_{q=1}^{N}\mu_q, \tag{16a}$$

$$\sigma^2 = \frac{1}{N}\sum_{q=1}^{N}\sigma_q^2 + \left(\mu_q - \mu\right)^2, \tag{16b}$$

where N is the number of base networks. We consider the mean of this distribution as the 'best-fitting prediction' and standard deviation as the scatter on the predicted halo/galaxy property.

The ensemble of NNs models the scatter in galaxy (halo) quantities conditioned on the halo (galaxy) quantities it has been trained on, which may be of any number. In Section 3.1 we described our prescription for calculating the bivariate simulation scatter, and its uncertainty, between two properties. This scatter then serves as an upper limit if a third property is added to the feature set, potentially allowing for a more precise prediction.

We compare the bivariate scatter to the NNs' scatter. If the NNs are aware of only the property on which the bivariate scatter is conditioned, then we we expect the two to be approximately equal as both are fitting a Gaussian conditioned on the same quantity. However, if the networks learn from other informative properties as well, then the networks' predictions will be not only more precise, i.e. with lower scatter, but this scatter will also be less uncertain.

3.5 Decision tree ensemble

To check the results of our network against a completely different architecture, we also consider non-parametric regression by means of a decision tree ensemble. A decision tree is built from the 'root' node, which contains the entire training set, by recursively splitting the training set into more homogeneous 'child' nodes. This is achieved by minimizing some loss function L at each split. Typically, in regression the split is chosen such that it minimizes the weighted variance of the samples assigned to the left and right child nodes. If the ith sample contains several features then the split is chosen along the one that minimizes L. The splitting is done recursively until a termination condition is met, e.g. the node's variance is zero, it contains only a single sample or a maximum tree depth is reached. Prediction on unseen samples is performed by propagating an input feature vector to a terminal leaf and, in regression, returning the mean value of all training sample targets assigned to the particular terminal leaf.

If the maximum depth of a decision tree is not capped it can 'memorize' the entire training set. Because of this over-fitting such unregulated decision trees do not generalize well to unseen samples. To account for this the tree can either be 'pruned' by removing branches with only a small increase in homogeneity, or randomized. A popular algorithm is the Random Forest (RF) which builds an ensemble of decision trees (Breiman 2001). RF injects randomness into the ensemble by training the trees on bootstrap aggregated samples and randomly selecting the number of features that are considered for an optimal split. The RF's prediction is the average of the values returned by its constituent trees.

Another ensemble algorithm choice is the Extremely Randomised Trees (Extra-Trees; Geurts, Ernst & Wehenkel 2006). Unlike RF, Extra-Trees trees are all grown on the same original sample and randomness is injected in the splitting. After randomly selecting k out of m features, each feature's threshold is *randomly* selected, instead of calculating the optimal threshold, with the best threshold being used to split the node. By picking the thresholds at random the computational cost of training is significantly reduced, while also reducing the model's variance at the cost of a small increase in bias. Moreover, in most applications – including ours – Extra-Trees is as good a predictor as RF, and we will therefore use it as an alternative to our NN model. While Extra-Trees and the other aforementioned tree-based models yield purely deterministic predictions, there exist decision tree models that model the statistical uncertainty (e.g. Chipman, George & McCulloch 2010; Duan et al. 2019). However, we do not consider these here.

We use the SCIKIT-LEARN[6] (Pedregosa et al. 2011) implementation of Extra-Trees EXTRATREESREGRESSOR. We find the optimal hyperparameters via a grid search, choosing from N_ESTIMATORS =

[4]https://www.tensorflow.org
[5]https://www.tensorflow.org/probability/
[6]https://scikit-learn.org/

64, 128, MIN_SAMPLES_SPLIT = 8, 16, 32, MAX_SAMPLES = 0.5, 1 and keeping the default values for the remaining hyperparameters.

3.6 Feature selection and importance

To assess the goodness-of-fit of the mean NN ensemble or Extra-Trees predictions on the test set, we use the R^2 coefficient of determination, defined as

$$R^2 = 1 - \frac{\sum_i (y_i - \mathcal{F}(\boldsymbol{x}_i))^2}{\sum_i (y_i - \hat{y})^2}, \tag{17}$$

where the sums run over the (test set) samples, $\mathcal{F}(\boldsymbol{x}_i)$ is the predicted value, and $\hat{y}$ is the mean target value on the test set. R^2 is upper bounded by 1, which signifies perfect predictions (zero residuals). Lower values indicate worse predictions. If $R^2 = 0$ the model only learns to predict the mean value $\hat{y}$ and $R^2 < 0$ signifies a non-informative model whose predictions are imbalanced with respect to the true values on average.

We select important features for predicting a given target by first considering an original set of N features. Then, we train Extra-Trees on all N features individually and select the one that yields the highest R^2 on the test set. We keep this feature and, one at a time, add additional properties from the remaining features. We keep whichever new property yields the maximum R^2 on the test set when used in conjunction with the features already included. Repeating this process, we end up with an ordered list of features and their R^2 increments. We consider as informative all features whose R^2 increment is more than 0.005.

We use the Spearman's rank correlation coefficient ρ_s to determine whether the residuals are correlated with any other property that was not used in training. The Spearman coefficient is defined as the Pearson correlation coefficient between rank variables of two vectors, which are formed by replacing the vector's values by their increasing ranks. The Pearson coefficient of two vectors is given by their covariance divided by the product of their standard deviations. The Spearman coefficient is bounded to $[-1, 1]$, such that $+1$ signifies a correlated monotonic relation and -1 an anticorrelated monotonic relation.

Having found an ordered list of features that maximizes the R^2 increments as described above, we calculate the Spearman correlation coefficient between residuals of a model trained on the first n features and the next feature to be added. This will allow us to further interpret the significance of each added feature.

4 CORRELATIONS OF THE SIMULATED GALAXY–HALO CONNECTIONS

4.1 Halo to galaxy mapping

We use the $z = 0$ halo properties as features and for simplicity only the following galaxy properties as targets: stellar mass $M_\star$, total gas mass $M_{\rm gas}$, and stellar-half mass radius $R_{1/2}$ (see Table 1). Furthermore, in TNG we will also consider learning the mapping from halo properties to the neutral hydrogen mass $M_{\rm H\,I}$.

For each galaxy target, we begin by training Extra-Trees to determine the important halo features separately in H-AGN and TNG central and satellite galaxies, as described in Section 3.6. Since the scatter in the galaxy–halo connection is not crucial for this part, we opt for Extra-Trees rather than the NN ensemble as it provides similar goodness-of-fit of the best-fitting predictions while being computationally cheaper. Having determined the important features, we train the NN ensemble, first on the most important feature only, which we find to be $M_{\rm h}$ in all cases, and then we add the remaining important features for each target. We calculate the R^2 scores of the mean predictions on the test set, which we then compare to the equivalent inference using Extra-Trees, finding that the two typically agree to within 1 per cent for well-predicted properties and to within $\sim$10 per cent if $R^2 \lessapprox 0.5$. Where there is disagreement, the NN may outperform Extra-Trees. This gives us confidence to employ the NN ensemble to measure the scatter of the relations. We compare the NN ensemble scatter to the bivariate scatter around the mean in the relation $\log Y|\log M_{\rm h}$, where Y represents one of the galaxy target properties. We illustrate the result by plotting mean and standard deviation of the predicted scatter of test set samples binned by $\log M_{\rm h}$.

We now proceed to describe our results for each separate galaxy target that we consider.

4.1.1 Predicting the stellar mass

We begin by predicting the stellar mass $M_\star$. Using our feature selection routine outlined in Section 3.6 with Extra-Trees we find the maximum possible R^2 to be 0.92/0.95 (0.91/0.90) in H-AGN/TNG centrals (satellites), which are calculated using all informative halo properties that yield $\Delta R^2 > 0.005$. The left column of Fig. 2 shows how the R^2 depends on the cumulative set of training features. We now describe the informative features in decreasing order of importance.

In H-AGN central galaxies the only important feature beyond $M_{\rm h}$ is $c_{\rm h}$. As indicated by the 'plus' marker symbol in Fig. 2, the residuals of a model trained on $M_{\rm h}$ are positively correlated with $c_{\rm h}$. In TNG centrals the secondary important features are $\lambda_{\rm h}$, $R_{\rm h}$, and $c_{\rm h}$. We note that after including $\lambda_{\rm h}$ and $R_{\rm h}$ the increment in R^2 due to $c_{\rm h}$ is larger than the previous two increments. So $c_{\rm h}$ provides noticeable gains in constraining $M_\star$, but only once we have provided information on both $\lambda_{\rm h}$ and $R_{\rm h}$; it is not as useful a feature if $M_{\rm h}$ is the only other input provided to the model. We have not explored the detailed cause for this behaviour, but note that it appears only in the TNG data set and not in H-AGN.

In TNG $M_{\rm h}$ by itself is a better predictor of $M_\star$ compared to H-AGN, corresponding to the tighter $M_\star$–$M_{\rm h}$ relation shown in Fig. 1. For satellite galaxies, in H-AGN the important features beyond $M_{\rm h}$ are $R_{\rm h}$ and $c_{\rm h}$. Similarly, in TNG satellites the other relevant features are $R_{\rm h}$, $c_{\rm h}$, $M^P_{\rm vir}$, and $\lambda_{\rm h}$, the importances of which are similarly shown in Fig. 2. In both simulations we find that after accounting for the $M_{\rm h}$ dependence the halo size $R_{\rm h}$ is anticorrelated with $M_\star$ ('circle' marker symbol).

We now move to a study of the scatter in the galaxy–halo correlations. We indicate the bivariate scatter $\sigma_{\log M_\star|\log M_{\rm h}}$ as the black dashed line in Fig. 3 for both H-AGN and and TNG. The simulations agree that above $M_{\rm h} \approx 10^{12}\,{\rm M}_\odot$ the scatter is $\sim$0.15 and $\sim$0.2 dex for centrals and satellites, respectively. Below this mass the scatter starts growing until $M_{\rm h} \approx 10^{11}\,{\rm M}_\odot$, where it artificially peaks due to the lower limit on $M_\star$ that we impose in both simulations. In H-AGN centrals the scatter peaks $\sim$0.05 dex higher than in TNG centrals, reflecting the tighter $M_\star$–$M_{\rm vir}$ relation shown in Fig. 1. The rise in vertical scatter is largely a consequence of the steepening of the $M_\star$–$M_{\rm h}$ relation at the faint end (see Fig. 1).

