



How to empirically model star formation in dark matter haloes – I. Inferences about central galaxies from numerical simulations

Yangyao Chen ^{1,2}★, H. J. Mo,² Cheng Li¹ and Kai Wang ^{1,2}

¹Department of Astronomy, Tsinghua University, Beijing 100084, China

²Department of Astronomy, University of Massachusetts, Amherst, MA 01003-9305, USA

Accepted 2021 March 1. Received 2021 January 20; in original form 2020 September 25

ABSTRACT

We use TNG and EAGLE hydrodynamic simulations to investigate the central galaxy–dark matter halo relations that are needed for a halo-based empirical model of star formation in galaxies. Using a linear dimension reduction algorithm and a model ensemble method, we find that for both star-forming and quenched galaxies, the star formation history (SFH) is tightly related to the halo mass assembly history (MAH). The quenching of a low-mass galaxy is mainly due to the infall-ejection process related to a nearby massive halo, while the quenching of a high-mass galaxy is closely related to the formation of a massive progenitor in its host halo. The classification of star-forming and quenched populations based solely on halo properties contains contamination produced by sample imbalance and overlapping distributions of the two populations. Guided by the results from hydrodynamic simulations, we build an empirical model to predict the SFH of central galaxies based on the MAH of their host haloes, and we model the star-forming and quenched populations separately. Our model is based on the idea of adopting star formation templates from hydrodynamic simulations to reduce model complexity. We use various tests to demonstrate that the model can recover SFHs of individual galaxies, and can statistically reproduce the galaxy bimodal distribution, stellar mass–halo mass and star formation rate–halo mass relations from low to high redshift, and assembly bias. Our study provides a framework of using hydrodynamic simulations to discover, and to motivate the use of, key ingredients to model galaxy formation using halo properties.

Key words: hydrodynamics – galaxies: formation – galaxies: haloes – galaxies: stellar content.

1 INTRODUCTION

In the Λ CDM cosmology, galaxies are luminous objects that form and evolve in the gravitational potential wells of their dark matter haloes in the cosmic density field. A key step to understand how galaxies form and evolve is therefore to understand how galaxies are related to dark matter haloes (see Mo, van den Bosch & White 2010; Wechsler & Tinker 2018, and references therein). Because the dark matter is invisible, direct observation is inaccessible. Meanwhile, numerical simulations based on first principles have also limitations because of the use of subgrid physical processes that are not resolved. Because of these difficulties, a variety of other methods, generally referred to as empirical models, have been developed to link galaxies with dark matter haloes. The details of these models, such as the model architectures and model parameters, can be constrained by observations, such as the galaxy stellar mass functions, two-point correlation functions, and so on. Examples of such models include (sub)halo abundance matching (Mo, Mao & White 1999; Vale & Ostriker 2004; Guo et al. 2010), clustering matching (Guo et al. 2016), age matching (Hearin & Watson 2013; Hearin et al. 2014; Meng et al. 2020), halo occupation distribution (Jing, Mo & Borner 1998; Berlind & Weinberg 2002), conditional luminosity function (Yang, Mo & van den Bosch 2003), conditional colour–magnitude

diagrams (Xu et al. 2018), and those based on star formation histories (SFHs; Lu et al. 2011, 2014a,b, 2015; Moster, Naab & White 2018; Behroozi et al. 2019; Moster et al. 2020). Although these empirical models are able to reproduce a large set of observations, some basic questions remain unresolved.

First, when modelling the galaxy–halo connection, it is not totally clear which halo quantities are the best to use as features to make the link to galaxies, and which set of galaxy quantities is the best in constraining the link. To reproduce the primary galaxy properties, such as stellar mass and luminosity, it is likely that the basic properties of haloes, such as halo mass and peak circular velocity, are the main features to use (see e.g. Reddick et al. 2013, for an extensive study). However, when higher order galaxy properties are concerned, a systematic approach is yet to be found to identify the best set of halo features for the purpose. Moster et al. (2020) showed an example of using random forest regressor to find the important halo properties related to galaxy stellar mass and star formation rate (SFR). This motivates, but has not been used in, the construction of a deep and dense model.

Secondly, the total sets of galaxy and halo properties, which are high dimensional, are too complex to be useful because it is both difficult to incorporate them into models and to interpret their roles in model predictions. For example, the details of the formation histories of individual haloes are complex, so are the star formation and merger histories of individual galaxies. Yet, it is necessary to include them in the modelling, as they carry important information connected

* E-mail: yangyaochen.astro@foxmail.com

to the current state of an object, such as halo concentration (e.g. Navarro, Frenk & White 1997; Jing 2000; Wechsler et al. 2002; Zhao et al. 2003a,b, 2009; MacCìd, Dutton & Van Den Bosch 2008; Jeon-Daniel et al. 2011), halo bias (e.g. Mo & White 1996; Sheth, Mo & Tormen 2001; Gao, Springel & White 2005; Wechsler et al. 2006; Bett et al. 2007; Gao & White 2007; Hahn et al. 2007; Jing, Suto & Mo 2007; Li, Mo & Gao 2008; Faltenbacher & White 2010; Wang et al. 2011), galaxy colour or SFR (e.g. Hearin & Watson 2013; Hearin et al. 2014; Lim et al. 2016; Wang et al. 2018; Meng et al. 2020; Shi et al. 2020), and galaxy structure (e.g. Kauffmann et al. 2003; Shen et al. 2003; Bernardi et al. 2007; Gao & Fan 2020; Yoon & Park 2020). Attempts have been made to use various formation redshifts to describe halo assembly histories (AHs; see e.g. Navarro et al. 1997; van den Bosch 2002; Li et al. 2008; Zhao et al. 2009; Jeon-Daniel et al. 2011; Wang et al. 2011; Shi et al. 2018). However, as shown in Chen et al. (2020), the information provided by these formation times is incomplete, and there is strong degeneracy among them.

Thirdly, because of the complexity in the galaxy–halo connection, it is important to know how we construct a model that is general but can still facilitate clear physical interpretations of the results it produces. The ‘performance-interpretation’ trade-off is a common problem in model construction. For example, the empirical model of Lu et al. (2014a) used a physically motivated relation between SFR and halo mass and redshift, while Behroozi et al. (2019) used the growth of peak circular velocity to rank the SFR. In both models, the physical meaning of the galaxy–halo connection is clear, but both may have missed other potentially important factors as well as nuanced processes that can cause uncertainties in the relations. In contrast, an empirical model based on densely connected neural networks, such as the one developed by Moster et al. (2018), usually uses multiple hidden layers to get a good representation of the halo properties and regresses them on stellar mass and SFR. The model is accurate, as long as there are sufficient constraints from observations, but the representation of haloes is in a ‘black box’, making it difficult to interpret the results. Some other empirical models invoke multiple ingredients, for example, by separating central galaxies from satellites and/or star forming from quenched galaxies, and use observations to constrain the joint distribution of model parameters (e.g. Lu et al. 2014a, 2015; Moster et al. 2018; Behroozi et al. 2019). Even in such models, it is still challenging to show which ingredients dominate the prediction error, and whether the discrepancy with observations owes to the incapability of the model or to the incompleteness of the observation constraints.

In this paper, we carry out a systematic investigation of the ingredients that are needed to construct a powerful empirical model of galaxy formation based on dark matter haloes. We use inferences from hydrodynamic simulations to motivate a potentially useful architecture to build such an empirical model. To address the first problem described above, we adopt a model ensemble algorithm, the gradient boosted decision trees (GBDT; see Appendix B), which can be used not only to capture complicated patterns between variables and to keep a good balance between bias and variance, but also to identify the most important variables that explain model predictions. We address the second problem by using a linear dimension reduction algorithm, the principal component analysis (PCA; see Appendix A), which can effectively reduce the dimension of the halo AH and the galaxy SFH and yet retain large amounts of information of the histories for the empirical modelling. Finally, we address the third problem by building a deep model that incorporates components of both dimension reduction and ensemble regressor and classifier.

Each component in the model is motivated physically and can be optimized separately. This approach makes the model capable of dealing with complex patterns in parameter space, and yet transparent to interpret. The ensemble regressor and linear dimension reduction method has already been used to study the relationship among halo properties in Chen et al. (2020). Here, we extend it to studying the galaxy–halo connection. As the first in a series, this paper focuses on central galaxies in dark matter haloes. We identify important ingredients that should be included in an empirical model, and demonstrate the limit such a model can reach in describing the stellar masses and SFHs of individual galaxies. Our model is built on the inferences from two hydrodynamic simulations, the Illustris-TNG (e.g. Nelson et al. 2019) and EAGLE (e.g. The EAGLE team 2017).

This paper is organized as follows. In Section 2, we describe the simulation data we use, and define the halo properties and samples used in our analysis. In Section 3, we use both the GBDT and PCA to study the relations of galaxy stellar mass and SFR with halo properties for both star-forming galaxies and quenched galaxies. We also identify halo properties that cause a galaxy to quench. In Section 4, we build an empirical model that predicts the SFH of galaxies in dark matter haloes, testing its performance in several steps. We summarize and discuss our results in Section 5.

2 THE DATA

2.1 The Illustris-TNG and EAGLE simulations

The Illustris-TNG simulation (Marinacci et al. 2018; Naiman et al. 2018; Nelson et al. 2018, 2019; Springel et al. 2018; Pillepich et al. 2018b) is a suite of cosmological, hydrodynamical simulations implemented with the moving-mesh code *Arepo* (Springel 2010). The cosmological parameters are taken from the Planck 2015 results (Planck Collaboration XIII 2015): Hubble constant $H_0 = 100h \text{ km s}^{-1} \text{ Mpc}^{-1}$ with $h = 0.6774$, cosmological constant $\Omega_{\Lambda,0} = 0.6911$, matter density $\Omega_{M,0} = 0.3089$, baryon density $\Omega_{B,0} = 0.0486$, and initial power spectrum with normalization $\sigma_8 = 0.8159$ and index $n_s = 0.9667$. The simulated physical processes for galaxy formation include gas cooling, star formation, stellar feedback, metal enrichment, black hole feedback, and so on. The details can be found in the two method papers, Weinberger et al. (2017) and Pillepich et al. (2018a). A total of 100 snapshots are saved for each of the simulation runs. Haloes are identified with the friends-of-friends (FoF) algorithm (Davis et al. 1985) and subhaloes are identified with the SUBFIND algorithm (Springel et al. 2001; Dolag et al. 2009). Subhalo merger trees are constructed by the SUBLINK algorithm (Rodríguez-Gomez et al. 2015). To achieve a balance between sample size and resolution, we choose to use the TNG100-1 run (hereafter TNG), which has a box size with co-moving volume $(106.5 \text{ Mpc})^3$, 2×1820^3 resolution units, a target baryon mass resolution of $1.4 \times 10^6 M_\odot$, and dark matter particle mass of $7.5 \times 10^6 M_\odot$.

The EAGLE project (Schaye et al. 2014; Crain et al. 2015; McAlpine et al. 2016; The EAGLE team 2017) consists of a suite of cosmological hydrodynamic simulations performed with the GADGET-3 tree-SPH code, which is an extension of the GADGET-2 code (Springel 2005). The cosmological parameters are taken from the Planck 2013 results (Planck Collaboration I 2014): $h = 0.6777$, $\Omega_{\Lambda,0} = 0.693$, $\Omega_{M,0} = 0.307$, $\Omega_{B,0} = 0.04825$, $\sigma_8 = 0.8288$, and $n_s = 0.9611$. The subgrid processes simulated include gas cooling and heating, star formation, stellar evolution, metal enrichment, stellar feedback, and black hole feedback. A total of 29 snapshots are saved for each of the runs. Haloes and subhaloes are also identified with

FoF and SUBFIND algorithms. Subhalo merger trees are constructed using the D-TREES algorithm (Jiang et al. 2014). To achieve a balance between sample size and resolution, we choose to use EAGLE Ref-L0100N1504 (thereafter EAGLE), which has a box size with a co-moving volume of $(100 \text{ Mpc})^3$, a total number of particles of 2×1504^3 , initial baryonic particle mass of $1.81 \times 10^6 M_\odot$, and dark matter particle mass of $9.70 \times 10^6 M_\odot$.