We train the NN ensemble model on the features found above to be important. In all cases, we verify that the best-fitting predictions and scatter are consistent with the true values on the test set by inspecting the fraction of objects with true target properties lying within the scatter interval. We plot the NN predicted scatter as the blue and orange bands in Fig. 3, again binned by halo mass for direct comparison with the bivariate scatter. In all cases, training

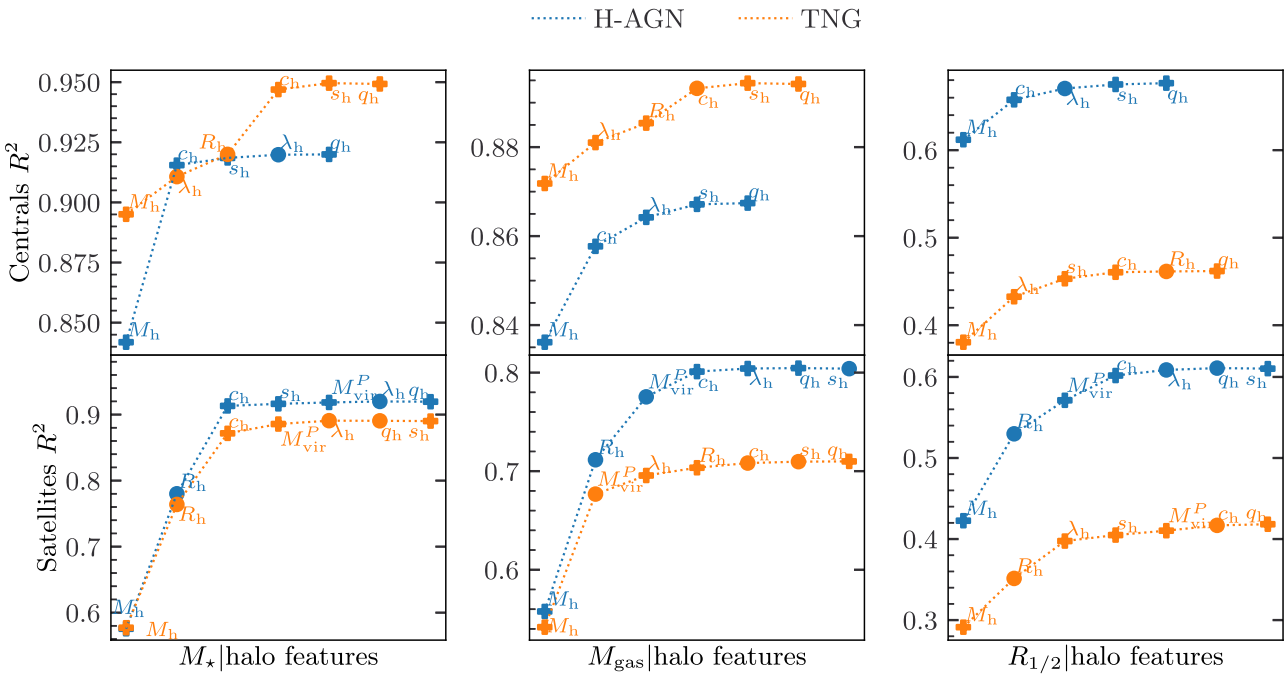

Figure 2. Test set R^2 scores from Extra-Trees as a function of the cumulative number of features used to predict $M_\star$ (*left* column), $M_{\rm gas}$ (*central* column) and $R_{1/2}$ (*right* column) in H-AGN and TNG central (*top* row) and satellite (*bottom* row) galaxies. The results from the NN ensemble are the same to within 1 per cent. Features are added in the order that maximizes the R^2 increments as described in Section 3.6. The labels show the halo feature being added to the set of features to the left of it. We indicate whether the residuals of a model trained on the previous set of features have a positive or negative Spearman correlation coefficient with the added feature by marking the point with a 'plus' or 'circle', respectively.

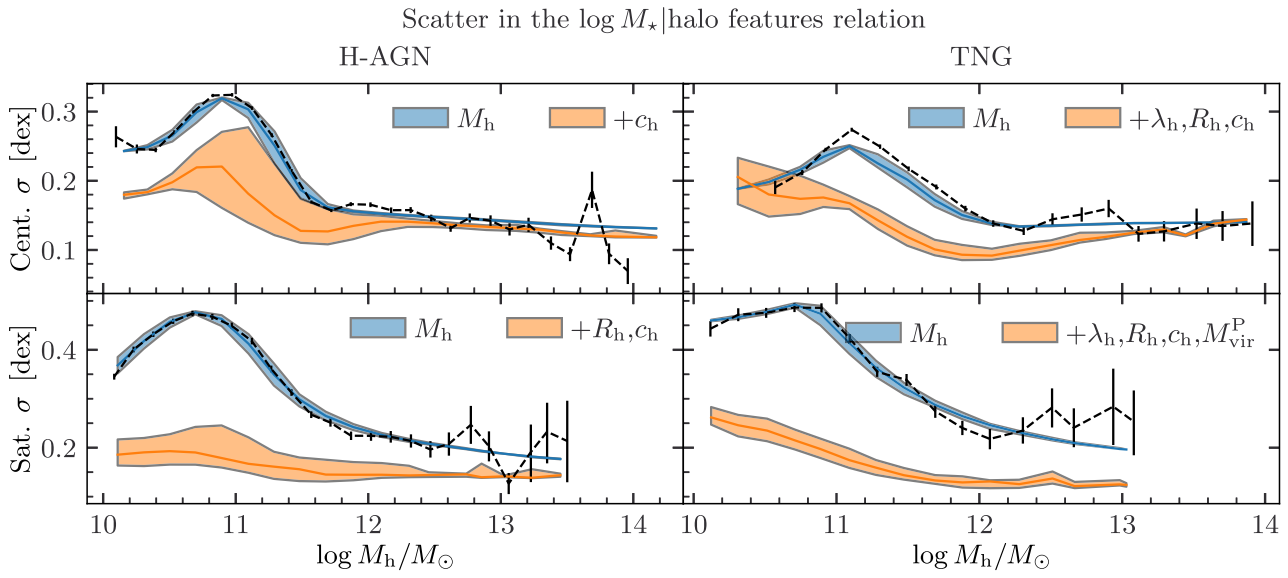

Figure 3. The NNs' predicted scatter in log $M_\star$ binned by log $M_{\rm h}$ in H-AGN (*left* column) and TNG (*right* column) for central (*top* row) and satellite (*bottom* row) galaxies. The legend indicates cumulative halo features used to train the networks, with the bands enclosing the 16th to 84th percentile of predictions for the scatter for galaxies in a given bin of log $M_{\rm h}$. The black dashed line is the $\sigma_{\log M_\star|\log M_{\rm h}}$ bivariate scatter calculated directly from the simulated points, with a 1σ error bar obtain by fitting Gaussian distributions within fixed-size log $M_{\rm h}$ bins as described in Section 3.1. The networks correctly recover the bivariate scatter if trained on $M_{\rm h}$ only (blue band). The orange band includes all other features deemed informative for the prediction of $M_\star$, resulting in a significantly lower uncertainty in $M_\star$ and indicating the extent to which these features correlate with residuals of the $M_\star$–$M_{\rm h}$ relation.

the ensemble on $M_{\rm h}$ alone correctly recovers the scatter in the simulations' log $M_\star|\log M_{\rm h}$ relations, showing that the ensemble is able to capture the non-linear $M_\star$–$M_{\rm h}$ relation.

In H-AGN centrals the addition of $c_{\rm h}$ reduces the scatter below $M_{\rm h} \approx 10^{12}$ M$_\odot$, whereas in TNG centrals the correlation of $M_\star$–$M_{\rm h}$ residuals with $\lambda_{\rm h}$, $R_{\rm h}$ and $c_{\rm h}$ explains up to 0.05 dex of the scatter across the considered halo mass range. More so in H-AGN satellites, $R_{\rm h}$ and $c_{\rm h}$ reduce scatter across all halo masses, particularly at lower halo masses (up to 0.2 dex), with similar behaviour observed in TNG. Therefore, even though the bivariate scatter $\sigma_{\log M_\star|\log M_{\rm h}}$ in satellites

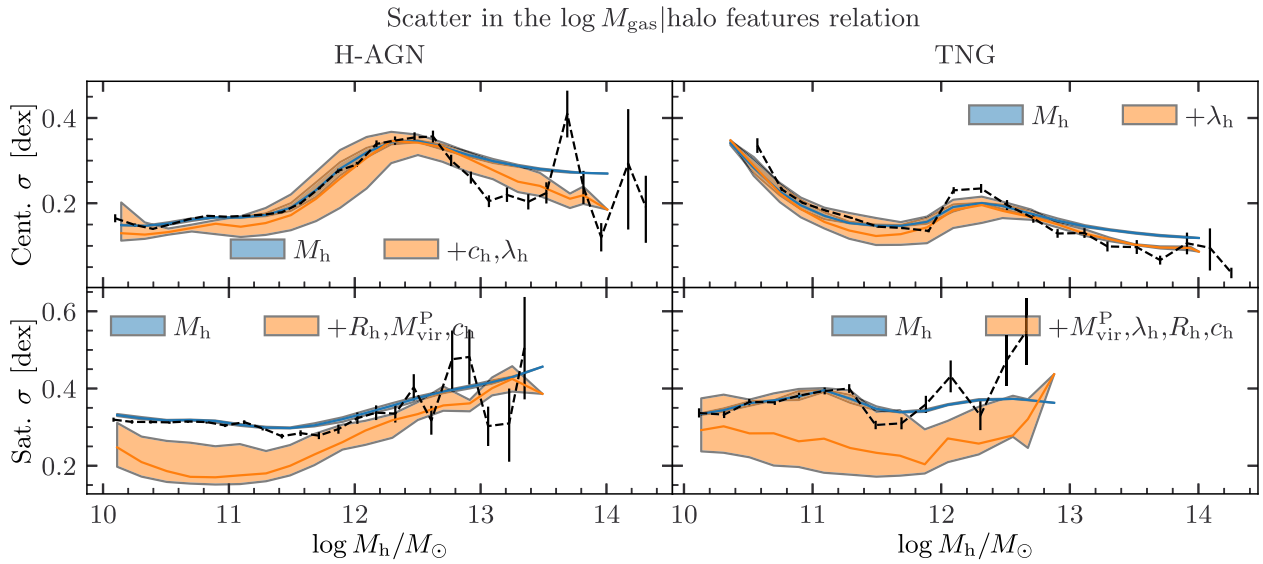

Figure 4. NNs' predicted scatter in log $M_{\rm gas}$ binned by log $M_{\rm h}$ in H-AGN (*left* column) and TNG (*right* column) for central (*top* row) and satellite galaxies (*bottom* row). The figure uses the same notation as Fig. 3. In central galaxies the addition of halo properties beyond $M_{\rm h}$ yields no substantial reduction of scatter, while in satellite galaxies the scatter is reduced only marginally.

is larger than in centrals, this can be attributed to residual correlations with secondary halo properties such that the scatter after taking these into account is similar.

4.1.2 *Predicting the gas mass*

We now move on to predicting the total gas mass, $M_{\rm gas}$. Using all informative halo features we find R^2 of 0.86/0.89 (0.79/0.70) in H-AGN/TNG centrals (satellites). In the central panel of Fig. 2 we show the R^2 score as a function of the cumulative feature set that maximizes the R^2 increments. As with $M_\star$ (and throughout), the most important halo feature is $M_{\rm h}$.

In H-AGN central galaxies the two important features for predicting $M_{\rm gas}$ beyond halo mass are $c_{\rm h}$ and $\lambda_{\rm h}$, whereas in TNG centrals the only other relevant feature is $\lambda_{\rm h}$. In both cases the importances of beyond-$M_{\rm h}$ properties are however relatively low ($\Delta R^2 \lesssim 0.02$). For satellites, we find that in H-AGN the other important features are $R_{\rm h}$, $M^{\rm P}_{\rm vir}$, and $c_{\rm h}$, while in TNG these are $M^{\rm P}_{\rm vir}$, $\lambda_{\rm h}$, $R_{\rm h}$, and $c_{\rm h}$. In H-AGN, $R_{\rm h}$ is the second most important feature ($\Delta R^2 \approx 0.15$), whereas in TNG its importance is near-negligible after accounting for $M^{\rm P}_{\rm vir}$ and $\lambda_{\rm h}$. As we describe in Section 2.2.1, in H-AGN $R_{\rm h}$ is the radius of the outermost shell within which the virial theorem remains satisfied. As a consequence, for satellite galaxies $R_{\rm h}$ is not uniquely determined by $M_{\rm h}$. On the other hand, in TNG we define $R_{\rm h}$ to be distance to the furthest gravitationally bound particle and, thus, it is not uniquely determined by $M_{\rm h}$ in either parent haloes or subhaloes. This makes it potentially informative even for a model that has already been trained on $M_{\rm h}$.