The output galaxy and halo catalogues in TNG and EAGLE present a variety of quantities, such as halo mass, galaxy stellar mass, and SFR. Using the method described by Rodriguez-Gomez et al. (2015), we construct merger trees for FoF haloes from the subhalo merger trees. Whenever a needed galaxy or halo property is not included in the public catalogue, we calculate it using the particle/cell data. The FoF halo and subhalo properties used in our analysis are listed below.

(i) M_{halo} : ‘top-hat’ mass of the FoF halo within a radius where the overdensity is that given by the spherical collapse model (Bryan & Norman 1998).

(ii) M_{subs} : for a central subhalo (defined as the most massive subhalo in TNG, and the most bound subhalo in EAGLE), it is the total mass bounded to all subhaloes in an FoF halo; for a satellite subhalo, it is the mass bounded to the subhalo itself.

(iii) v_{max} : peak circular velocity of a subhalo, $\sqrt{GM(<r)/r}$, where $M(<r)$ is the total mass within a radius r .

(iv) z_{imm} : the redshift of last major merger of an FoF halo, where a major merger is defined as a merger event with the mass ratio between the small and large progenitors larger than one-third.

(v) z_{infall} : last in-fall redshift of an FoF halo, defined as the lowest redshift at which a progenitor of the central subhalo of the subhalo merger tree is not the most massive subhalo in the hosting FOF halo.

(vi) $z_{\text{mb}, 1/2}$: the highest redshift at which the main-branch progenitor of an FoF halo in the FoF halo merger tree assembled half of its final mass M_{halo} .

(vii) $z_{\text{mb}, \text{core}}$: the highest redshift at which the main-branch progenitor of an FoF halo in the FoF halo merger tree reached a fixed mass $M_{\text{h}, \text{core}} = 10^{11.5} h^{-1} M_\odot$.

(viii) c : the concentration parameter of the Navarro–Frenk–White (NFW) profile (Navarro et al. 1997) of an FoF halo.

(ix) q_{axis} : the shape parameter, $(a_2 + a_3)/(2a_1)$, of an FoF halo, where $a_1 \geq a_2 \geq a_3$ are the lengths of the three axes of the inertia ellipsoid. Only particles within $2.5r_s$ are used, where r_s is the characteristic radius of the NFW profile.

(x) λ_s : the dimensionless spin parameter of an FoF halo, defined as

$$\lambda_s = \frac{\|\mathbf{j}\|}{\sqrt{2}M_{\text{halo}}R_{\text{vir}}V_{\text{vir}}}, \quad (1)$$

where \mathbf{j} , R_{vir} , and V_{vir} are the total angular momentum, virial radius, and virial velocity, respectively. Only particles within $2.5r_s$ are used.

(xi) $\langle \dot{M}_{\text{halo}} \rangle$: FoF halo accretion rate, defined as

$$\langle \dot{M}_{\text{halo}} \rangle = \left\langle \frac{dM_{\text{subs}}}{dt} \right\rangle_{\text{dyn}} - 4\pi R_{\text{vir}}^2 \rho(R_{\text{vir}}) \left\langle \frac{dR_{\text{vir}}}{dt} \right\rangle_{\text{dyn}}, \quad (2)$$

where $(dx/dt)_{\text{dyn}} = [x(t) - x(t - t_{\text{dyn}})]/t_{\text{dyn}}$ is the average growth rate of a quantity x . This rate is defined using the main branch of the subhalo merger tree rooted in the central subhalo, and the dynamical time is for the FoF halo, $t_{\text{dyn}} = \sqrt{R_{\text{vir}}^3/(GM_{\text{subs}})}$.

(xii) d_{ngb} : distance of a halo to its nearest FoF halo whose M_{halo} is larger than that of the halo in consideration.

Both M_{halo} and v_{max} are provided by the TNG and EAGLE halo catalogues. We refer the reader to Chen et al. (2020) for detailed definitions and physical meanings of z_{imm} , $z_{\text{mb}, 1/2}$, $z_{\text{mb}, \text{core}}$, c , q_{axis}

and λ_s . The details of $\langle \dot{M}_{\text{halo}} \rangle$ can be found in Moster et al. (2018). We use the following quantity as our time (redshift) variable: $\delta_c(z) \equiv \delta_{c,0}/D(z)$, where $\delta_{c,0} = 1.686$ is the critical overdensity for spherical collapse, and $D(z)$ is the linear growth factor at z . We use the transfer function given by Eisenstein & Hu (1998), and the linear growth factor given by Carroll, Press & Turner (1992).

The galaxy properties used for our analysis are the following:

(i) M_* : the stellar mass of a galaxy, which is the sum of the masses of stellar particles within a radius that is two times the stellar half mass radius for TNG, or within 30 physical kpc for EAGLE.

(ii) SFR: the SFR of a galaxy, defined as the sum of the SFR of gas cells/particles within the same radius as that used for M_* .

(iii) $M_{*, \text{int}}$: the stellar mass that has ever formed in the history, i.e. $\sum_n \text{SFR}_n \Delta t_n$, where SFR_n and Δt_n are the SFR at the n th snapshot and the time interval spanned by this snapshot, respectively. So defined, $M_{*, \text{int}}$ is different from M_* in that the mass-loss due to stellar evolution and mass change due to merger are not taken into account. However, if a merger event triggers a change in the *in situ* SFR, its effect is indirectly contained in $M_{*, \text{int}}$.

(iv) sSFR: the specific SFR, defined as SFR/M_* .

Due to the limited resolution of the simulations, the SFRs have large fluctuations among different snapshots. To make the results more stable, whenever necessary we smooth the data by averaging the SFRs in adjacent snapshots. We use five adjacent snapshots for the smoothing for TNG and two for EAGLE. The resulting SFR and sSFR are referred to as the smoothed SFR and sSFR, respectively.

The galaxy–halo connection is expected to depend not only on the current status of galaxies and haloes, but also on their histories. We therefore define a number of ‘history’ quantities to describe the formation histories of galaxies and haloes. The halo AH (or mass assembly history, MAH) of a subhalo is defined as the set of v_{max} values (a vector) in the main branch of the subhalo merger tree rooted in the subhalo in question. Such a set is denoted as $\mathbf{v}_{\text{max}}^1$, and has a dimension the same as the number of snapshots spanned by the merger tree. The galaxy SFH describes the amount of star formation in its history. As we are interested in both the SFR and the cumulative quantities, M_* and $M_{*, \text{int}}$, the SFH of a galaxy (or of a hosting subhalo) may refer to the set of values for SFR, or M_* , or $M_{*, \text{int}}$, along the main branch of the subhalo merger tree, depending on the context. We denote the SFH described by these three quantities as **SFR**, \mathbf{M}_* , and $\mathbf{M}_{*, \text{int}}$, respectively, which are vectors with the same dimension as \mathbf{v}_{max} . To avoid ambiguity, we refer \mathbf{v}_{max} , **SFR**, \mathbf{M}_* , and $\mathbf{M}_{*, \text{int}}$ as the v_{max} history, the SFR history, the M_* history, and the $M_{*, \text{int}}$ history, respectively.

All of the four history vectors are in the space of a too high dimension to be useful. Here, we apply the same dimension reduction technique as used in Chen et al. (2020) to reduce the dimension of the history quantities. We provide a brief description of this method and its performance in Appendix A. After such dimension reduction, each of these histories becomes a set of principal components (PCs), which we denote as $\mathbf{PC} = (\text{PC}_1, \text{PC}_2, \dots)$, with a subscript to distinguish different physical quantities. The same technique was used in Chaves-Montero & Hearin (2020) to find the principal direction of galaxy distribution in

¹To avoid confusion, we use ‘log’ to denote 10-based logarithm, bold-roman characters to denote vectors. We use 1σ , 2σ , and 3σ regions to denote those covering 68 per cent, 95 per cent, and 99.7 per cent data points, respectively, in the space of any dimension.

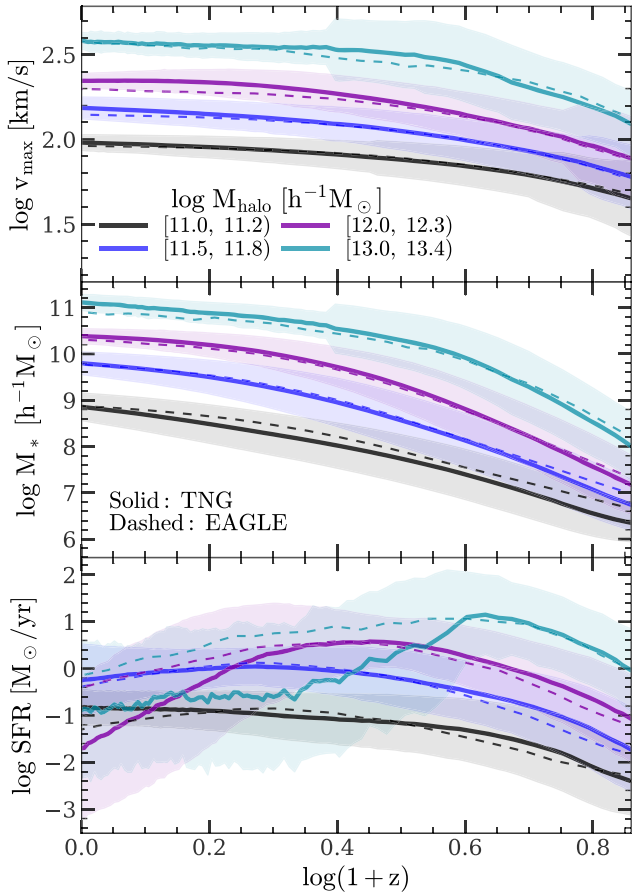


Figure 1. Halo mass assembly histories (MAHs) of central galaxies, characterized by v_{\max} (top), M_* (middle), and SFR (bottom) as functions of $\log(1+z)$. In each panel, results are shown for four different ranges of final ($z=0$) halo masses, and separately for TNG and EAGLE in the solid and dashed lines, respectively. The 1σ scatter is shown for the TNG simulation only.

the colour space and to relate the principal colour component to the SFH.

Fig. 1 shows the v_{\max} history, the SFR history, and the M_* history for central subhaloes of different masses obtained from TNG and EAGLE. Despite the difference between the two simulations, some common patterns do exist. First, the histories of M_* are very similar to those of v_{\max} , both increasing with cosmic time, but the increase being slower at lower redshift. Secondly, the galaxy with a higher v_{\max} also has a higher M_* on average. Thirdly, the SFR increases with time at high redshift but decreases at low redshift. This can also be seen from the fast-to-slow increase of M_* with time and indicates that many of the galaxies become quenched at low redshift. All these suggest that the galaxy SFH is tightly correlated with halo AH, as we will quantify in the following sections.