We calculate the bivariate scatter in the $\sigma_{\log M_{\rm gas}|\log M_{\rm h}}$ relation and plot it in Fig. 4 as the black dashed line. Focusing on H-AGN central galaxies first, we find a peak in scatter of $\sim$0.35 dex at $M_{\rm h} \approx 10^{12.5}$ $M_\odot$. In TNG central galaxies the scatter similarly peaks at this halo mass, though its scatter is lower by up to 0.1 dex. At this mass the contributions of the hot diffuse and star forming together with the cold gas are approximately equal and, since both trace $M_{\rm h}$ with a different slope and scatter, the total scatter is inflated. Below this mass the hot diffuse gas component is typically subdominant. The location of the peak is consistent with the halo mass at which AGN feedback first dominates galaxy evolution; Anbajagane et al. (2022a, see fig. 7) confirm this for the TNG galaxy formation model by extracting the energetics of different baryonic processes (gas cooling, star formation, and AGN feedback) as a function of halo mass. The presence of AGN feedback is known to stop gas cooling within the halo and also expel some of the gas from the halo potential well. In TNG centrals we also note a large increase in scatter at low halo masses due to outlier simulated galaxies with artificially low star forming gas, likely due to the simulation's finite resolution. There is no prominent peak in either H-AGN nor TNG satellites where the scatter is approximately constant at $\sim$0.35 dex in both cases.

The coloured bands in Fig. 4 show the NN ensemble scatter in log $M_{\rm gas}$ binned by halo mass. In both H-AGN and TNG we first train the NN ensemble on $M_{\rm h}$ only, which correctly recovers the bivariate scatter $\sigma_{\log M_{\rm gas}|\log M_{\rm h}}$. In centrals, adding the other informative properties yields no significant decrease in either H-AGN nor TNG, showing that the scatter in $\sigma_{\log M_{\rm gas}|\log M_{\rm h}}$ is not significantly attributable to residual correlations with other halo features. On the other hand, in both H-AGN and TNG satellites the addition of secondary properties can reduce the scatter by up to 0.1 dex.

4.1.3 *Predicting the neutral hydrogen mass*

In TNG we also try predicting the neutral hydrogen mass $M_{\rm H\,I}$. With Extra-Trees we find maximum R^2 values of 0.47 and 0.49 in central and satellite galaxies, respectively, using all informative dark matter features. We show the R^2 score as a function of the cumulative feature set in Fig. 5. In central galaxies the important halo properties besides $M_{\rm h}$ are $\lambda_{\rm h}$, $s_{\rm h}$, and $c_{\rm h}$, though it is mainly $\lambda_{\rm h}$ that improves the predictions ($\Delta R^2 \approx 0.075$). In satellite galaxies, the relevant secondary features are $M^{\rm P}_{\rm vir}$, $\lambda_{\rm h}$, and $c_{\rm h}$. While for central galaxies $\lambda_{\rm h}$ is the second most important feature and is positively correlated with $M_{\rm H\,I}$ after accounting for the $M_{\rm h}$ dependence, in satellites it is the

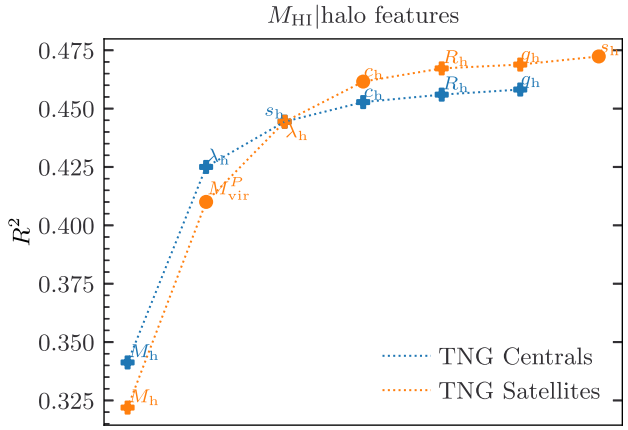

Figure 5. Test set R^2 scores from Extra-Trees as a function of the cumulative number of used features for predicting $M_{\rm HI}$ in TNG central and satellite galaxies. Notation follows that of Fig. 2. Note that the prediction is poor ($R^2 \ll 1$) no matter how many halo features are used.

third most important feature and is negatively correlated with $M_{\rm HI}$ after accounting for both $M_{\rm h}$ and $M^{\rm P}_{\rm vir}$.

The scatter in the relation $\log M_{\rm HI}|\log M_{\rm h}$ is approximately constant at 0.40 dex (0.45 dex) in centrals (satellites), which is typically at least twice as large as the scatter in stellar mass. Because of the low R^2 and relatively low importance of features beyond $M_{\rm h}$, training the NN ensemble on such features can reduce the $M_{\rm HI}$–$M_{\rm h}$ scatter by at most 0.05 dex.

4.1.4 *Predicting the stellar half-mass radius*

Predicting the stellar half-mass radius $R_{1/2}$ from halo properties, using Extra-Trees we arrive at a maximum R^2 of 0.67/0.47 (0.60/0.41) in H-AGN/TNG centrals (satellites) when using all informative halo features ($\Delta R^2 > 0.005$). The right column in Fig. 2 shows the R^2 scores as a function of the cumulative feature set that maximizes the R^2 increments.

As we show in Fig. 2, in H-AGN centrals the other important halo properties for predicting $R_{1/2}$ are $c_{\rm h}$ and $\lambda_{\rm h}$. In TNG centrals these are $\lambda_{\rm h}$, $s_{\rm h}$ and $c_{\rm h}$. However, in both simulations $R_{1/2}$ is poorly constrained and the addition of secondary features results in $\Delta R^2 \lesssim 0.05$. Looking at satellites, in H-AGN we see that the secondary important features are $R_{\rm h}$, $M^{\rm P}_{\rm vir}$, $c_{\rm h}$, and $\lambda_{\rm h}$, whereas in TNG these are $R_{\rm h}$, $\lambda_{\rm h}$, $s_{\rm h}$, and $M^{\rm P}_{\rm vir}$. In both cases the satellite $R_{1/2}$ is predicted worse than in central galaxies and $R_{\rm h}$ is the second most important feature, which is anticorrelated with $R_{1/2}$ after accounting for $M_{\rm h}$. These correlations are discussed further in the context of semi-empirical galaxy formation models in Section 5.

The distribution $R_{1/2}$–$M_{\rm h}$ differs in H-AGN and TNG (see Fig. 1) and, as a consequence, so does the scatter $\sigma_{\log R_{1/2}|\log M_{\rm h}}$ which we show in Fig. 6. In H-AGN the scatter is constant at ∼0.1 dex in both centrals and satellites, though the scatter in satellites is higher by ∼10 per cent. On the other hand, in TNG the distribution is wider, with the scatter peaking at ∼0.18 dex at $M_{\rm h} \approx 10^{12}\,{\rm M}_\odot$. This peak can be attributed to the fact that above it the galaxy stellar size is proportional to the halo virial mass with a relatively small scatter, whereas below it the relation's slope flattens and widens. At the peak's location the distribution of galaxy size at fixed virial mass has a heavy tail, deviating from a Gaussian distribution. This peak is also consistent with the AGN feedback discussed above.

We train the NN ensemble on the previously identified important features and show the predicted scatter in $\log R_{1/2}$ in Fig. 6. In central galaxies calculating the NN scatter on all important halo properties yields no significant improvement over the bivariate scatter $\sigma_{\log R_{1/2}|\log M_{\rm h}}$. In H-AGN satellites the network learns to attribute ∼10 per cent of the scatter to secondary halo properties; however, in TNG satellites the inclusion of such features does not reduce the scatter.

4.2 Galaxy to halo mapping

Our machinery may easily be adapted to learn the inverse of the galaxy–halo connection as we have considered it above, i.e. the mapping from galaxy to halo properties. We consider two sets of galaxy properties as features. The first set includes all galaxy properties summarized in Table 1 except absolute magnitudes, while the second consists only of absolute magnitudes as the directly observable properties. With the first set we predict both the halo mass $M_{\rm h}$ and spin $\lambda_{\rm h}$, whereas with the second we predict only $M_{\rm h}$. We briefly discuss predicting the halo concentration $c_{\rm h}$ in Section 6.

As in Section 4.1, we first train an Extra-Trees algorithm to determine the important galaxy features and then the NN ensemble to learn the scatter in the relations. We compare this to the bivariate scatter in $\log M_{\rm h}|\log M_\star$. We condition this on $\log M_\star$ as we expect it *a priori* to be the most important property.

4.2.1 *Predicting the halo mass*

Training Extra-Trees on the galaxy properties to predict the halo mass $M_{\rm h}$ yields a maximum R^2 of 0.95/0.98 (0.80/0.95) in H-AGN/TNG centrals (satellites) when all informative features are considered. In Fig. 7 we show the R^2 scores as a function of the cumulative feature set.

As we show in Fig. 7, in H-AGN centrals the important galaxy properties beyond $M_\star$ are Z and $M_{\rm gas}$. On the other hand, in TNG the most important property that maximizes R^2 on its own is $M_{\rm gas}$ and the other relevant features are $M_\star$ and $\lambda_{\rm g}$. Going to satellites, in H-AGN the most important galaxy feature is $M_\star$ and the other informative properties, ordered by decreasing importance, are Z, $M_{\rm gas}$, SFR, $q_{\rm g}$, and $\lambda_{\rm g}$. In TNG satellites, the most important feature is again $M_{\rm gas}$, with the other relevant features being $M_\star$, $\lambda_{\rm g}$, $R_{1/2}$, and $s_{\rm g}$. We see that in H-AGN central and satellite galaxies stellar metallicity is the second most important feature and is anticorrelated with $M_{\rm h}$ after removing the $M_\star$ dependence, and that in all cases it is beneficial to consider both $M_\star$ and $M_{\rm gas}$. This is in contrast to the halo-to-galaxy prediction where a single feature, $M_{\rm h}$, was unambiguously the most important.

We show the bivariate scatter $\sigma_{\log M_{\rm h}|\log M_\star}$ in Fig. 8. In H-AGN the scatter in centrals is ∼0.2 dex at both low and high $M_\star$, although in between the scatter decreases and at $M_\star \approx 10^{10.3}\,{\rm M}_\odot$ it reaches ∼0.15 dex. In TNG the scatter in centrals behaves similarly, though it is consistently lower than in H-AGN, particularly at lower stellar masses. In satellites the scatter at fixed stellar mass grows nearly monotonically from ∼0.2 dex at low $M_\star$ to ∼0.4 dex at high $M_\star$ in both H-AGN and TNG. The increase in scatter from low to high $M_\star$ can be understood from the simulated $M_\star$–$M_{\rm vir}$ relations (see Fig. 1), which flatten at high $M_\star$. Similarly, in centrals it reaches a minimum at $M_\star \approx 10^{10.3}\,{\rm M}_\odot$ where the two parts of the broken power law meet.

The bands in Fig. 8 show the predicted NN ensemble scatter in $\log M_{\rm h}$, binned by $M_\star$. In all cases, the networks visibly learn from properties beyond $M_\star$, since in TNG $M_{\rm gas}$ by itself is marginally even more important than $M_\star$. In both H-AGN and TNG central galaxies the addition of the informative features beyond $M_\star$ nearly halves the predicted scatter. In satellites the addition of the important secondary

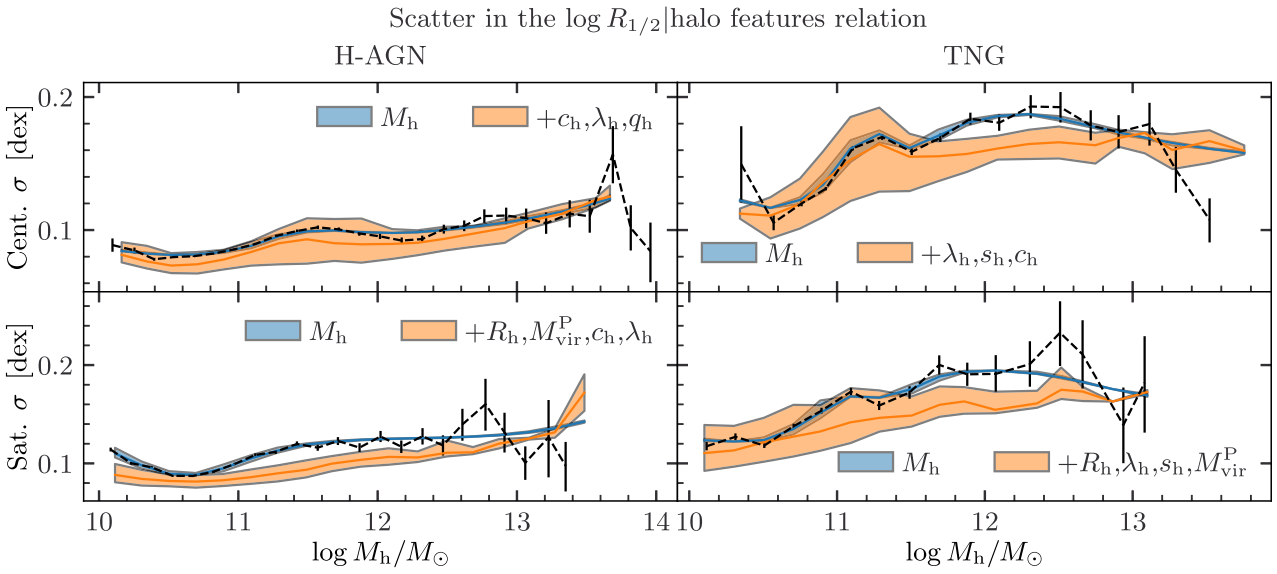

Figure 6. NNs' predicted scatter in $\log R_{1/2}$ binned by $\log M_h$ in H-AGN (*left column*) and TNG (*right column*) for central (*top row*) and satellite galaxies (*bottom row*). The figure uses the same notation as Fig. 3. Similarly as for M_{gas}, the addition of properties beyond M_h only marginally reduces the uncertainty in the $R_{1/2}$ prediction for satellite galaxies.