2.2 The galaxy samples

In this paper, we focus on the formation of central galaxies at $z=0$. We thus select all central galaxies (the ones hosted by central subhaloes) in TNG and EAGLE. The grey shade in Fig. 2 shows the galaxy distribution in the $(\log M_*, \log \text{sSFR})$ plane, where sSFR is the smoothed sSFR (see Section 2.1). It is clear that there are two distinct populations: a star-forming main sequence in which the

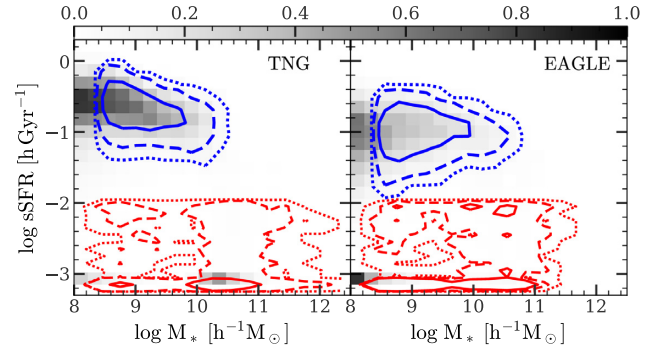


Figure 2. Distribution of central galaxies at $z=0$ in the plane of $\log \text{sSFR}$ versus $\log M_*$, for TNG (left) and EAGLE (right). The grey shade shows the normalized distribution of the full sample. Galaxies above $10^{8.5} h^{-1} M_\odot$ are divided into the star-forming sample, S'_{SF} (the blue contours), and the quenched sample, S_{Q} (the red contours). The solid, dashed, and dotted contours enclose 1σ , 2σ and 3σ regions, respectively. Galaxies with $\text{sSFR} < 10^{-3} h \text{ Gyr}^{-1}$ are stacked at the bottom of the panel. See Section 2.2 and Table 1 for the definitions of different samples.

sSFR is high and almost independent of M_* ; a quenched population with low sSFR for which the star formation activity may not even be resolved by the simulations. Since the quenching of galaxies in star formation is expected to be regulated by feedback processes, the presence of the two distinct populations indicates that the physical processes operating in them are different. Motivated by this, we separate galaxies into two samples as specified below.

- (i) The star-forming sample S_{SF} . This sample includes galaxies at $z=0$ with $M_* \geq 10^{8.5} h^{-1} M_\odot$ and the smoothed $\text{sSFR} \geq 10^{-2} h \text{ Gyr}^{-1}$, but with all galaxies that lie outside the 90 percent contour of the distribution in the $(\log M_*, \log \text{sSFR})$ plane eliminated.
- (ii) The quenched sample S_{Q} . This includes all galaxies at $z=0$ with $M_* \geq 10^{8.5} h^{-1} M_\odot$ and $\text{sSFR} < 10^{-2} h \text{ Gyr}^{-1}$.

In some of the following analysis, where a complete sample is needed, we use a sample S'_{SF} , which includes all $z=0$ central galaxies with $M_* \geq 10^{8.5} h^{-1} M_\odot$ and $\text{sSFR} \geq 10^{-2} h \text{ Gyr}^{-1}$. We also have analysis for which the properties of galaxies at a higher redshift z_0 are needed. In such cases, we apply the same separation criteria to the galaxies at the desired redshift, and construct samples $S_{\text{SF}, z=z_0}$ and $S_{\text{Q}, z=z_0}$ accordingly. In Sections 3.2 and 3.3, we have to further divide these samples according to stellar mass. We will describe the subsamplings when they are used. We summarize the samples used in this paper in Table 1. The two $z=0$ samples defined above are shown by the blue and red contours in Fig. 2.

We checked our results by using a higher M_* limit and a different sSFR threshold for the separation for the two populations, and by excluding post-merger systems. We found that our conclusions are not sensitive to the criteria adopted.

3 RELATION BETWEEN GALAXY AND HALO PROPERTIES IN SIMULATIONS

Because of the bimodal distribution of galaxies as seen in Section 2.2, we discuss the galaxy–halo relations separately for the star-forming and quenched populations. In this section, we first discuss the relation for the star-forming main sequence. We then examine how galaxies get quenched by looking at their halo properties. Finally, we present the galaxy–halo relation for the quenched population.

Table 1. Galaxy samples used in this paper. Detailed definitions can be found in Section 2.2. All of the samples are central galaxies selected by stellar mass and smoothed sSFR from TNG and EAGLE.

Sample	Description
S_{SF}	Star-forming sample consisting of all central galaxies at $z = 0$ with $M_* \geq 10^{8.5} h^{-1} M_{\odot}$ and $\text{sSFR} \geq 10^{-2} h \text{ Gyr}^{-1}$, and with 10 per cent outliers eliminated.
S'_{SF}	The same as S_{SF} but without eliminating the outliers.
$S_{\text{SF}, z=z_0}$	The same as S_{SF} but selected at $z = z_0$.
S_{Q}	Quenched sample consisting of all central galaxies at $z = 0$ with $M_* \geq 10^{8.5} h^{-1} M_{\odot}$ and $\text{sSFR} < 10^{-2} h \text{ Gyr}^{-1}$.
$S_{\text{Q}, z=z_0}$	The same as S_{Q} but selected at $z = z_0$.

3.1 Galaxy–halo relation for the main-sequence sample

To quantify the correlation strength between halo and galaxy quantities, we use the model ensemble method GBDT to build a regressor, $y = f(\mathbf{x})$, which maps the set of halo quantities \mathbf{x} to a galaxy quantity y . Using many predictor variables available, we can build a series of regressors with an increasing number of predictors. As the number of predictors increases, the overall performance, R^2 , also increases. At each step, the amount of increase in R^2 caused by including a new variable $x \in \mathbf{x}$ can be used to judge whether this variable has any contribution to the target. When all of the predictors are included, the importance value, \mathcal{I} , output from the final regressor, can be used to judge the relative contributions from individual predictors. The details of GBDT, R^2 , and \mathcal{I} can be found in Appendix B.

Fig. 3 shows the galaxy–halo relations for the star-forming samples, S_{SF} and $S_{\text{SF}, z=2}$, in both TNG and EAGLE. Here, the smoothed SFR and the first three PCs of the v_{max} history are used (see Section 2.1 and Appendix A). The results can be summarized as follows: (i) For both low- z and high- z (the left and middle panels), both M_* and the SFR are tightly correlated with v_{max} . As shown in the right-hand panel, even only v_{max} is used as the sole predictor, the value of R^2 is still quite large (≥ 0.9 for M_* , ≥ 0.7 for SFR at $z = 0$ and ≥ 0.9 for SFR at $z = 2$). (ii) At $z = 0$, the relation between SFR and v_{max} has larger scatter than that between M_* and v_{max} . The smaller R^2 for SFR shown in the right-hand panel also confirms this. This indicates that the factors regulating the star formation activity becomes more diverse as galaxies evolve from high z to low z . (iii) For both M_* and SFR, and for both redshifts, adding more halo quantities into the predictor set does not significantly improve the regression performance R^2 . In all cases, R^2 is significantly larger than 50 per cent when only v_{max} is used. This indicates that the evolution of both M_* and SFR is dominated by v_{max} . The large contribution (\mathcal{I}) from v_{max} also validates this argument.

The tight M_* – v_{max} and SFR– v_{max} relations indicate that the star-forming main sequence is a well-defined population that is largely determined by the halo potential well represented by v_{max} . Other halo properties, such as the MAH, are only secondary factors that produce relatively small variance in the sequence. To see which halo quantities are most responsible for the variance, we first define the residual value $\Delta \log \text{sSFR}$ for the smoothed sSFR as follows:

(i) We build a GBDT regressor that maps $\log M_*$ to $\log \text{sSFR}$. The predicted value of such a regressor is denoted as $\log \text{sSFR}(\log M_*)$, which can be viewed as the mean value of $\log \text{sSFR}$ at a given stellar mass.

(ii) We subtract the $\log \text{sSFR}$ of each galaxy by the mean value at the corresponding stellar mass to get the residual, $\Delta \log \text{sSFR} = \log \text{sSFR} - \log \text{sSFR}(\log M_*)$.

We relate the residual defined this way to halo quantities, as described below.

To see the effect of any halo property, x , on the main-sequence residual, we form two subsamples for galaxies of a given stellar mass. The first one consists of the 16 per cent with the highest x , while the second consists of the 16 per cent with the lowest x . If x does have an effect on the variance of the main sequence, these two subsamples should have different mean $\Delta \log \text{sSFR}$. We do this for both the S_{SF} and $S_{\text{SF}, z=2}$ samples using $x = \text{PC}_{v_{\text{max}}, 1}$, the first PC of the v_{max} history and $x = \langle \dot{M}_{\text{halo}} \rangle$, the halo accretion rate. The mean $\Delta \log \text{sSFR}$ for the two subsamples at given stellar mass are shown in Fig. 4 in comparison with the standard deviation of $\log \text{sSFR}$ at the same stellar mass. Clearly, the means of $\Delta \log \text{sSFR}$ in the two subsamples are different, and the effect of $\text{PC}_{v_{\text{max}}, 1}$ is significant in both TNG and EAGLE. Compared to the total main-sequence scatter, the effect appears relatively small at $z = 0$ and becomes larger at higher z . Thus, using $\text{PC}_{v_{\text{max}}, 1}$ alone can only explain a small portion of the residual at $z = 0$, and a larger portion at $z = 2$. The R^2 values using only $\text{PC}_{v_{\text{max}}, 1}$ shown in the right-hand panels are significantly less than 50 per cent, confirming that the prediction power of $\text{PC}_{v_{\text{max}}, 1}$ is limited. The results also show that the effect of $\langle \dot{M}_{\text{halo}} \rangle$ is smaller than that of $\text{PC}_{v_{\text{max}}, 1}$ at both $z = 0$ and $z = 2$ in both TNG and EAGLE. This indicates that the halo accretion rate is not as relevant as $\text{PC}_{v_{\text{max}}}$ in affecting the SFR, and is not a powerful proxy to separate galaxies according to the sSFR galaxies. This is consistent with O’Donnell, Behroozi & More (2021) who used the SDSS and an empirical model to demonstrate that the halo accretion rate does not significantly correlate with the current SFR, although some simulation-based investigations reached the opposite conclusion (e.g. Wetzel & Nagai 2015).

Again, because of the large number of halo quantities and potentially complex patterns in the feature space, we use GBDTs to relate the main-sequence residual, $\Delta \log \text{sSFR}$, to halo quantities. The cumulative R^2 from each of the regressors, and the contribution from each halo property in the final regressor using all halo properties, are shown in the right-hand panel of Fig. 4. At $z = 0$, the explained variance, i.e. R^2 , is only about 10 per cent in TNG and about 30 per cent in EAGLE, even when a large set of halo properties are used. At $z = 2$, R^2 for both TNG and EAGLE are still far less than 50 per cent. These poor performances in terms of R^2 indicate that there is no dominant set of halo properties that can fully explain the variance in the main sequence. Thus, once the main trend of SFR with respect to the halo mass or to v_{max} is already taken into account, an empirical model should avoid using these halo properties to assign the SFR of a galaxy based on deterministic ranking or to directly predict the main-sequence residual. Part of the main-sequence residual has to be modelled as a random component with correct statistical properties. We will discuss how to build such a model in Section 4.1.