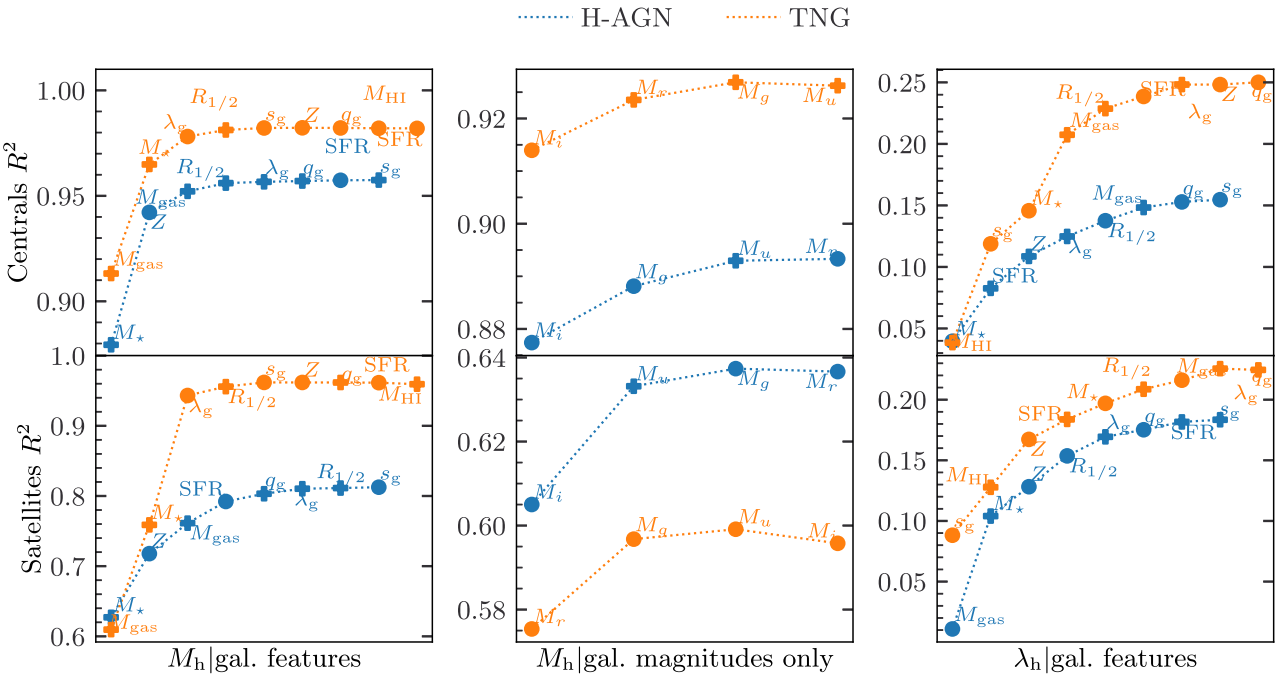

Figure 7. Test set R^2 scores from Extra-Trees as a function of the cumulative number of features used to predict M_h from all simulation features and from magnitudes only (*left* and *central* columns, respectively), and to predict λ_h (*right* column), for central (*top* row) and satellite (*bottom* row) galaxies. The notation follows Fig. 2. The magnitudes-only predictions of M_h are substantially worse in satellite galaxies and in all cases we find λ_h to be very poorly predicted ($R^2 \lesssim 0.25$).

features flattens the increasing scatter at the bright end, partially in H-AGN and completely in TNG.

4.2.2 *Predicting halo mass from photometric observables*

Training Extra-Trees on the *u*, *g*, *r*, and *i* magnitude bands to predict M_h, we find maximum R^2 values of 0.89/0.92 (0.64/0.60) in H-AGN/TNG central (satellite) galaxies. We show the R^2 as a function of the cumulative features used in Fig. 7. In all cases we find the *i* and *r* magnitude bands to maximize the R^2 increment, although the magnitudes are strongly mutually correlated. The *i*-band is the most important because it exhibits the lowest scatter in its mass-to-light ratio: in both simulations we observe that $M_\star$ correlates near-perfectly with M_i (up to $|\rho_s| \approx 0.99$) and weakest with the *u*-band magnitudes ($|\rho_s| \lesssim 0.90$).

From the right column of Fig. 7 we see again that in H-AGN centrals the best band for predicting halo mass is *i*, with some addi-

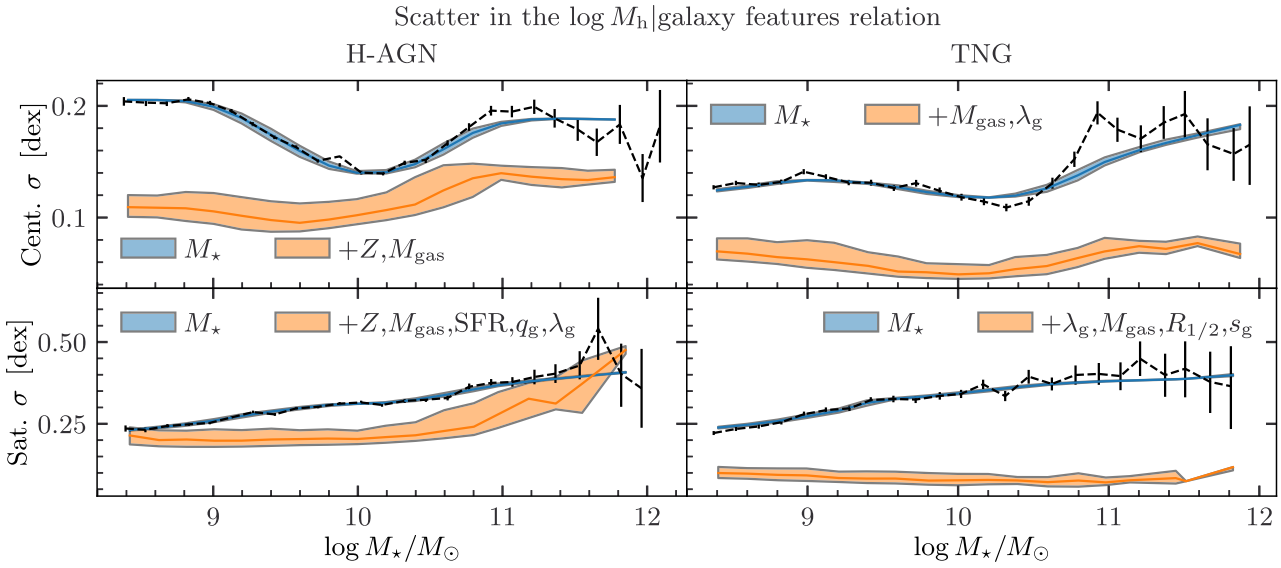

Figure 8. NNs' predicted scatter in log $M_{\rm h}$ binned by log $M_\star$ in H-AGN (*left* column) and TNG (*right* column) for central (*top* row) and satellite (*bottom* row) galaxies, training on galaxy properties. The figure's notation follows that of Fig. 3. In all cases galaxy properties other than $M_\star$ are substantially informative for predicting $M_{\rm h}$.

tional information from the *g* and *u*-bands. The *r*-band is considered unimportant because of its strong correlation with particularly the *i*-band. In TNG the most important band is similarly *i*, with only minor contribution from *r*. Moving to H-AGN satellites, we see that the most important magnitude is still M_i with some minor importance again in M_u and and M_g. In TNG satellites the most important band is now *r*, with additional minor information in *g* and *u*. In all cases we find that the *i* or *r*-band magnitudes are the most important. In both simulations the direct observables capture nearly all the $M_{\rm h}$ information of central galaxies, while in satellites the R^2 is significantly lower.

In neither case does training the NN ensemble on more than the most important magnitude band show significant benefit in reducing the scatter in halo mass. For central galaxies the dependence of scatter on the magnitude is similar to that on stellar mass, which we show in Fig. 8. On the other hand, in satellite galaxies the directly observable properties capture substantially less information than the full set ($\Delta R^2 \approx 0.3$), which is in part reflected by the fact that the scatter grows near linearly towards brighter galaxies even if all magnitudes are included in the training set. The main halo mass can be predicted from magnitudes with an uncertainty of ∼0.15 dex, whereas the subhalo mass scatter runs from ∼0.2 dex at the faint end to ∼0.4 dex at the bright end.

4.2.3 Predicting the halo spin

Finally, we ask whether the halo spin can be inferred from galactic properties only. We find that if we consider all informative galaxy properties the maximum R^2 is 0.15/0.25 (0.18/0.23) in H-AGN/TNG central (satellite) galaxies, suggesting that in all cases $\lambda_{\rm h}$ is very poorly predicted. We show the R^2 as a function of the cumulative feature set in the right-hand panel of Fig. 7.

Starting with central galaxies, the most important properties are $M_\star$ and $M_{\rm H\,I}$ in H-AGN and TNG, respectively, although their R^2 are close to 0, meaning that having access to only them the model at best learns to predict the average spin. While in H-AGN centrals no additional property yields a significant R^2 increase on its own, in TNG the addition of $s_{\rm g}$ after $M_{\rm H\,I}$ increases the R^2 by ≈0.08. Similar behaviour is observed with satellites, although in the case of TNG we find $s_{\rm g}$ to be the best predictor on its own with $R^2 \approx 0.1$. Because of the low R^2 score even when training on all galaxy properties we do not pursue modelling the scatter in the galaxy–halo spin connection.

5 IMPLICATIONS FOR SEMI-EMPIRICAL MODELS

Semi-empirical and semi-analytic models of galaxy formation stipulate correlations between galaxy and halo properties which may or may not be present in the full-physics results of hydrodynamical simulations. Provided they accurately describe the real Universe, such simulations therefore provide an opportunity to test the models. Our ML analysis contributes to this by uncovering the most significant correlations of the simulations' galaxy–halo connections, while more detailed information can be obtained by testing the models explicitly on the simulated data. Here we focus on two classes of semi-empirical model: (1) subhalo abundance matching (SHAM) as a method for associating luminosity or stellar mass to haloes given their mass and mass accretion history, and (2) models that attempt to predict galaxy size from halo properties, in particular the AM partition model of Fall & Efstathiou (1980), Mo et al. (1998), and the galaxy size–virial radius relation of Kravtsov (2013).