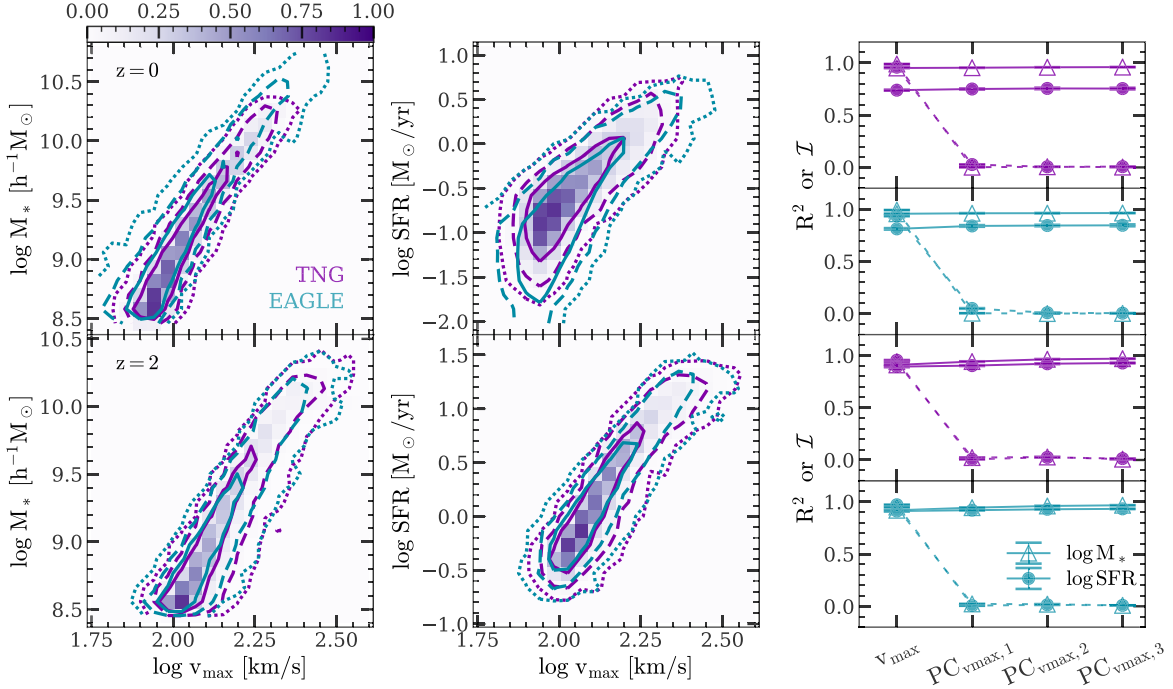


Figure 3. The halo-galaxy relations of the star-forming galaxies at two redshifts (upper row, sample S_{SF} ; lower row, sample $S_{\text{SF},z=2}$). In each row, the left-hand and middle panel show the correlation of M_* and the smoothed SFR with v_{max} . The solid, dashed, and dotted contours enclose 1σ , 2σ and 3σ regions, respectively. The purple shade represents the normalized distribution for TNG. The GBDT regression results are shown in the right-hand panels, where four halo properties are used to predict $\log M_*$ (triangles) and $\log \text{SFR}$ (circles). The solid lines are cumulative R^2 , and dashed lines are the importance \mathcal{I} of predictors in the regressor that uses the halo properties labelled along the x -axis (see Appendix B for the definitions of R^2 and \mathcal{I}). The error bars are computed by bootstrap resampling. Results from TNG and EAGLE are plotted in purple and green, respectively.

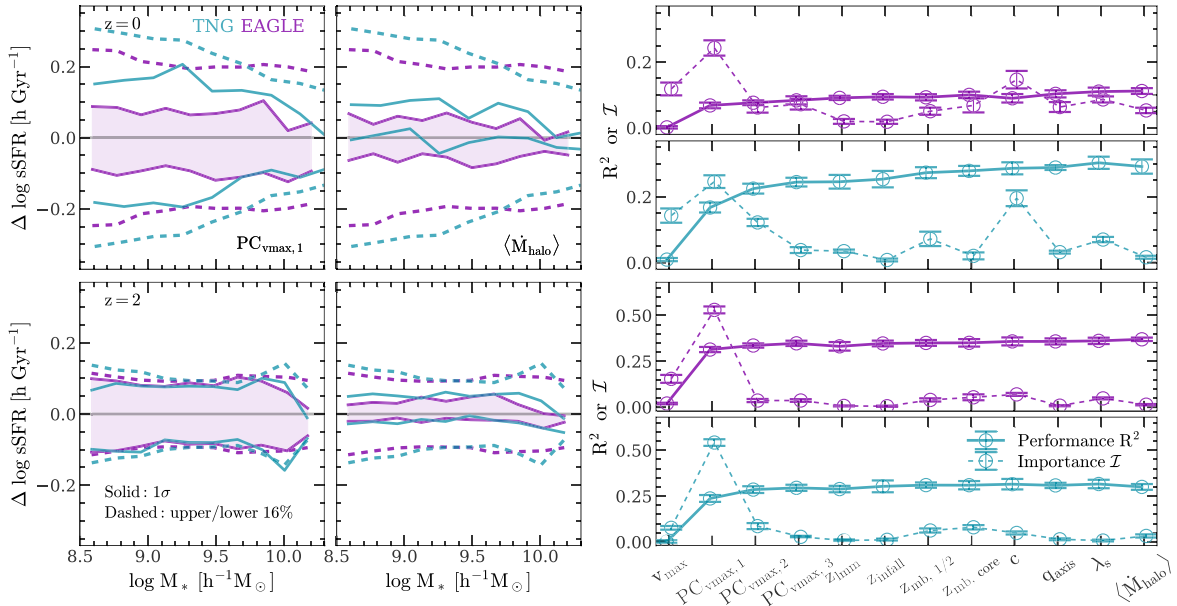


Figure 4. The two columns on the left display the sSFR residual $\Delta \log \text{sSFR}$ of star-forming galaxies as a function of $\log M_*$ for sample S_{SF} (upper row), and sample $S_{\text{SF},z=2}$ (lower row). Results for TNG and EAGLE are plotted in purple and green, respectively. In the first column, the solid lines represent the mean value of $\Delta \log \text{sSFR}$ for subsamples of galaxies whose $\text{PC}_{v_{\text{max},1}}$ are among the highest 16 per cent and the lowest 16 per cent of the full sample, respectively. The solid lines in the middle column show the results for subsamples selected by the halo accretion rate $\langle \dot{M}_{\text{halo}} \rangle$ instead of $\text{PC}_{v_{\text{max},1}}$. In both columns, the standard deviation of the full sample is plotted as the dashed lines. The right-hand panels show the results of the regression of $\Delta \log \text{sSFR}$ on halo properties (solid, cumulative R^2 ; dashed, importance \mathcal{I} of predictors in the regressor using halo properties indicated along the x -axis). See Appendix B for the definitions of R^2 and \mathcal{I} . Errors are computed by bootstrap resampling.

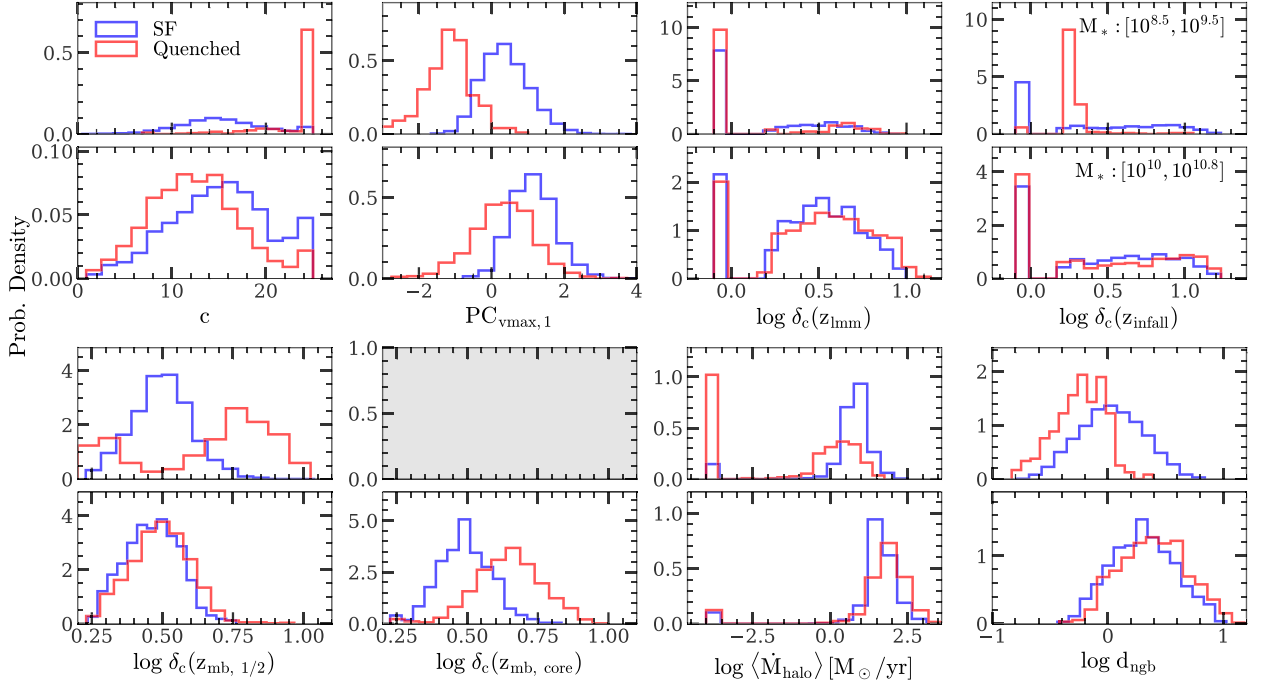


Figure 5. The distributions of halo properties for TNG galaxies with $\text{sSFR} \geq 10^{-2} \text{Gyr}^{-1} h$ (blue) and $< 10^{-2} \text{Gyr}^{-1} h$ (red). Each of the eight vertical pairs of panels shows the distribution of one halo property. In each pair, the upper panel is for low-mass galaxies and the lower is for high-mass galaxies, as indicated in the upper right pair of panels (in units of $h^{-1} M_{\odot}$). Low-mass galaxies do not have $z_{\text{mb, core}}$ measurements, and so no result is shown for this property.

3.2 Halo quantities that drive quenching

Before moving to the quenched population, let us first examine why a galaxy is quenched. To be specific, we want to see which halo quantities can be used to predict whether a galaxy is quenched or not, and whether the prediction is deterministic or stochastic. This is crucial to empirical modelling. For example, in order for a halo-based model to predict the correct bimodal distribution for galaxies, we need a careful model design so that the halo properties can really be used to distinguish between the star-forming and quenched populations. Since low-mass galaxies and massive ones may be quenched through different processes (for example, supernova feedback may be more efficient in a low-mass galaxy, while AGN feedback may be stronger in a massive galaxy that can host a more powerful central supermassive black hole), it is necessary to answer these questions separately for galaxies with different masses. We therefore define four subsamples according to both M_* and the smoothed sSFR, among all of the $z = 0$ TNG galaxies. First, we separate these galaxies into two subsamples with $10^{8.5} \leq M_*/(h^{-1} M_{\odot}) \leq 10^{9.5}$ and $10^{10} \leq M_*/(h^{-1} M_{\odot}) \leq 10^{10.8}$, respectively. We then split each of the two subsamples into two subsets according to $\text{sSFR} \geq 10^{-2} h \text{Gyr}^{-1}$ and $\text{sSFR} < 10^{-2} h \text{Gyr}^{-1}$, respectively. These four subsamples are referred as the low-mass active sample, low-mass passive sample, high-mass active sample, and high-mass passive sample, respectively. The choice of the stellar mass intervals is a compromise between minimizing the effect of M_* and preventing each subsample from being too small.

For each of these four samples, we plot the distributions of different halo quantities in Fig. 5. Here, $\log \delta_c(z_{\text{lmm}})$ for a halo without any major merger is set to be a small negative value, and the same applies to z_{infall} . The value of $\langle \dot{M}_{\text{halo}} \rangle$ is set to be $10^{-4} M_{\odot} \text{yr}^{-1}$ when the measured value is small or negative. Galaxies in the low-mass

subsamples do not have the $z_{\text{mb, core}}$ measurement, and so they do not appear in the panel of $z_{\text{mb, core}}$.

For low-mass galaxies, the active and passive populations have totally different distributions in z_{infall} . The active population has a flat z_{infall} distribution, while the passive one has a sharply peaked distribution. This difference strongly suggests that passive low-mass galaxies have undergone a very recent infall-ejection process, while high-mass galaxies do not. The distributions of other halo properties confirm this. For example, passive galaxies on average have smaller d_{ngb} , consistent with the fact that the distance of such a galaxy to a massive halo must be small for the infall event to occur. Due to interactions in the infall-ejection process, the MAH of the halo can change significantly, which may change the distributions in $\text{PC}_{\text{vmax},1}$, $z_{\text{mb}, 1/2}$, and $\langle \dot{M}_{\text{halo}} \rangle$. The halo density profile may also change in this process, which explains why the distribution of c for passive galaxies is also distinct from that of the star-forming population.