5.1 Subhalo abundance matching

SHAM posits a nearly monotonic relation between the luminosity or stellar mass of a galaxy and a halo quantity called the 'proxy' which is typically a mass or rotation velocity at a particular epoch of the halo's formation. While the earliest models used present-day virial mass as the proxy (Conroy, Wechsler & Kravtsov 2006; Behroozi et al. 2010; Moster et al. 2010), more recent work has shown that fits to galaxy clustering are improved by instead using a quantity that scales with haloes' concentration (Reddick et al. 2013; Lehmann et al. 2017); this reflects, through the correlation of concentration and halo formation time, the fact that galaxies formed in haloes at $z > 0$. SHAM also provides reasonable fits to galaxies' dynamical scaling relations, although there are indications that the galaxy mass–halo mass relation it implies is too curved and that

the neglect of additional correlations in the galaxy–halo connection generates excess scatter in circular velocity at fixed mass (Trujillo-Gomez et al. 2011; Desmond & Wechsler 2015, 2017; Desmond 2017a, b; Li et al. 2022).

In Fig. 2 we show the halo features that galaxy stellar mass most strongly depends on in the H-AGN and TNG simulations. While the halo mass is clearly the strongest predictor, we find that for satellites $R_{\rm h}$ is a more important feature than concentration, indicating that SHAM may be able to provide an improved fit to the galaxy–halo connections in the simulations by incorporating this variable. The fact that $M_{\rm H\,I}$ is relatively poorly predicted even when all our halo properties are considered (Fig. 5) explains why SHAM models have found little success in modelling the neutral hydrogen gas of galaxies (Guo et al. 2017; Dutta, Khandai & Rana 2021; Stiskalek et al. 2021). The models require a large scatter in the SHAM matching to replicate the poor correlation of neutral hydrogen galaxy mass with halo properties. Our results indicate that this situation could not be greatly improved no matter which additional halo variables considered in this work were included in the SHAM proxy, and also that the optimal model likely depends much more strongly on simulation-specific physics than does one based on $M_\star$. Of course we calculate the correlations only at $z = 0$; stronger correlations may be found by considering halo properties e.g. at the epoch of peak mass (McGibbon & Khochfar 2022).

Our analysis also provides insight into 'inverse SHAM', where galaxy properties are used to estimate the halo mass of an object. From Fig. 7 we see that, as expected, $M_\star$ in H-AGN and $M_{\rm gas}$ in TNG are the best predictors of $M_{\rm h}$ for centrals and satellites. The gas mass tracks the halo mass particularly well in massive haloes where the gas-to-halo mass approaches the ratio of the cosmic baryon-to-dark matter density. Interestingly, we find stellar metallicity Z to be the next most important variable for both satellites and centrals in H-AGN, while stellar mass plays this role in TNG. Especially for satellite galaxies these secondary correlations can be relatively important, indicating significant galaxy assembly bias. Our ML model for performing the function of SHAM or inverse SHAM would outperform these models by a considerable margin for satellites, suggesting the need for refinement of the model to reproduce the results of hydrodynamical simulations with high fidelity. This could be achieved within the framework of 'conditional abundance matching' (Hearin et al. 2014).

5.2 Galaxy size models

Next we consider models for galaxy size, which traditionally employ assumptions about AM. Specifically, the assumption that galaxy and halo specific angular momenta are comparable has been a staple of semi-analytic modelling for over two decades, and would be expected in the absence of significant transfer or loss of AM during galaxy formation. The model may be generalized by taking the specific AM of galaxies and haloes to be proportional rather than identical, where the coefficient of proportionality $j_{\rm d}$ would describe a quasi-universal transfer of AM on average. This allows galaxy disc sizes $R_{1/2}$ to be calculated as a function of halo spin $\lambda_{\rm h}$, halo mass $M_{\rm h}$, and flat rotation curve velocity $V_{\rm c}$,

$$R_{1/2} = \frac{\lambda_{\rm h} G M_{\rm h}^{3/2}}{2 V_{\rm c} |E|^{1/2}} \left(\frac{j_{\rm d}}{m_{\rm d}} \right) \approx \frac{j_d}{m_d} \lambda_{\rm h} R_{\rm h}, \tag{18}$$

where $m_{\rm d}$ is a fixed fraction of the virial mass that settles into the galaxy disc and E is the total energy of the halo (Mo et al. 1998). Using a realistic $M_\star$–$M_{\rm vir}$ relation it has been shown that taking $j_{\rm d} \simeq 0.6$–1 gives the correct normalization for the $M_\star$–$R_{1/2}$ relation (Desmond & Wechsler 2015; Posti et al. 2018; Mancera Piña et al. 2021), although it overpredicts the scatter in sizes at fixed stellar mass and implies an unobserved strong anticorrelation between residuals of the Tully–Fisher and mass–size relations (de Jong & Lacey 2000; Gnedin et al. 2007; Desmond & Wechsler 2015; Desmond et al. 2019). The model of Kravtsov (2013) is motivated by that of Mo et al. although it is formally separable from it: it suggests that $R_{1/2}$ is ∼15 per cent of R_{200} regardless of $M_\star$. Unlike Mo et al., this may but does not have to imply a correlation between $R_{1/2}$ and $\lambda_{\rm h}$.

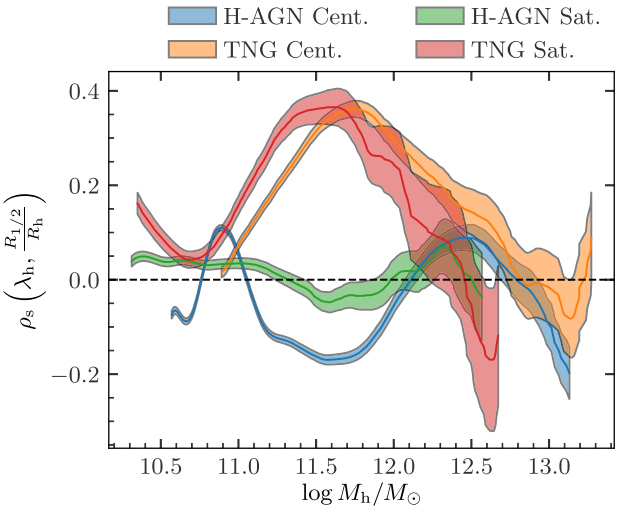

Figure 9. Spearman correlation coefficient ρ_s between the halo spin $\lambda_{\rm h}$ and the galaxy-to-halo size ratio $R_{1/2}/R_{\rm h}$ as a function of $\log M_{\rm h}$ in H-AGN and TNG central and satellite galaxies. ρ_s is calculated in a moving window of width 0.5 dex in $\log M_{\rm h}$. Shaded bands indicate the $1\,\sigma$ uncertainties in bootstrap resamples.

Several studies have tested the Mo et al. model against hydrodynamical simulations, typically finding far less correlation between $R_{1/2}$ and $\lambda_{\rm h}$ than the model predicts (Sales et al. 2009, 2012; Desmond et al. 2017; Stevens et al. 2017; Jiang et al. 2019; Yang et al. 2021; Rohr et al. 2022). This accords with the empirical deficiencies of the model as described above. Our Fig. 2 shows that the prediction of $R_{1/2}$ is little helped by knowledge of $\lambda_{\rm h}$: in H-AGN it is the third most important halo feature for centrals and the fifth for satellites, while for TNG, although it is the second and third most important variable, respectively, the R^2 score of the $R_{1/2}$ prediction never reaches 0.5. This implies that galaxy size is not strongly correlated with *any* halo property, indicating no significant correlations of $R_{1/2}$ beyond the $R_{1/2}$–$M_\star$ relation. This was also found in the EAGLE simulation (Desmond et al. 2017), and argued for on the basis of galaxies' dynamical scaling relations in Desmond (2017a) and Desmond et al. (2019). In addition, Fig. 7 shows that λ_h cannot be predicted reliably from any set of galaxy properties.

To test this further, in Fig. 9 we calculate the Spearman coefficient of the $\lambda_{\rm h}$–$R_{1/2}/R_{\rm h}$ relation in a moving bin of $\log M_{\rm h}$ of width 0.5 dex, separately for centrals and satellites in each simulation. It is expected that $\lambda_{\rm h}$ has a linear correlation with $R_{1/2}/R_{\rm h}$ in the model of Mo et al. (equation 18), but instead we find that the correlation is always weak ($\rho_{\rm s} < 0.4$). Indeed, in H-AGN it is *negative* for centrals at $M_{\rm h} \approx 10^{12}\,{\rm M}_\odot$, indicating that faster-spinning haloes host relatively *smaller* galaxies.

Finally, we consider the correlations of the residuals of the $R_{1/2}$–$M_\star$, $\lambda_{\rm h}$–$M_\star$, and $M_{\rm h}$–$M_\star$ relations to remove the mean trend of these

variables with $M_\star$. The residuals are defined by

$$\Delta \log X|M_\star \equiv \log(X) - \langle \log(X)|\log(M_\star)\rangle, \quad (19)$$

where $\langle \log(X)|\log(M_\star)\rangle$ is the expectation for $\log(X)$ at given $\log(M_\star)$ from fitting a shallow forest of decision trees on subsets of the entire data set. We plot the resulting correlations for of $\Delta\log\lambda_\mathrm{h}|M_\star - \Delta\log R_{1/2}|M_\star$ in Fig. 10 and $\Delta\log M_\mathrm{h}|M_\star - \Delta\log R_{1/2}|M_\star$ in Fig. 11, indicating in each panel the Spearman's rank correlation coefficient and its uncertainty from bootstrap resamples of the data. We find no particularly significant correlations in either simulation, in contrast to the expectations from the models of Mo et al. and Kravtsov, respectively.

We conclude that, in accordance with observations and other hydrodynamical simulations, the model of Mo et al. is not at play in H-AGN, while in TNG the correlation strengths fall far short of those expected from AM partition. We also find that the model of Kravtsov essentially reduces to the R_eff–$M_\star$ plus $M_\star$–M_vir relations; this is further shown by the fact that the linear correlation $R_{1/2}$–$M_\star$ is considerably stronger than that of $R_{1/2}$–M_h, although their Spearman correlation coefficients are comparable due to the strong $M_\star$–M_h relation. Lastly, we consider predicting the residuals $\Delta\log R_{1/2}|M_\star$ from the halo properties, finding that in all cases the R^2 scores of the predictions are $\lesssim 0.1$. A corollary is that $R_{1/2}$ provides little to no useful information for inferring M_h except through its correlation with $M_\star$, as also evidenced by Fig. 2 (contra Rohr et al. 2022). More detailed numerical studies of AM transport in forming galaxies are needed to provide a physical understanding of these results.

6 DISCUSSION

6.1 The galaxy–halo connections in the H-AGN and Illustris-TNG simulations

The set of galaxy and halo properties from a simulation populates a region of a high-dimensional space. As a simple example, consider the three-dimensional space of $M_\star$–M_h–c_h, with $M_\star$ considered as the dependent variable. A typical, point-wise supervised ML algorithm learns the best-fitting surface parameterized by M_h and c_h on which to find values of $M_\star$ consistent with the training data, such that the true values of $M_\star$ scatter minimally but by an unknown amount around it. On the other hand, our NN ensemble with a Gaussian loss function quantifies this scatter in addition to the optimal manifold, which it learns from the training data of simulated points in the $M_\star$–M_h–c_h space. This allows our method to capture fully the range of possible values of $M_\star$, by means of a (Gaussian) probability distribution, from unseen M_h and c_h values.

The scatter that we are interested in learning represents the statistical uncertainty that describes the stochasticity of the simulated galaxy–halo connections, and which will tend towards a constant, typically non-zero value in the limit of infinite number of training samples. This is approximately summed in quadrature with an additional uncertainty due to the finite number of training samples. Going back to the example of the optimal manifold, parameterized by M_h and c_h, the statistical uncertainty represents the scatter around the best-fitting manifold. By means of bootstrap resampling and training the models on subsets of the data we have verified that the contribution to the scatter from the finite size of the training set is subdominant to the intrinsic scatter. An important source of systematic uncertainty is the dependence of the results on the astrophysical and cosmological parameters assumed in the simulations. While we have some information on this from our use of two different simulations, a systematic investigation is lacking (see also Section 6.2).