Although the distributions of halo properties for the star-forming and passive populations are significantly different, it is still challenging to design an ideal classifier to tell whether a low-mass galaxy is quenched or not using halo properties alone. The problem lies in the sample imbalance: the fraction of the passive population among all low-mass galaxies is less than 3 per cent in TNG, and less than 7 per cent in EAGLE. No matter how the classification boundary is drawn, there is always a large contamination in the population classified as the quenched population by star-forming galaxies.

The situation for high-mass galaxies is more complicated. Among all of the halo properties shown in Fig. 5, the only three which show large differences between star-forming and passive populations are $z_{\text{mb, core}}$, $\text{PC}_{\text{vmax},1}$, and c . Because haloes with mass larger than $M_{\text{h, core}}$ may likely contain bright AGNs to quench star formation and be more dominated by hot model accretion, an earlier formation of a large progenitor, i.e. a higher $z_{\text{mb, core}}$, may be indicative of a

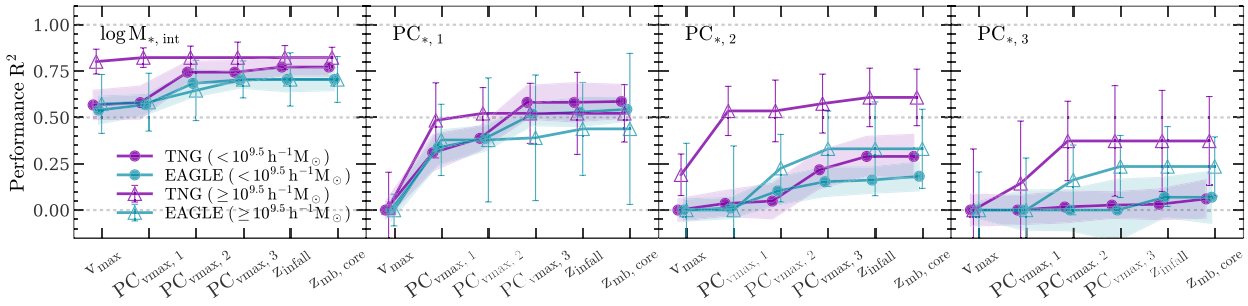


Figure 6. Cumulative R^2 of regressions of different stellar properties on halo properties for the quenched galaxies (sample S_Q). In each panel, one galaxy property (as indicated at the upper left corner) is regressed on six halo properties for both TNG (purple) and EAGLE (green) and for both low-mass (circles) and high-mass (triangles) galaxies. The stellar mass ranges (M_*) are indicated in the first panel. See Appendix B for the definitions of R^2 and \mathcal{I} . The error bars and the shaded regions are the errors estimated from bootstrap resampling.

higher probability for the galaxy to quench. Indeed, we can see this in the distribution of $z_{\text{mb,core}}$, and, implicitly, in the distributions of $\text{PC}_{v_{\text{max}},1}$. Compared with the quenched galaxies, the host haloes of star-forming galaxies are more concentrated because haloes with smaller $z_{\text{mb,core}}$ are less massive (Li et al. 2008) and therefore more concentrated. However, the distribution in $z_{\text{mb,core}}$, $\text{PC}_{v_{\text{max}},1}$, or c has significant overlap between the star-forming and quenched populations. Thus, even though the star-forming and passive samples are more balanced for massive galaxies than for low-mass ones, it is still difficult to distinguish the two populations for individual galaxies on the basis of the properties of their host haloes. We also try to distinguish the star-forming and quenched populations by building GBDTs and using the combination of multiple halo quantities as features, including $M_{\text{mb,core}}$, $\text{PC}_{v_{\text{max}},1}$, and c , with their effects shown in Fig. 5, and M_{halo} and v_{max} . Because of the degeneracy between halo properties, we find that including all these features makes no obvious improvement over using only $M_{\text{mb,core}}$. These indicate again that a halo-based empirical model may not be able to predict galaxy quenching per individual halo. What we can do is to build a model capable of correctly predicting the statistical properties of galaxies for a large ensemble of haloes.

3.3 The relations in the quenched sample

Once a galaxy is quenched, its SFR is lower and may even become too low to be resolved by the simulations. However, as we see from Fig. 1, a quenched galaxy may have a high SFR in the past when it was still in the main sequence. The tight main-sequence relation seen in Section 3.1 therefore indicates that the SFH of a galaxy may be related to the MAH of its host halo. Motivated by this, we use the integrated stellar mass, $M_{*,\text{int}}$, to represent the galaxy SFH. Compared to the current stellar mass, $M_{*,\text{int}}$ at any given redshift can be viewed as the stellar mass that has formed throughout the history before the redshift (see Section 2.1 for definition). We have tested that our conclusion does not depend on this choice because almost all galaxies in samples S_{SF} and S_Q have $M_* \leq M_{*,\text{int}} \leq 2M_*$.

So defined, the $M_{*,\text{int}}$ history is a direct quantity that ‘remember’ the history of star formation of a galaxy. Therefore, the $M_{*,\text{int}}$ history should be connected to the halo MAH. Using the same PCA as used for the v_{max} history, we reduce the dimensions of the $M_{*,\text{int}}$ histories by representing them with several PCs, denoted as $\text{PC}_* = (\text{PC}_{*,1}, \text{PC}_{*,2}, \dots; \text{see Section 2.1 and Appendix A})$.

We relate $M_{*,\text{int}}$ and PC_* of the quenched sample, S_Q , to the following set of halo properties: v_{max} , the PCs of the v_{max} history, z_{infall} and $z_{\text{mb,core}}$, applying the GBDTs separately for low-mass

($M_* < 10^{9.5} h^{-1} M_{\odot}$) and high-mass ($M_* \geq 10^{9.5} h^{-1} M_{\odot}$) subsamples. The inclusion of z_{infall} and $z_{\text{mb,core}}$ is motivated by the results presented in Section 3.2, where it is shown that these two quantities are responsible for the quenching of low-mass and high-mass galaxies, respectively. The results are shown in Fig. 6. The prediction of $M_{*,\text{int}}$ has a R^2 larger than 50 per cent, even if we use only v_{max} as the predictor. This indicates that $M_{*,\text{int}}$ can be well reproduced with halo properties, and is dominated by v_{max} . The SFH PCs are harder to predict. By using all of the six halo quantities, the R^2 for each of the three SFH PCs is still far less than 100 per cent, indicating that a large portion of factors affecting the SFH are still missing in the model and that the details of how a galaxy forms may be influenced by many nuanced factors. For the $\text{PC}_{*,1}$, R^2 is significant and $\text{PC}_{v_{\text{max}},1}$ is the most important factor, as seen from the big increase of the cumulative R^2 it produces. The second and third PCs of the v_{max} history are also important for the $\text{PC}_{*,1}$ of low-mass galaxies. As shown in Section 3.2, the quenching of low-mass galaxies is mainly due to the infall process in which their SFHs have large variances and more halo PCs are needed to capture them. For the $\text{PC}_{*,2}$ and $\text{PC}_{*,3}$ of high-mass galaxies, the TNG and EAGLE show large differences. The R^2 for EAGLE is much lower, and require more halo PCs to capture the variances. For low-mass galaxies, both TNG and EAGLE have low R^2 for the predictions of $\text{PC}_{*,2}$ and $\text{PC}_{*,3}$, indicating that high-order variations in the SFH are typically more difficult to model. In all the cases, z_{infall} and $z_{\text{mb,core}}$ do not provide a significant contribution to R^2 when PCs of the v_{max} history are already used. This indicates that information carried by these two characteristic redshifts are already contained in the PCs of the v_{max} history. We will use these results to help build our empirical model, as described in Section 4.1.

4 THE EMPIRICAL MODEL OF STAR FORMATION IN DARK MATTER HALOES

The results presented in Sections 3.1 and 3.3 show that the properties of the SFH of a central galaxy are well captured by the halo properties when the galaxy is in the star-forming main sequence, and that the $M_{*,\text{int}}$ history can be well captured by halo properties even for quenched galaxies. Based on this tight galaxy–halo connection, we propose an empirical model to populate haloes with central galaxies. Because of the differences between the star-forming and quenched populations, we model them separately. In this section, we first discuss the design of the model and present a detailed description of all of the model ingredients. We then use five cases to test the model step by step.

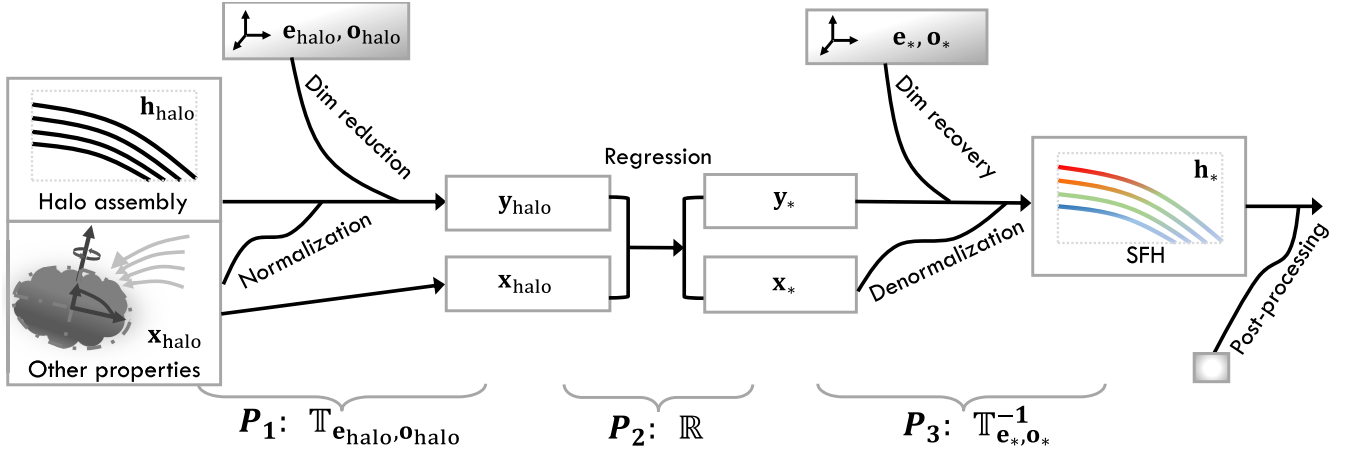


Figure 7. The outline of the empirical model for the star formation of central galaxies in dark matter haloes. The MAH, \mathbf{h}_{halo} , and other properties of haloes, \mathbf{x}_{halo} , are transformed into the galaxy star formation history, \mathbf{h}_* , through three procedures, P_1 , P_2 , and P_3 . A post-processing is performed in the end. See Section 4.1 for the detailed description of the model.

4.1 The model

The structure of the model is designed on the basis of the following considerations. (i) We favour a simple model with a small number of parameters to a complicated black-box model. A simpler model is easier to understand, and can provide more transparent insights into the relation between galaxies and dark matter haloes. In addition, a simpler model is less prone to overfitting problems. We thus choose the use of PCA to reduce the size of the parameter space for both haloes and galaxies. (ii) The model should be expressive and flexible to absorb a variety of observation constraints and to provide a wide range of outcomes to compare with future observations. Because of this, we choose to build the model in a deeper way rather than directly mapping halo properties to galaxy properties. The model should include the full pipeline of the feature extraction, the regression, and the post-processing, each of which is simple enough while the joint of them is sufficient to capture the complicated patterns in the galaxy–halo connection. (iii) To optimize such a model, no standard approach is available. Here, we choose to break the model into several pieces and optimize them stepwise. This optimization borrows the idea from the ‘greedy algorithm’ described in many textbooks of algorithm-design (e.g. Cormen et al. 2009; Sedgewick & Wayne 2011).

We outline the model in Fig. 7. The overall purpose of the model is to predict the SFH, \mathbf{h}_* , and other properties, \mathbf{x}_* , of a given central galaxy, from the MAH of its host halo, \mathbf{h}_{halo} , and a set of other halo properties, \mathbf{x}_{halo} . Here, \mathbf{x}_{halo} and \mathbf{x}_* are defined at a given redshift z_0 , and \mathbf{h}_{halo} and \mathbf{h}_* are histories defined over a redshift range between z_0 and $z_1 > z_0$. We will specify the definitions of these variables later. To achieve our goal, we break the model construction into three procedures, P_1 , P_2 , and P_3 , which are described one-by-one in the following.

In the first procedure, P_1 , we reduce the dimension of the halo MAH, \mathbf{h}_{halo} . The purpose is to make the representation of a halo simpler so that the mapping from it to galaxy properties is easier to establish. P_1 consists of the following steps.

(i) We choose $\mathbf{h}_{\text{halo}} = v_{\text{max}}$ as the halo MAH variable and use only v_{max} for \mathbf{x}_{halo} , $\mathbf{x}_{\text{halo}} = (v_{\text{max}})$. We have checked other halo properties, such as halo virial mass and mass bound to subhaloes and found that v_{max} is the best. This is consistent with the test results of subhalo abundance matching in Reddick et al. (2013), but here we extend it by including the MAH as a secondary halo property. We choose only PCs of the v_{max} history as the history variable, because we have

already seen that the v_{max} history is tightly related both to the SFR for galaxies in the main sequence, and to the history of $M_{*, \text{int}}$ even for quenched galaxies.

(ii) We normalize the halo MAH by $\tilde{\mathbf{h}}_{\text{halo}} = \mathbf{h}_{\text{halo}}/h_{\text{halo}, z=z_0}$, where $h_{\text{halo}, z=z_0}$ is the component of \mathbf{h}_{halo} that corresponds to $z = z_0$. The purpose is to prevent the dimension reduction from being too much concentrated in low redshift.

(iii) We apply the PCA to $\tilde{\mathbf{h}}_{\text{halo}}$, which gives a set of eigenvectors, $\mathbf{e}_{\text{halo}, i}$ ($i = 1, 2, 3, \dots$), and a mean offset, \mathbf{o}_{halo} (see Appendix A). After a shift of \mathbf{o}_{halo} and a projection with \mathbf{e}_{halo} , we get a new vector \mathbf{y}_{halo} , which is the set of PCs we want to obtain. Our test shows that using the first two PCs is sufficient for modelling the SFH, and that including more PCs does not lead to much gain in the model performance.

After procedure P_1 , a halo can be described by a small set of variables (\mathbf{y}_{halo} , \mathbf{x}_{halo}), which is sufficiently simple. As shown in Sections 3.1 and 3.3, this set also gives a good prediction for the SFH. We denote the total transformation in procedure P_1 as $\mathbb{T}_{\mathbf{e}_{\text{halo}}, \mathbf{o}_{\text{halo}}}$:

$$(\mathbf{y}_{\text{halo}}, \mathbf{x}_{\text{halo}}) = \mathbb{T}_{\mathbf{e}_{\text{halo}}, \mathbf{o}_{\text{halo}}}(\tilde{\mathbf{h}}_{\text{halo}}, \mathbf{x}_{\text{halo}}), \quad (3)$$

where \mathbf{x}_{halo} , which is not involved in the transformation, is included to simplify descriptions in the following. The halo properties (\mathbf{y}_{halo} , \mathbf{x}_{halo}) are then fed into procedure P_2 .

Before entering P_2 , we must decide how to represent a galaxy. One of the quantities of interest is $M_{*, \text{int}}$, and we denote the set of stellar properties as $\mathbf{x}_* = (M_{*, \text{int}})$ in this case. The SFH is a large vector, too complicated to model. It is therefore necessary to represent the SFH also by a set of PCs. For both star-forming and quenched galaxies, the SFH is well correlated with halo properties (see Sections 3.1 and 3.3), so we define the SFH as $\mathbf{h}_* = \log \mathbf{M}_{*, \text{int}}$. The normalization is performed as $\tilde{\mathbf{h}}_* = \mathbf{h}_* - h_{*, z=z_0}$. Note that we also tested using other galaxies properties to represent SFH, e.g. SFR for star-forming galaxies, but found little difference in terms of the model performance. Once $\tilde{\mathbf{h}}_*$ is obtained, we use the same method as we did for haloes to reduce the dimension of SFH into a set of PCs, \mathbf{y}_* , given by the set of eigenvectors, $\mathbf{e}_{*, i}$ ($i = 1, 2, 3, \dots$), and the mean offset, \mathbf{o}_* . The normalization and the projection into the new frame are jointly referred as the transformation $\mathbb{T}_{\mathbf{e}_*, \mathbf{o}_*}$, so that $(\mathbf{y}_*, \mathbf{x}_*) = \mathbb{T}_{\mathbf{e}_*, \mathbf{o}_*}(\tilde{\mathbf{h}}_*, \mathbf{x}_*)$. The real modelling process is actually the inverse, namely we first predict the PCs of the SFH and \mathbf{x}_* according to halo properties, and then do the reverse transformation to obtain the SFH. This is what we do in procedures P_2 and P_3 .

Table 2. Five test cases for the empirical model used in this paper. The exact definitions can be found in Section 4.2. This table lists the target galaxies we want to model and compare with the simulation, the simulation which the SFH template is taken from, and how to model the star-forming and quenched populations.

Test case	T_{TNG}	$T_{\text{EAGLE+TNG-modes}}$	T_{EAGLE}	T_{join}	T'_{join}
Target galaxies	TNG	EAGLE	EAGLE	EAGLE	EAGLE
SFH template	TNG	TNG	EAGLE	TNG	EAGLE
Treatment of bimodal populations		Separately		Jointly	

Procedure P_2 is simple: we build a regressor to predict stellar properties of a galaxy, $(\mathbf{y}_*, \mathbf{x}_*)$, according to halo properties. Denoting the regressor as \mathbb{R} , we have

$$(\mathbf{y}_*, \mathbf{x}_*) = \mathbb{R}(\mathbf{y}_{\text{halo}}, \mathbf{x}_{\text{halo}}). \quad (4)$$

Procedure P_3 is just the reverse of $\mathbb{T}_{\mathbf{e}_*, \mathbf{o}_*}$:

$$(\mathbf{h}_*, \mathbf{x}_*) = \mathbb{T}_{\mathbf{e}_*, \mathbf{o}_*}^{-1}(\mathbf{y}_*, \mathbf{x}_*), \quad (5)$$

which includes the dimension recovering and denormalization. The dimension recovering transforms \mathbf{y}_* back to $\tilde{\mathbf{h}}_*$ by the inverse of the PCA using \mathbf{e}_* and $\mathbf{o}_{*,i}$. The de-normalization transforms $\tilde{\mathbf{h}}_*$ to \mathbf{h}_* using \mathbf{x}_* . Note that \mathbf{x}_* is not changed in the transformation.

Putting all these procedures together, we have a mapping from halo properties, $(\mathbf{h}_{\text{halo}}, \mathbf{x}_{\text{halo}})$, to galaxy properties, $(\mathbf{h}_*, \mathbf{x}_*)$:

$$(\mathbf{h}_*, \mathbf{x}_*) = \mathbb{T}_{\mathbf{e}_*, \mathbf{o}_*}^{-1} \mathbb{R} \mathbb{T}_{\mathbf{e}_{\text{halo}}, \mathbf{o}_{\text{halo}}}(\mathbf{h}_{\text{halo}}, \mathbf{x}_{\text{halo}}). \quad (6)$$

Note that the model has some degrees of freedom to be fixed. The dimension reduction templates for haloes, $\mathbf{e}_{\text{halo},i}$ and \mathbf{o}_{halo} , are always known because we populate haloes in dark-matter simulations. On the other hand, the regressor, \mathbb{R} , in P_2 needs to be modelled for real applications. The template for galaxies, \mathbf{e}_* and \mathbf{o}_* , also needs to be modelled. The main advantage of our model is that, we may borrow some unknown parts from hydrodynamic simulations, so that the degrees of freedom of the model can be reduced dramatically. For example, although galaxy SFHs in different simulations may differ significantly, they may still be represented accurately by a small number of PCs with eigen-functions obtained from one set or a combination of multiple sets of simulations, thus reducing the dimension of each individual SFH from infinity (to describe a continuous history) to a small number. We will discuss the details in Section 4.2 and show the results in Section 4.3.

Since we only take $M_{*,\text{int}}$ at $z = z_0$ as the normalization for galaxy SFH, a small discrepancy in the prediction of $M_{*,\text{int}}$ may give rise to a large difference in SFH at high redshift. To make the model more precise at high redshift, we break the halo MAH and the SFH of central galaxies at $z = 0$ into two pieces: the first is between $z_0 = 0$ and $z_1 = 1.5$, and the second is above $z_0 = 1.5$. We run the model separately for these two redshift ranges, and join the modelled \mathbf{h}_* with a proper smoothing at $z \sim 1.5$. This, of course, doubles the model complexity but gives a more accurate prediction of the SFH, which may be needed when high- z data are available to constrain the model.

For any galaxy, once $M_{*,\text{int}}$ is known, we can differentiate it with respect to cosmic time to obtain **SFR**.

As shown in Section 3.1, the SFR of a galaxy cannot be totally determined by its halo properties even for galaxies in the main sequence. The residual of the main sequence is hard to predict even with the use of a large set of halo properties. So far our model has not taken the scatter in the SFR into account. To make the model for the star-forming galaxies more realistic, we add a Gaussian random component with a zero mean and a covariance Σ to the modelled $\log \text{sSFR}$. The covariance is obtained by fitting that of

the residual between the modelled $\log \text{sSFR}$ and the simulated one. The logarithm of each diagonal element of Σ is fitted by a sigmoid function σ versus $\log(1+z)$, and each off-diagonal element is fitted by a linearly decayed correlation strength versus the number of snapshots between any two elements, $i-j$:

$$\log \Sigma_{i,j} = \sigma[\log(1+z_i)]\sigma[\log(1+z_j)]\text{Lin}(i-j), \quad (7)$$

where the four free parameters in the sigmoid function and the two free parameters in the linear function are all free parameters to be determined by the fit. We found that the correlation length is always quite small, with the correlation quickly decreasing to a negligible value. We also found that the sigmoid behaviour of the variance does not depend strongly on halo mass.

All the processes after P_3 are collectively referred to as the post-processing.

4.2 Testing the model with simulations

We now apply our model to simulations and test its performance by comparing the model prediction with the simulated SFH of galaxies. We define five test cases, denoted as T_{TNG} , $T_{\text{EAGLE+TNG-modes}}$, T_{EAGLE} , T_{join} , and T'_{join} , respectively. This design allows us to test our model both in ideal cases, where all of the model ingredients are known, and in more realistic cases, where some of the model ingredients need to be modelled. We summarize the test cases in Table 2.

The test cases defined here use dark matter haloes in full hydrodynamic runs. We checked our results by matching these subhaloes with those in the corresponding dark-matter-only (DMO) runs, and rerunning our model on the DMO subhaloes. We found no obvious changes in our results, although the uncertainty in the modelled stellar properties increases moderately. This may be expected, as the halo structure on the scale relevant to our modelling (e.g. where v_{max} is defined) may not be affected significantly by the baryonic effects. We should also emphasize that we use hydro simulations to guide our model design, rather than to establish the exact mapping between haloes and galaxies.