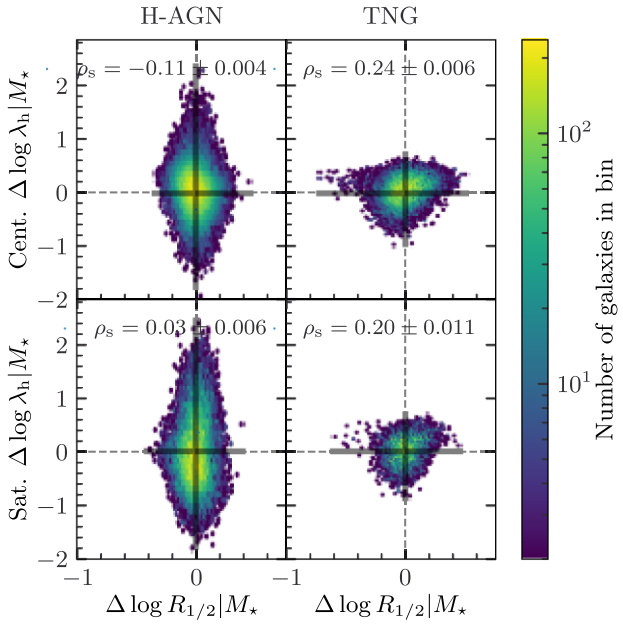

Figure 10. Correlation between the residuals of the λ_h–$M_\star$ and $R_{1/2}$–$M_\star$ relations in H-AGN (*left* column) and TNG (*right* column) central (*top* row) and satellite (*bottom* row) galaxies. These residuals would be strongly correlated in the Mo et al. (1998) model. Within each panel we indicate the Spearman correlation coefficient and its uncertainty estimated by bootstrapping.

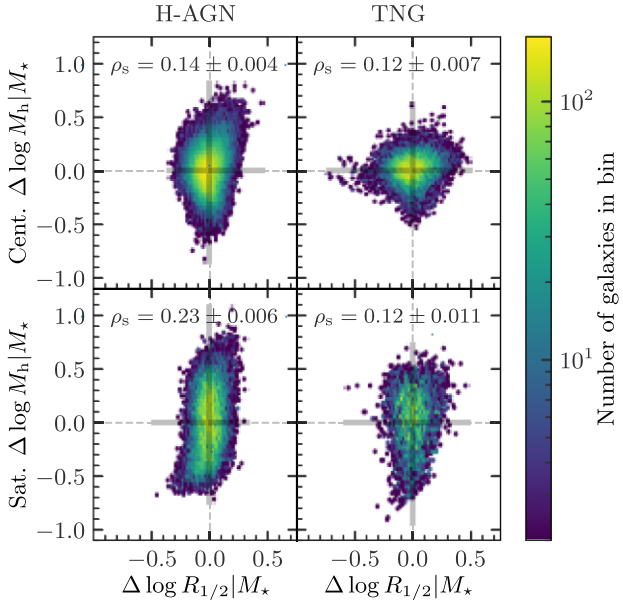

Figure 11. Same as Fig. 10, except for the M_h–$M_\star$ and $R_{1/2}$–$M_\star$ residual correlations. These would be strongly correlated in the Kravtsov (2013) model.

We have trained our models on informative features only, by adding a feature only if it improves the R^2 score on the test set by at least $\Delta R^2 = 0.005$. In Figs 2, 7, and 5 we observe that training a model on additional, non-informative features may marginally worsen the goodness-of-fit, despite Extra-Trees being known for resistance to the presence of non-informative features. Furthermore, with our NN ensemble we observe that the variation between the models artificially increases, with the NNs being less likely to converge to

a 'good' minimum of the loss function if trained on uninformative features as well. This makes it preferable to invest time in determining which features are important for predicting a given target to achieve optimal performance, as opposed to inputting all available features into the model, many of which may act only as noise with respect to the target quantity. We now discuss our results for learning to predict $M_\star$, $M_{\rm gas}$, $R_{1/2}$, $M_{\rm H\,I}$, and $M_{\rm h}$.

In Fig. 2 we show that $M_\star$ can be predicted from the halo properties in both central and satellite galaxies with similar, high accuracy ($R^2 > 0.9$), which is consistent with previous studies (Kamdar et al. 2016; Chen et al. 2021; de Santi et al. 2022; Lovell et al. 2022). While in H-AGN the only relevant property beyond $M_{\rm h}$ is $c_{\rm h}$, yielding $\Delta R^2 \approx 0.07$, in TNG the combination of $\lambda_{\rm h}$, $R_{\rm h}$, and $c_{\rm h}$ is required. On the other hand, in satellite galaxies we find that in both simulations the halo size $R_{\rm h}$ along with $c_{\rm h}$ is required to match the goodness-of-fit of central galaxies. Secondary halo features account for ∼50 per cent of the scatter in the $M_\star$–$M_{\rm h}$ relation as shown in Fig. 3, indicating that their inclusion may benefit semi-empirical models such as SHAM, especially in satellites. That the features required to predicted masses are similar for both simulations is consistent with other work (Chen et al. 2021; Villaescusa-Navarro et al. 2021a), although we find that the relevant features for other properties (e.g. $R_{1/2}$) are simulation-dependent.

In both H-AGN and TNG central galaxies the total gas mass $M_{\rm gas}$ is well constrained from $M_{\rm h}$ alone ($R^2 \approx 0.87$), with the importance of other features being near-negligible ($\Delta R^2 \approx 0.02$). In satellite galaxies secondary properties are important, allowing $\Delta R^2 \approx 0.2$. We show the importance of such features in Fig. 2, although these can only explain some of the $M_{\rm h}$–$M_{\rm gas}$ scatter in satellite galaxies (Fig. 4). In all cases we find the stellar mass to be better constrained than the total gas mass of galaxies from the $z = 0$ halo properties.

Predictions of the stellar half-mass radius $R_{1/2}$ are strongly simulation-dependent as the halo-to-$R_{1/2}$ mapping in H-AGN yields R^2 values higher by up to 0.2 than in TNG, and the simulations have different secondary feature importances. This indicates a lack of convergence of state-of-the-art hydrodynamical simulations in the physics relevant to galaxy sizes. In both simulations $R_{1/2}$ is however poorly constrained, as was also found in the IllustrisTNG300 simulation (de Santi et al. 2022). We find that no property beyond $M_{\rm h}$ can account for a significant amount of the $R_{1/2}$–$M_{\rm h}$ scatter and that $R_{1/2}$ can be predicted primarily because of its correlation with $M_\star$, which in turn correlates with $M_{\rm h}$. We discuss this in further detail in Section 5. We also study the neutral hydrogen mass relation in TNG, finding it to be poorly constrained from ($z = 0$) halo features alone ($R^2 \leq 0.5$), though halo spin $\lambda_{\rm h}$ and, in the case of satellites, the parent halo virial mass are significant (Fig. 5). It is possible that the halo-to-$R_{1/2}$ or $M_{\rm H\,I}$ connections could be better constrained if the feature set was augmented with environmental properties and/or halo variables at $z > 0$.

As expected from the results above, we find the central-galaxy-to-$M_{\rm h}$ connection to be near perfectly constrained in both H-AGN and TNG ($R^2 \approx 0.95, 0.98$, respectively). The satellite connection is not as well constrained ($R^2 \approx 0.8, 0.9$), however, as we show in Fig. 7. Properties beyond $M_\star$ account for up to 50 per cent of the scatter in the $M_{\rm h}$–$M_\star$ relation (Fig. 8), such that the central and satellite halo masses can be constrained with an uncertainty of ~ 0.1 dex and ~ 0.1–0.2 dex, respectively. This is comparable to the scatter found in similar studies which analyse the CAMELS simulations (Shao et al. 2021; Villanueva-Domingo et al. 2021). The central halo mass can be nearly as well predicted from magnitudes only, although in satellites the predictions are again worse (Fig. 7). We find the i-band magnitude to be the most important, as expected from its strongest correlation with $M_\star$ compared to the other bands. Although, if magnitudes only are used as inputs, we find that the scatter in $M_{\rm h}$ starts significantly growing towards brighter galaxies. The predictive power of raw photometry is comparable to that found by von Marttens et al. (2021), who also studied the TNG simulations but with a Tree-based Pipeline Optimization Tool (Olson et al. 2016). They found that the total mass could be predicted from magnitudes with $R^2 \lesssim 0.86$ and from all features with $R^2 \approx 0.98$, again in agreement with our results (see Section 4.2.1).

We also consider predicting the halo spin, finding in both simulations that no combination of galaxy properties can predict $\lambda_{\rm h}$ to better than $R^2 \approx 0.25$. Fig. 2 shows that the stellar half-mass radius $R_{1/2}$ is never a particularly important feature for predicting $\lambda_{\rm h}$. This provides further evidence that the model of Mo et al. (1998) is not present in either H-AGN or TNG, which implies that $R_{1/2}$ should be proportional to $\lambda_{\rm h}$ (equation 18).

Lastly, we also test learning to predict the halo concentration $c_{\rm h}$ from galaxy properties, finding that the results are strongly dependant on the concentration definition. In H-AGN the concentration is poorly constrained from the galaxy features with $R^2 \approx 0.53, 0.38$ in central and satellite galaxies, respectively. In TNG these numbers are $R^2 \approx 0.12, 0.32$. This suggests that in case of TNG centrals the model can do little better than predict the average value of concentration. As per Section 2.2.2, we define the concentration to be the ratio between $R_{\rm h}$ and $r_{\rm s}$. However, while in H-AGN $R_{\rm h}$ is the virial radius of the halo, in TNG it is the distance to the furthest gravitationally bound DM particle. A consequence is that the halo concentration correlates much more weakly with halo mass in TNG than in H-AGN and is therefore less predictable from galaxy variables.

6.2 Future work

6.2.1 Applications

Our method has a range of applications. The most obvious is to elucidate the galaxy–halo connections in hydrodynamical simulations by quantifying the intrinsic scatter of various multiparameter correlations. This in turns promotes understanding of the effects of the feedback and subgrid prescriptions implemented in the simulations, and informs the construction of semi-empirical and semi-analytic models that attempt to capture the salient parts of this physics. In common with other ML analyses, our techniques also allow hydrodynamical simulation to be bypassed in connecting galaxy and halo variables. This can be used either to paste galaxies on to haloes produced in N-body simulations in a manner that mimics their association in full-physics simulations, or, in the opposite direction, to infer haloes from observed galaxies. The former is useful for making mock universes from cheap DM-only simulations – which can then be used either for forecasting upcoming cosmological galaxy surveys (e.g. Reddick et al. 2014; Zhang et al. 2019; Kasmanoff et al. 2020; Wechsler et al. 2022) or for validating and developing analysis pipelines for these surveys (e.g. DeRose et al. 2019) – while the latter is useful for estimating the distribution of DM associated with the observed distribution of light. This has been explored in the context of predicting halo masses of an observed galaxy and group catalogue from the Sloan Digital Sky Survey, finding that it outperforms conventional halo abundance matching or dynamical mass modelling (Calderon & Berlind 2019)

The advantage of using ML applied to simulations is that it provides a powerful way of capturing *all* relevant correlations, regardless of their (unknown) functional form, at the cost of reduced

generality in that the model is specific to the simulation it was trained on. We have seen important similarities and differences between the galaxy–halo connections of H-AGN and IllustrisTNG, indicating that some parts of the model are robust to simulation-specific physics while others are not. In the future, techniques such as Domain Adaptation and Transfer Learning could be used to generalize the ML models and extend their functionality across multiple data sets and hence create less simulation-specific results (e.g. Ćiprijanović et al. 2021).

6.2.2 *Simulations and features*

As described in Section 6.1, our NNs learn the high-dimensional manifold characterizing the mean galactic and halo properties, and how far from this manifold objects typically reside. Since the location and shape of the manifold depends on cosmological and astrophysical parameters (see e.g. Villaescusa-Navarro et al. 2022b), in this work we have compared two simulations with different cosmologies and subgrid physics models. One way to extend our framework is to consider a wider range of training sets which systematically vary such differences, which would improve understanding of the sensitivity of the results to the simulation inputs. A useful meta-sample comes from the CAMELS project (Villaescusa-Navarro et al. 2021c, 2022a), a suite of 2000 magneto-hydrodynamic simulations with varying initial conditions, cosmological and baryonic parameters, and two different subgrid physics models, designed with ML applications in mind (e.g. Villaescusa-Navarro et al. 2021b). The CAMELS boxes are, however, small (25 Mpc/h in side length), increasing sample variance.