The first test case T_{TNG} relies only on the TNG data. It is conducted separately for both the star-forming sample, $S = S_{\text{SF}}$, and the quenched sample, $S = S_{\text{Q}}$ (see Section 2.2 and Table 1 for sample definitions). To test the performance of the model, we randomly split each of the TNG samples, S , into a training set and a test set, with a ratio of 3: 1 in the number of galaxies between them. The steps are the following:

- (i) Following procedure P_1 , we apply the PCA to the histories $\tilde{\mathbf{h}}_{\text{halo}}$ of the hosting haloes of galaxies in sample S , which gives the transformation, $\mathbb{T}_{\mathbf{e}_{\text{halo}}, \mathbf{o}_{\text{halo}}}$, and the low-dimension representation of the halo MAH, \mathbf{y}_{halo} .
- (ii) We apply the PCA to the SFH, $\tilde{\mathbf{h}}_*$, of the galaxies in sample S , which gives the transformation, $\mathbb{T}_{\mathbf{e}_*, \mathbf{o}_*}$, and the low-dimension representation of the SFH, \mathbf{y}_* .
- (iii) Using the training set, we train the GBDT regressor, \mathbb{R} , which maps $(\mathbf{y}_{\text{halo}}, \mathbf{x}_{\text{halo}})$ into $(\mathbf{y}_*, \mathbf{x}_*)$.

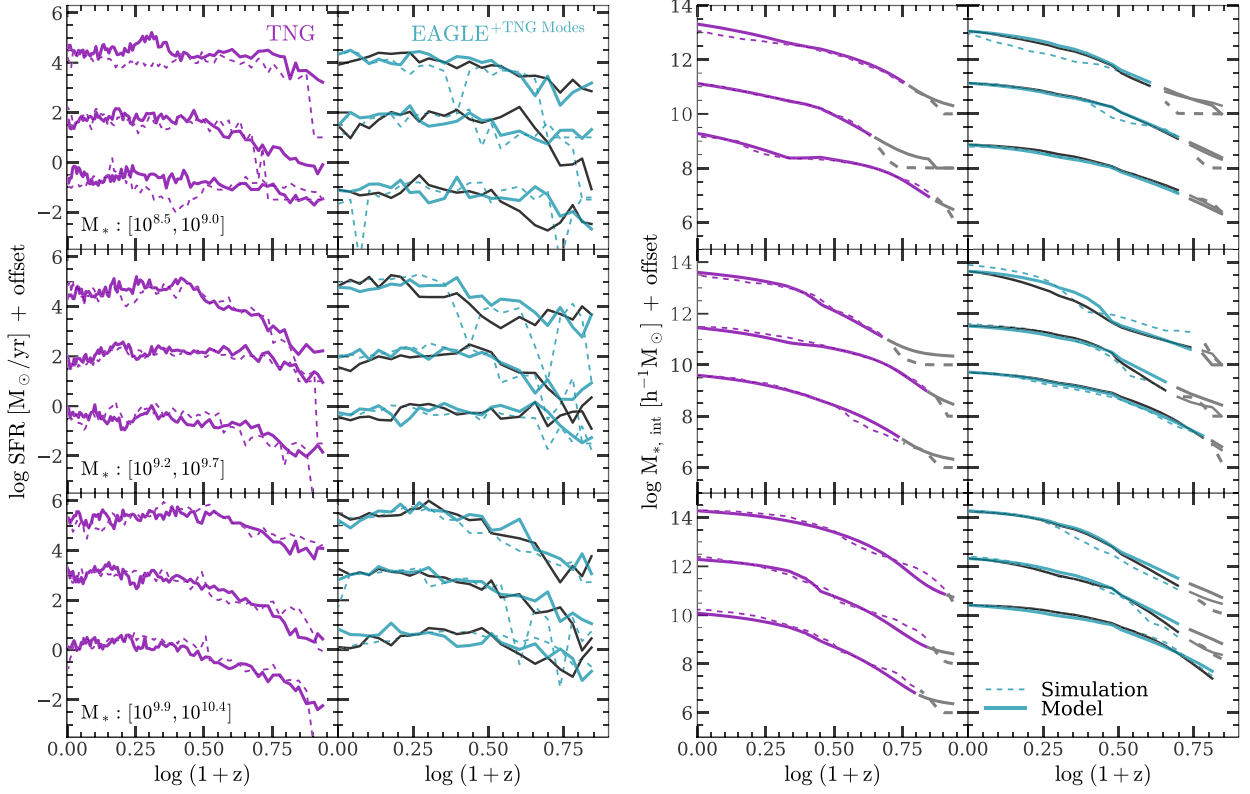


Figure 8. The modelled SFR and $M_{*,\text{int}}$ histories of star-forming galaxies at $z = 0$ (sample S_{SF}) compared with the simulation results for three test cases, T_{TNG} , $T_{\text{EAGLE+TNG-modes}}$, and T_{EAGLE} , and for three different stellar mass ranges as indicated in the leftmost panels (in units of $h^{-1}M_{\odot}$). In each panel, solid lines are from our model, while dashed lines are from simulations. Purple lines are for the test case T_{TNG} , green lines for $T_{\text{EAGLE+TNG-modes}}$, and black lines, plotted together with the green lines, for T_{EAGLE} . In each case, results are shown for three example galaxies, with arbitrary offsets vertically for clarity. Gray lines in the panels of $M_{*,\text{int}}$ histories show the parts of the histories where the stellar mass is below $10^7 h^{-1}M_{\odot}$ and which may not be well resolved in the simulations.

(iv) We apply the transformations obtained above to map halo properties to galaxy properties in the test set using $(\mathbf{h}_*, \mathbf{x}_*) = \mathbb{T}_{\mathbf{e}_*, \mathbf{o}_*}^{-1} \mathbb{R} \mathbb{T}_{\mathbf{e}_{\text{halo}}, \mathbf{o}_{\text{halo}}}(\mathbf{h}_{\text{halo}}, \mathbf{x}_{\text{halo}})$, and perform the post-processing.

After these steps, we obtain the modelled SFR and $M_{*,\text{int}}$ histories for the galaxies in the test set, and we compare them with the TNG data. Because the separation of the star-forming and quenched galaxies and all of the transformations are obtained directly from the simulation data, the performance of this test case can be viewed as the upper limit of our model. In this case, the deviation of the model output from the simulation is due to the intrinsic incapability of the model, which, in principle, can be improved by including more halo properties into \mathbf{x}_{halo} and using more PCs of \mathbf{h}_{halo} , provided that the training set is sufficiently large.

The second test case $T_{\text{EAGLE+TNG-modes}}$ relies both on the TNG and EAGLE and is designed to mimic the situation in real applications where some of the model ingredients are unknown. The test is also made for both the star-forming sample $S = S_{\text{SF}}$ and the quenched sample $S = S_{\text{Q}}$ in EAGLE (see Section 2.2 and Table 1). To test the model performance, we randomly split each of the EAGLE sample, S , into a training set and a test set, again with a 3:1 ratio in the number of galaxies between the two sets. The test is conducted through the following steps:

(i) In a real application, halo information is accessible. Therefore, we directly apply the PCA to the histories, $\tilde{\mathbf{h}}_{\text{halo}}$, of EAGLE haloes in sample S , which gives us the transformation, $\mathbb{T}_{\mathbf{e}_{\text{halo}}, \mathbf{o}_{\text{halo}}}$ and the PCs describing the halo MAH.

(ii) SFH is not accessible because it is the target of the model. This prevents us from getting a dimension reduction template $(\mathbf{e}_*, \mathbf{o}_*)$. Thus, some assumptions have to be made. We choose to use the eigenvectors, \mathbf{e}_* , that are built from the TNG in T_{TNG} , and we interpolate each of these eigenvectors to the redshifts of EAGLE’s snapshots. In doing so, we in effect borrow the template from the TNG for the analysis of EAGLE. As we will show later, using the TNG template to reduce the dimension of EAGLE SFH is, in terms of model performances, comparable to using the template from EAGLE itself. Thus, only \mathbf{o}_* remains to be modelled in real applications, and it can be modelled by using some parametric form to be constrained by observations. As more observations are added, the estimate of \mathbf{o}_* will be improved. Here, we want to test the upper performance limit of our model by using the real \mathbf{o}_* obtained directly from sample S of EAGLE, and we denote it by $\tilde{\mathbf{o}}_*$. Finally we obtain the transformation, $\mathbb{T}_{\mathbf{e}_*, \tilde{\mathbf{o}}_*}$.

(iii) In a real application, the mapping, \mathbb{R} , also needs to be modelled and constrained by observations. Again, because we want to gauge the upper limit of the model performance, we train \mathbb{R} by halo properties, $\mathbb{T}_{\mathbf{e}_{\text{halo}}, \mathbf{o}_{\text{halo}}}(\mathbf{h}_{\text{halo}}, \mathbf{x}_{\text{halo}})$, and galaxy properties, $\mathbb{T}_{\mathbf{e}_*, \tilde{\mathbf{o}}_*}(\mathbf{h}_*, \mathbf{x}_*)$, both from the training set of EAGLE. The trained regressor is denoted by $\tilde{\mathbb{R}}$.

(iv) We apply the transformation $\mathbb{T}_{\mathbf{e}_*, \tilde{\mathbf{o}}_*}^{-1} \tilde{\mathbb{R}} \mathbb{T}_{\mathbf{e}_{\text{halo}}, \mathbf{o}_{\text{halo}}}$ to the host haloes of the test galaxies, and perform the post-processing to get the final output.

In the end of all these steps, we obtain the model predictions for galaxy properties and compare them with the results of EAGLE.