Other existing simulation suites, of large box sizes with adequate resolution for studies of more massive haloes, already contain multiple runs that vary input parameters for AGN feedback (e.g. McCarthy et al. 2017) and cosmology (e.g. Magneticum Pathfinder),[7] or vary whole galaxy formation models (eg. Cui et al. 2018). Ensembles constructed by combining different suites have already been used to study the impact of galaxy formation (or lack thereof) on some galaxy/halo observables (e.g. Anbajagane et al. 2020; Lim et al. 2021; Ragagnin et al. 2021; Anbajagane et al. 2022b), and such ensembles also have the potential to serve as training sets. An obvious disadvantage here however is that while CAMELS provides a homogeneous parametrization for feedback, an ensemble constructed from disparate simulations does not. In principle, an 'effective', common parametrization of the simulation suites is possible, but given basic uncertainties in small-scale baryonic physics and its implementation this would not be simple to construct.

Future work should consider whether the inclusion of other galaxy or halo variables, such as summaries of the merger history as in Jo & Kim (2019), could further improve the NN's predictions. Also interesting is how the scatter in the galaxy–halo connection evolves as a function of redshift. Studies of the Illustris-TNG simulation using Extremely Randomised Trees (McGibbon & Khochfar 2022) have found that it is possible to make significantly more accurate predictions by including features from a range of redshifts. In particular, low-redshift observables are most important for determining the gas mass and star formation rate, although predictions of stellar mass and metallicity benefit from higher redshift information. There are hints of universal relations between total halo mass and galaxy properties which are independent of redshift (Shao et al. 2021); it would be interesting to see whether such relations exist for the scatter as well as the mean prediction.

Finally, we have not considered environmental properties of the galaxy or halo in this work, and have thus treated all objects equally, independent of their location in the simulation box. Since it has been shown that environmental overdensity and shear can improve clustering when predicting the halo occupation distribution (Delgado et al. 2021) and that the local environment is an important feature in predicting galactic properties (Agarwal et al. 2018), there is potentially useful information in such features which should be exploited in future work. However, we note that Lovell et al. (2022) report negligible improvements when predicting the majority of galaxy properties based on their host haloes in the EAGLE/C-EAGLE simulations with the local density included in the feature set.

6.2.3 *Methodology*

Although we have been able to determine which features are important for predicting certain properties and have attempted to decipher the correlations being utilized, as with many ML analyses interpreting the results can be challenging. Techniques such as symbolic regression, which learns analytic forms for the relationships between variables, have already been used to find analytic approximations to the mean relations between galaxy and halo properties (Delgado et al. 2021; Shao et al. 2021), and may profitably be applied also to the scatter.

In Section 3.1 we outlined how we calculate the bivariate scatter in a relation Y–X at fixed X by fitting a univariate Gaussian distribution to samples binned by X. In doing so we have assumed that a Gaussian distribution is sufficient to describe the distribution of samples within each bin. However, this assumption will be violated in case of e.g. heavy tails or the presence of outliers. These would increase the best-fitting scatter and make its interpretation non-trivial. Therefore, an optimal approach would allow for a more general distribution than a Gaussian. Similarly, in equation (13), we have assumed a Gaussian loss function, although in principle the Gaussian distribution can be replaced with any probability distribution. Moment Networks (Jeffrey & Wandelt 2020) could be used to determine the mean and scatter (i.e. the first two moments) of these relationships in a likelihood-free manner, without assuming a Gaussian distribution.

For each base network, q, we find the mean, μ_q, and scatter, σ_q, that maximize the likelihood, such that their combination through equations (16a) and (16b) should provide a good estimate for the true maximum likelihood. However, the distributions of μ_q and σ_q cannot be interpreted as posterior distributions for μ and σ, and thus we cannot assign an object-by-object uncertainty to the parameters μ and σ. In plots such as Fig. 3 we therefore assign uncertainties using the spread among galaxies in a given mass bin. Future work could use methods such as Bayesian Neural Networks (e.g. Valentin Jospin et al. 2020) to determine the uncertainties on the probability distributions themselves.

We have either manually tested which hyperparameters of the NN architecture (depth, width, activation function, optimizer, etc.) represent the data the best or performed a brute-force hyperparameter grid search in case of Extra-Trees. Although we do not expect our results to be heavily architecture-dependent (the several methods we tested gave similar results), we cannot rule out the possibility that there exists some superior hyperparameter combination. In principle one should search the entire space of hyperparameters for a way to do this automatically and efficiently (see e.g. Akiba et al. 2019 and references therein).

[7] http://www.magneticum.org/

Finally, it would be possible to predict several target variables simultaneously with a multivariate loss function, allowing the model to learn the correlations between the target variables. In combination with the above this would ultimately allow the most faithful characterization of the galaxy–halo connection.

7 CONCLUSION

We build an ensemble of neural networks with a Gaussian loss function to study the $z = 0$ galaxy–halo connection in the H-AGN and Illustris-TNG100 (TNG) simulations, with particular emphasis on modelling the intrinsic scatter of this connection. This is equivalent to an uncertainty in predicting a galaxy/halo property when some subset of halo/galaxy properties is known. We quantify the importance of various halo properties for the prediction of $M_\star$, $M_{\rm gas}$, $R_{1/2}$, and $M_{\rm H\,I}$, and of various galaxy properties in the prediction of $M_{\rm h}$ and $\lambda_{\rm h}$. We then examine how the addition of secondary properties decreases the predicted uncertainty relative to the uncertainty of a model trained on the most important property only. This enables us to quantify, for example, how much scatter in the $M_\star$–$M_{\rm h}$ relation can be attributed to residual correlations with e.g. the halo concentration $c_{\rm h}$, which we show in Fig. 3. We validate our bespoke neural network against an Extra-Trees decision tree ensemble. Lastly, we explore the implications of our findings for semi-empirical galaxy formation models. We find that while SHAM captures the most important correlations of the simulations it may be possible to improve it, and that galaxy size–halo spin (or halo size) models do not provide a good fit to the simulated data or aid in interpreting it.

Our main findings are:

(i) While the halo-to-stellar mass relation is well-constrained in central galaxies, with a relatively low importance of any other halo quantity, properties such as the halo size and concentration play an important role in satellite galaxies. After accounting for these properties the main halo and subhalo masses can be predicted nearly equally well ($R^2 \geq 0.9$). Secondary properties account for up to 0.1 dex and 0.2 dex of the scatter in the $M_\star$–$M_{\rm h}$ relation in central and satellite galaxies, respectively.

(ii) In both H-AGN and TNG, stellar mass is more predictable from halo properties than the total gas mass with R^2 higher by $\sim$0.05 and $\sim$0.10 in central and satellite galaxies, respectively. The halo-to-neutral hydrogen mass relation is poorly constrained with $R^2 \approx 0.5$.

(iii) Main halo masses can be predicted near perfectly in both simulations from galaxy properties ($R^2 \geq 0.95$), although not subhalo masses. Furthermore, while the main halo mass can be predicted as well from magnitudes only, that is not the case for subhaloes. The uncertainty in the halo mass is $\sim$0.1 and 0.2 dex in centrals and satellites, respectively, after accounting for all galaxy properties. However, if only directly observable properties (photometric magnitudes) are considered then the uncertainty grows towards the bright end, most notably in satellites where it reaches up to $\sim$0.4 dex.

(iv) We find that galaxy size-based semi-analytical models are not at play in either H-AGN nor TNG, although the quantitative correlations relevant to these models are simulation-dependent. In both simulations the correlation between galaxy and halo size can essentially be reduced to the separate correlations of galaxy size with stellar mass, stellar mass with halo mass, and halo mass with halo size.

By studying the H-AGN and TNG simulations we showcase the ability of ML to quantify all relevant correlations of the galaxy–halo connection, including their intrinsic scatter. By modelling the statistical uncertainty in predictions for galaxy properties based on halo ones, or vice versa, this can be used to introduce stochasticity into the generation of mock galaxy or halo samples and hence produce more realistic and useful catalogues.

ACKNOWLEDGEMENTS

We thank Julien Devriendt, Clotilde Laigle, and Benjamin P. Moster for input and discussions. We also thank Jonathan Patterson for smoothly running the Glamdring Cluster hosted by the University of Oxford, where most of the data processing was performed.

RS is supported by the Deutscher Akademischer Austauschdienst (DAAD) Study Scholarship. DJB is supported by the Science and Technology Facilities Council (STFC) and Oriel College Oxford. HD was supported by St John's College, Oxford, a McWilliams Fellowship at Carnegie Mellon University, and a Royal Society University Research Fellowship (grant no. 211046). DA is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1746045. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 693024).

We would like to thank the Horizon-AGN collaboration for allowing us to use the simulation data, and particularly Stephane Rouberol for smoothly running the Horizon Cluster hosted by the Institut d'Astrophysique de Paris where most of the processing of the raw simulation data was performed. Some of the numerical work also made use of the DiRAC Data Intensive service at Leicester, operated by the University of Leicester IT Services, which forms part of the STFC DiRAC HPC Facility (www.dirac.ac.uk). The equipment was funded by BEIS capital funding via STFC capital grants ST/K000373/1 and ST/R002363/1 and STFC DiRAC Operations grant ST/R001014/1. DiRAC is part of the National e-Infrastructure.

We thank the IllustrisTNG team for publicly releasing all simulation data. The IllustrisTNG simulations were undertaken with compute time awarded by the Gauss Centre for Supercomputing (GCS) under GCS Large-Scale Projects GCS-ILLU and GCS-DWAR on the GCS share of the supercomputer Hazel Hen at the High Performance Computing Center Stuttgart (HLRS), as well as on the machines of the Max Planck Computing and Data Facility (MPCDF) in Garching, Germany.

DATA AVAILABILITY

We make our NN implementation publicly available at https://github.com/Richard-Sti/scatternn. The Horizon-AGN data is available from the Horizon-AGN collaboration on request. The IllustrisTNG data is publicly available at https://www.tng-project.org/data/ and the public release products are described in Nelson et al. (2019). All other data will be shared on reasonable request to the authors.

REFERENCES

Abadi M. et al., 2015, preprint (arXiv:1603.04467)
Agarwal S., Davé R., Bassett B. A., 2018, MNRAS, 478, 3410
Akiba T., Sano S., Yanase T., Ohta T., Koyama M., 2019, preprint (arXiv:1907.10902)
Anbajagane D., Evrard A. E., Farahi A., Barnes D. J., Dolag K., McCarthy I. G., Nelson D., Pillepich A., 2020, MNRAS, 495, 686
Anbajagane D., Evrard A. E., Farahi A., 2022a, MNRAS, 509, 3441
Anbajagane D. et al., 2022b, MNRAS, 510, 2980
Aubert D., Pichon C., Colombi S., 2004, MNRAS, 352, 376

Bartlett D. J., Desmond H., Devriendt J., Ferreira P. G., Slyz A., 2021, MNRAS, 500, 4639
Baugh C. M., 2006, Rep. Prog. Phys., 69, 3101
Behroozi P. S., Conroy C., Wechsler R. H., 2010, ApJ, 717, 379
Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
Breiman L., 2001, Mach. Learn., 45, 5
Bryan G. L., Norman M. L., 1998, ApJ, 495, 80
Bullock J. S., Dekel A., Kolatt T. S., Kravtsov A. V., Klypin A. A., Porciani C., Primack J. R., 2001, ApJ, 555, 240
Calderon V. F., Berlind A. A., 2019, MNRAS, 490, 2367
Capano C., Stiskalek R., 2021, cdcapano/epsie. Zenodo
Chabrier G., 2003, PASP, 115, 763
Chen Y., Mo H. J., Li C., Wang K., Wang H., Yang X., Zhang Y., Katz N., 2021, MNRAS, 507, 2510
Cheng H.-T. et al., 2016, in Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. DLRS 2016. Association for Computing Machinery, New York, p. 7
Chipman H. A., George E. I., McCulloch R. E., 2010, Ann. Appl. Stat., 4, 266
Chisari N. et al., 2015, MNRAS, 454, 2736
Chisari N. E. et al., 2017, MNRAS, 472, 1163
Ćiprijanović A. et al., 2021, MNRAS, 506, 677
Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, 201
Cui W. et al., 2018, MNRAS, 480, 2898
de Jong R. S., Lacey C., 2000, ApJ, 545, 781
de los Rios M. E., Petač M., Zaldivar B., Bonaventura N. R., Calore F., Iocco F., 2021, preprint (arXiv:2111.08725)
de Santi N. S. M., Rodrigues N. V. N., Montero-Dorta A. D., Abramo L. R., Tucci B., Artale M. C., 2022, MNRAS, 514, 2463
Delgado A. M., Wadekar D., Hadzhiyska B., Bose S., Hernquist L., Ho S., 2021, preprint (arXiv:2111.02422)
DeRose J. et al., 2019, preprint (arXiv:1901.02401)
Desmond H., 2017a, MNRAS, 464, 4160
Desmond H., 2017b, MNRAS, 472, L35
Desmond H., Wechsler R. H., 2015, MNRAS, 454, 322
Desmond H., Wechsler R. H., 2017, MNRAS, 465, 820
Desmond H., Mao Y.-Y., Wechsler R. H., Crain R. A., Schaye J., 2017, MNRAS, 471, L11
Desmond H., Katz H., Lelli F., McGaugh S., 2019, MNRAS, 484, 239
Diemer B. et al., 2018, ApJS, 238, 33
Diemer B. et al., 2019, MNRAS, 487, 1529
Dietterich T. G., 2000, in Multiple Classifier Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, p. 1
Dillon J. V. et al., 2017, preprint (arXiv:1711.10604)
Dolag K., Borgani S., Murante G., Springel V., 2009, MNRAS, 399, 497
Donnari M. et al., 2019, MNRAS, 485, 4817
Duan T., Avati A., Ding D. Y., Thai K. K., Basu S., Ng A. Y., Schuler A., 2019, preprint (arXiv:1910.03225)
Dubois Y., Teyssier R., 2008, A&A, 477, 79
Dubois Y., Pichon C., Devriendt J., Silk J., Haehnelt M., Kimm T., Slyz A., 2013, MNRAS, 428, 2885
Dubois Y. et al., 2014, MNRAS, 444, 1453
Dutta S., Khandai N., Rana S., 2021, MNRAS, 511, 2585
Eisert L., Pillepich A., Nelson D., Klessen R. S., Huertas-Company M., Rodriguez-Gomez V., 2022, preprint (arXiv:2202.06967)
F. R. S. K. Pearson, 1901, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2, 559
Fall S. M., Efstathiou G., 1980, MNRAS, 193, 189
Geurts P., Ernst D., Wehenkel L., 2006, Mach. Learn., 36, 3
Gnedin O. Y., Weinberg D. H., Pizagno J., Prada F., Rix H.-W., 2007, ApJ, 671, 1115
Goodfellow I. J., Shlens J., Szegedy C., 2014, preprint (arXiv:1412.6572)
Gunn J. E., Gott J., Richard I., 1972, ApJ, 176, 1
Guo H., Li C., Zheng Z., Mo H. J., Jing Y. P., Zu Y., Lim S. H., Xu H., 2017, ApJ, 846, 61
Hadzhiyska B., Bose S., Eisenstein D., Hernquist L., 2021, MNRAS, 501, 1603
Hearin A. P., Watson D. F., Becker M. R., Reyes R., Berlind A. A., Zentner A. R., 2014, MNRAS, 444, 729
Hearin A. P., Zentner A. R., van den Bosch F. C., Campbell D., Tollerud E., 2016, MNRAS, 460, 2552
Ho M., Farahi A., Rau M. M., Trac H., 2021, ApJ, 908, 204
Hornik K., Stinchcombe M., White H., 1989, Neural Netw., 2, 359
Hotelling H., 1936, Biometrika, 28, 321
Jeffrey N., Wandelt B. D., 2020, preprint (arXiv:2011.05991)
Jiang F. et al., 2019, MNRAS, 488, 4801
Jo Y., Kim J.-h., 2019, MNRAS, 489, 3565
Kamdar H. M., Turk M. J., Brunner R. J., 2016, MNRAS, 457, 1162
Kasmanoff N., Villaescusa-Navarro F., Tinker J., Ho S., 2020, preprint (arXiv:2012.00186)
Kaviraj S. et al., 2017, MNRAS, 467, 4739
Kimm T., 2012, PhD thesis, Oxford University, UK
Kingma D. P., Ba J., 2014, preprint (arXiv:1412.6980)
Klambauer G., Unterthiner T., Mayr A., Hochreiter S., 2017, preprint (arXiv:1706.02515)
Kodi Ramanah D., Wojtak R., Ansari Z., Gall C., Hjorth J., 2020, MNRAS, 499, 1985
Kodi Ramanah D., Wojtak R., Arendse N., 2021, MNRAS, 501, 4080
Komatsu E. et al., 2011, ApJS, 192, 18
Kravtsov A. V., 2013, ApJ, 764, L31
Kravtsov A. V., Berlind A. A., Wechsler R. H., Klypin A. A., Gottlöber S., Allgood B., Primack J. R., 2004, ApJ, 609, 35
Laigle C. et al., 2019, MNRAS, 486, 5104
Lakshminarayanan B., Pritzel A., Blundell C., 2016, preprint (arXiv:1612.01474)
LeCun Y., Bottou L., Orr G. B., Muller K., 2012, in Montavon G., Orr G. B., Müller K. R., eds, Neural Networks: Tricks of the Trade. Vol. 7700, Springer, Berlin, Heidelberg
Lehmann B. V., Mao Y.-Y., Becker M. R., Skillman S. W., Wechsler R. H., 2017, ApJ, 834, 37
Li P., McGaugh S. S., Lelli F., Tian Y., Schombert J. M., Ko C.-M., 2022, ApJ, 927, 198
Lim S. H., Barnes D., Vogelsberger M., Mo H. J., Nelson D., Pillepich A., Dolag K., Marinacci F., 2021, MNRAS, 504, 5131
Lovell C. C., Wilkins S. M., Thomas P. A., Schaller M., Baugh C. M., Fabbian G., Bahé Y., 2022, MNRAS, 509, 5046
McCarthy I. G., Schaye J., Bird S., Le Brun A. M. C., 2017, MNRAS, 465, 2936
McGibbon R., Khochfar S., 2022, MNRAS, 513, 5423
Machado Poletti Valle L. F., Avestruz C., Barnes D. J., Farahi A., Lau E. T., Nagai D., 2021, MNRAS, 507, 1468
Mancera Piña P. E., Posti L., Fraternali F., Adams E. A. K., Oosterloo T., 2021, A&A, 647, A76
Man Z.-Y., Peng Y.-J., Shi J.-J., Kong X., Zhang C.-P., Dou J., Guo K.-X., 2019, ApJ, 881, 74
Marinacci F. et al., 2018, MNRAS, 480, 5113
Mo H. J., Mao S., White S. D. M., 1998, MNRAS, 295, 319
Moster B. P., Somerville R. S., Maulbetsch C., van den Bosch F. C., Macciò A. V., Naab T., Oser L., 2010, ApJ, 710, 903
Moster B. P., Naab T., Lindström M., O'Leary J. A., 2021, MNRAS, 507, 2115
Naiman J. P. et al., 2018, MNRAS, 477, 1206
Navarro J. F., Frenk C. S., White S. D. M., 1996, ApJ, 462, 563
Nelson D. et al., 2018, MNRAS, 475, 624
Nelson D. et al., 2019, Comput. Astrophys. Cosmol., 6, 2
Nix D., Weigend A., 1994, in Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94). Vol. 1, p. 55
Ntampaka M. et al., 2019, ApJ, 876, 82
Olson R. S., Bartley N., Urbanowicz R. J., Moore J. H., 2016, preprint (arXiv:1603.06212)
Paranjape A., Kovač K., Hartley W. G., Pahwa I., 2015, MNRAS, 454, 3030
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Pillepich A. et al., 2018a, MNRAS, 473, 4077
Pillepich A. et al., 2018b, MNRAS, 475, 648

Pillepich A. et al., 2019, MNRAS, 490, 3196
Planck Collaboration, 2016, A&A, 594, A13
Posti L., Pezzulli G., Fraternali F., Di Teodoro E. M., 2018, MNRAS, 475, 232
Power C., Navarro J. F., Jenkins A., Frenk C. S., White S. D. M., Springel V., Stadel J., Quinn T., 2003, MNRAS, 338, 14
Ragagnin A., Fumagalli A., Castro T., Dolag K., Saro A., Costanzi M., Bocquet S., 2021, preprint (arXiv:2110.05498)
Rasera Y., Teyssier R., 2006, A&A, 445, 1
Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, ApJ, 771, 30
Reddick R. M., Tinker J. L., Wechsler R. H., Lu Y., 2014, ApJ, 783, 118
Rohr E. et al., 2022, MNRAS, 510, 3967
Rumelhart D. E., Hinton G. E., Williams R. J., 1986, Nature, 323, 533
Sales L. V., Navarro J. F., Schaye J., Dalla Vecchia C., Springel V., Haas M. R., Helmi A., 2009, MNRAS, 399, L64
Sales L. V., Navarro J. F., Theuns T., Schaye J., White S. D. M., Frenk C. S., Crain R. A., Dalla Vecchia C., 2012, MNRAS, 423, 1544
Shao H. et al., 2021, ApJ, 927, 85
Springel V., 2010, MNRAS, 401, 791
Springel V., Hernquist L., 2003, MNRAS, 339, 289
Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726
Springel V. et al., 2018, MNRAS, 475, 676
Stevens A. R. H., Lagos C. d. P., Contreras S., Croton D. J., Padilla N. D., Schaller M., Schaye J., Theuns T., 2017, MNRAS, 467, 2066
Stiskalek R., Desmond H., Holvey T., Jones M. G., 2021, MNRAS, 506, 3205
Teyssier R., 2002, A&A, 385, 337
Trujillo-Gomez S., Klypin A., Primack J., Romanowsky A. J., 2011, ApJ, 742, 16
Tweed D., Devriendt J., Blaizot J., Colombi S., Slyz A., 2009, A&A, 506, 647
Valentin Jospin L., Buntine W., Boussaid F., Laga H., Bennamoun M., 2020, preprint (arXiv:2007.06823)
Villaescusa-Navarro F. et al., 2021a, preprint (arXiv:2109.09747)
Villaescusa-Navarro F. et al., 2021b, preprint (arXiv:2109.10360)
Villaescusa-Navarro F. et al., 2021c, ApJ, 915, 71
Villaescusa-Navarro F. et al., 2022a, preprint (arXiv:2201.01300)
Villaescusa-Navarro F. et al., 2022b, ApJ, 929, 132
Villanueva-Domingo P. et al., 2021, preprint (arXiv:2111.08683)
von Marttens R. et al., 2021, preprint (arXiv:2111.01185)
Wechsler R. H., DeRose J., Busha M. T., Becker M. R., Rykoff E., Evrard A., 2022, ApJ, 931, 145
Weinberger R. et al., 2017, MNRAS, 465, 3291
Yang H., Gao L., Frenk C. S., Grand R. J. J., Guo Q., Liao S., Shao S., 2021, preprint (arXiv:2110.04434)
Zemp M., Gnedin O. Y., Gnedin N. Y., Kravtsov A. V., 2011, ApJS, 197, 30
Zhang X., Wang Y., Zhang W., Sun Y., He S., Contardo G., Villaescusa-Navarro F., Ho S., 2019, preprint (arXiv:1902.05965)

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