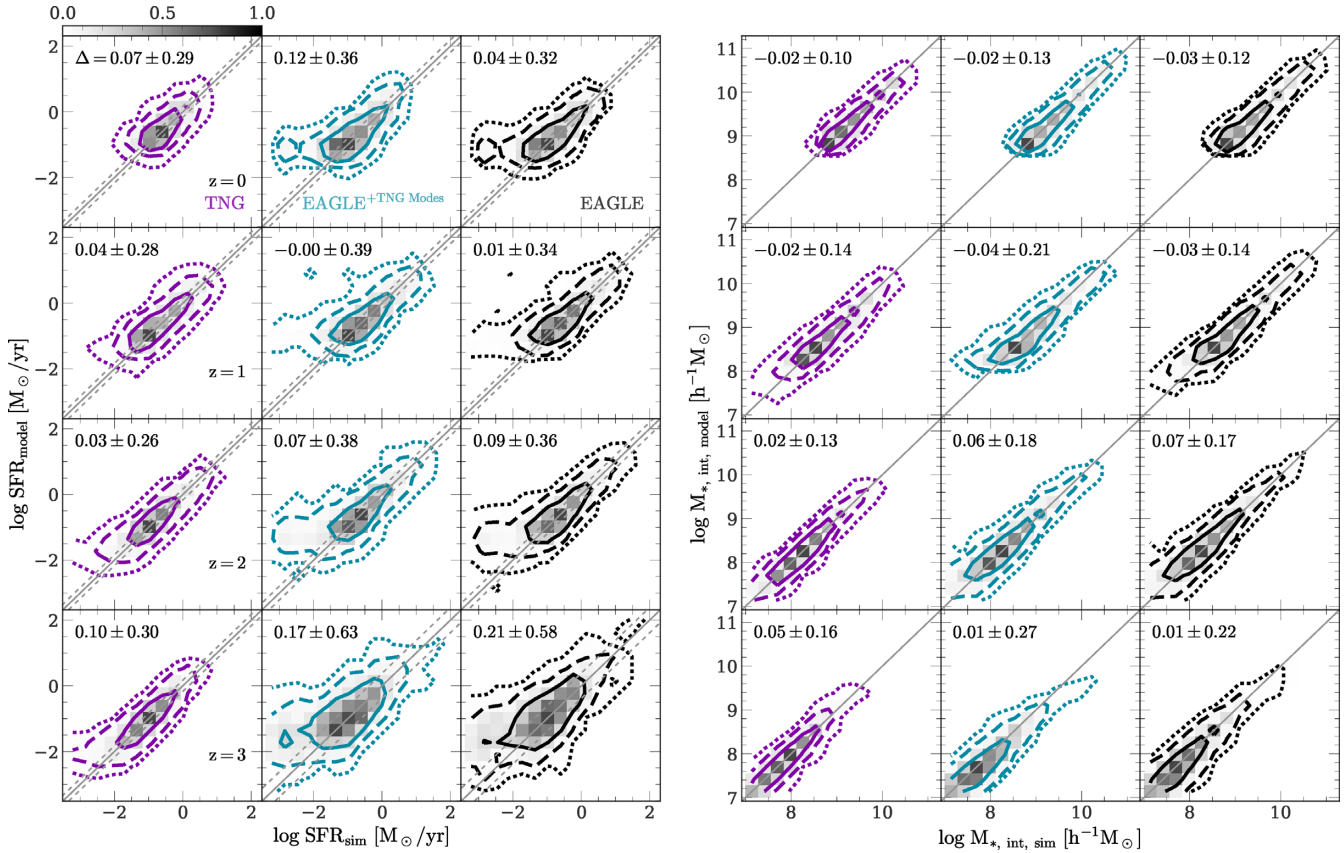


Figure 9. The modelled SFR (the left three columns) and $M_{*, \text{int}}$ (the right three columns) in the histories of $z = 0$ star-forming galaxies (sample S_{SF}) in comparison with the simulation results for three test cases, T_{TNG} , $T_{\text{EAGLE+TNG-modes}}$, and T_{EAGLE} , and for different redshifts as indicated in the left-hand panels. The M_* -weighted average of residual and standard deviation (Δ) are indicated in the upper left corner of each panel. The modelled SFR includes the randomly added noise whose standard deviation is shown by the dashed grey curves in each of the SFR panels. Solid, dashed and dotted contours enclose 1σ , 2σ , and 3σ regions, respectively.

Since in $T_{\text{EAGLE+TNG-modes}}$ some of the transformation ingredients are borrowed from TNG, the model performance is inevitably worse than that using the true transformation. To see the effect caused by the imperfect transformation, we design a third test case, T_{EAGLE} , which is identical to T_{TNG} , except that both S_{SF} and S_{Q} are taken from EAGLE.

Finally, we design a more realistic testing case, T_{join} , in which the separation of star-forming and quenched galaxies is also to be modelled. This test is conducted for both the star-forming sample $S = S'_{\text{SF}}$ and the quenched sample $S = S_{\text{Q}}$ in EAGLE. We again use a 3:1 split between the training and test sets. The testing steps are the following:

- (i) We apply the same modelling as in $T_{\text{EAGLE+TNG-modes}}$ to S'_{SF} and S_{Q} , and obtain two models that map halo properties to the galaxy SFH separately for star-forming and quenched galaxies.
- (ii) Using the combination of the training set in S'_{SF} and S_{Q} , we train a GBDT classifier which classifies a $z = 0$ galaxy into the star-forming or the quenched population according to its halo properties, v_{max} at $z = 0$, z_{infall} , $z_{\text{mb, core}}$, and the first three PCs of the v_{max} history. The inclusion of z_{infall} and $z_{\text{mb, core}}$ is motivated by the fact that these two properties are important in affecting galaxy quenching (see Section 3.2).
- (iii) We apply the classifier to the combination of the test sets in both S'_{SF} and S_{Q} . A galaxy is then classified either as star-forming

or quenched. We apply the two trained models to star-forming and quenched populations, respectively.

As mentioned in Section 3.2, the separation of star-forming and quenched galaxies is far from perfect, which can lead to significant contamination in both the star-forming and quenched samples classified. This limits the performance of models based on halo properties. However, as we will show later, although the reconstruction of the SFH for individual galaxies is contaminated by imperfect classification, the statistical properties of the whole population are unbiased. The final outputs of the two models in this test case consist of properties of both star-forming and quenched galaxies at $z = 0$, and are compared to the EAGLE data. As described above, the separation of star-forming and quenched galaxies, as well as the transformation, all mimic real applications in this case.

We again want to see if the use of the template from EAGLE itself can make an improvement in the model performance. To this end, we define a fifth test case, T'_{join} , which is identical to T_{join} , except that the dimension reduction template is from EAGLE itself in its first step.

4.3 The results

We now show the results of the five test cases, T_{TNG} , $T_{\text{EAGLE+TNG-modes}}$, T_{EAGLE} , T_{join} , T'_{join} . In the first three cases, the star-forming sample S_{SF} and the quenched sample S_{Q} are modelled separately. The

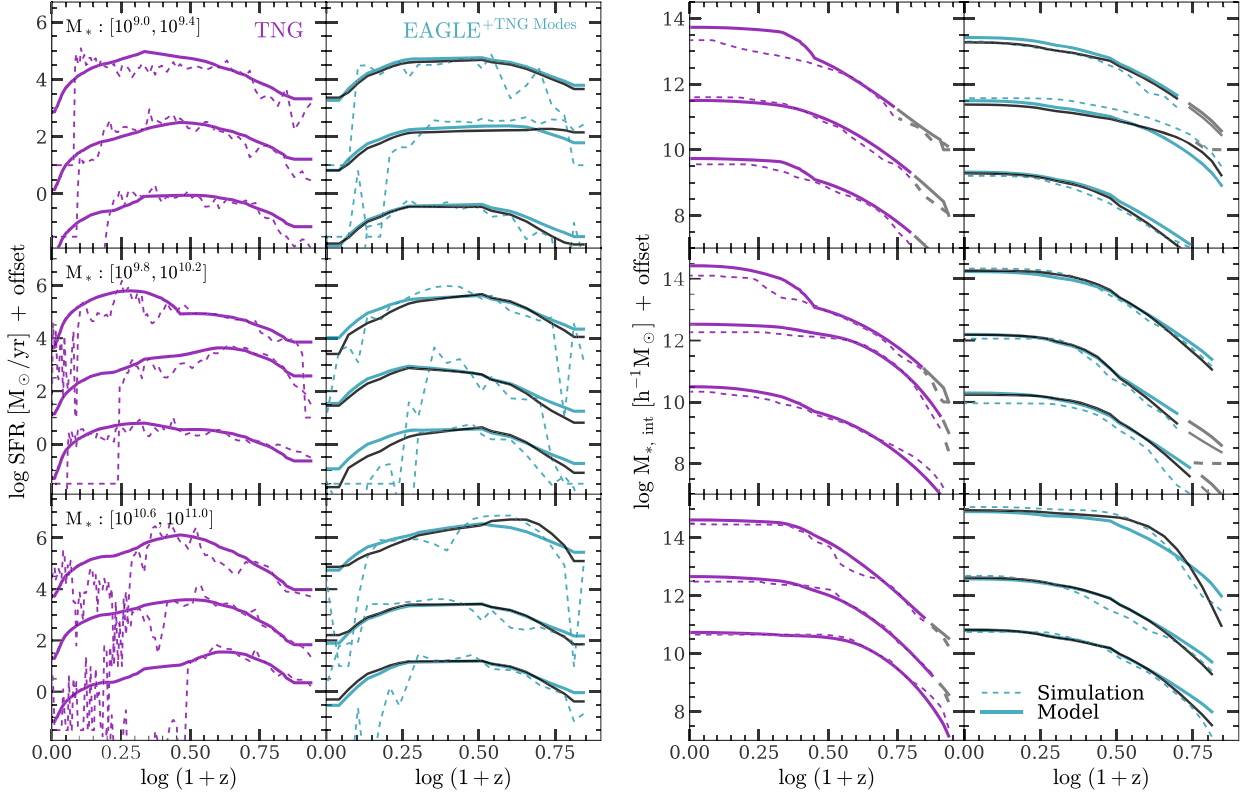


Figure 10. The same as Fig. 8, but for the quenched galaxy sample S_Q .

modelled SFHs of star-forming galaxies, represented by SFR and $M_{*,\text{int}}$ at each snapshot, are shown in Fig. 8 in comparison with the simulated one. The simulated SFR histories of individual galaxies show small fluctuations on small time-scale, which are not captured well by the PCs in the model, but can be modelled by adding a random component in the post-processing. In all of the three test cases, the model successfully reproduces the overall trend of the simulated SFR histories for individual galaxies. The difference between the model and the simulation is small at low z , and becomes slightly larger at higher z where the SFR becomes too low to model accurately. The modelled SFR histories in $T_{\text{EAGLE+TNG-modes}}$ is as good as those in T_{TNG} and T_{EAGLE} at low redshift, and becomes slightly worse at high z in some cases. The $M_{*,\text{int}}$ histories are more smooth, but the overall conclusion for the SFR histories also holds for the $M_{*,\text{int}}$ histories. All of these indicate that the model can reproduce both the SFR and $M_{*,\text{int}}$ histories for star-forming galaxies.

To quantify the goodness of the model in describing the data of star-forming galaxies, we compare in Fig. 9 the simulated SFR and $M_{*,\text{int}}$ at several redshifts from 0 to 3. We also compute the mean and standard deviation of the M_* -weighted average of residual between the simulation and the model. The results can be summarized as follows: (i) The residual between modelled and simulated log SFR and log $M_{*,\text{int}}$ has no obvious bias at all redshifts, (ii) The residual between modelled and simulated SFR and $M_{*,\text{int}}$ slightly increases with redshift. The scatter is about 0.3 dex for the SFR and 0.1 dex for the $M_{*,\text{int}}$ at $z = 0 \sim 2$ and increases at $z > 2$, (iii) The random noise, which cannot be modelled by the halo MAH, is moderate at low z , and becomes significant at $z = 3$. Because numerical simulations usually have more limited output time resolution at higher redshift, the SFHs of galaxies at higher redshift are expected to contain more noise. These suggest that the full potential of the empirical model is

limited by the resolution and output frequency of the hydrodynamic simulation used. Consistent with this, the results in Section 3.1 show that the residual in the sSFR cannot be fully explained by halo properties, and (iv) The bias and scatter for both SFR and $M_{*,\text{int}}$ at all redshifts are only slightly larger in $T_{\text{EAGLE+TNG-modes}}$ than in T_{EAGLE} , indicating that the borrow of template does not introduce large error in the model. All these confirm that the model is powerful in describing the SFH of star-forming galaxies.

The modelled SFHs obtained from the quenched sample S_Q in the first three test cases are shown in Fig. 10 in comparison with the simulation results. Compared with the results for star-forming population, the SFR and $M_{*,\text{int}}$ histories are as well reproduced over a wide range of redshift. Case $T_{\text{EAGLE+TNG-modes}}$, which uses the TNG template, also gives results comparable to cases where EAGLE template itself is used. The only exception is at low redshift when these galaxies are quenched and the SFRs decrease quickly to very low values for the model to predict accurately. However, even in this case, the predicted $M_{*,\text{int}}$ histories still closely follow the simulated ones.

For the quenched sample S_Q , we also show, in Fig. 11, the comparisons between the model predictions and the simulated results for both SFR and $M_{*,\text{int}}$ at several redshifts between 0 and 3. At low redshift, the SFR of quenched galaxies cannot be predicted accurately, so that both the bias and scatter are large. As we go to higher redshift, the bias and scatter decrease. These indicate that, even for a galaxy that is quenched at $z = 0$, it is still possible to infer its SFH from its halo MAH. In all of the three test cases, the modelled $M_{*,\text{int}}$ is tightly correlated with the simulation results, with almost no bias at low z and small bias at $z = 3$, and with small scatter at all redshifts. Again, the use of TNG template to model EAGLE galaxies in $T_{\text{EAGLE+TNG-modes}}$ is as good as using EAGLE's own template in

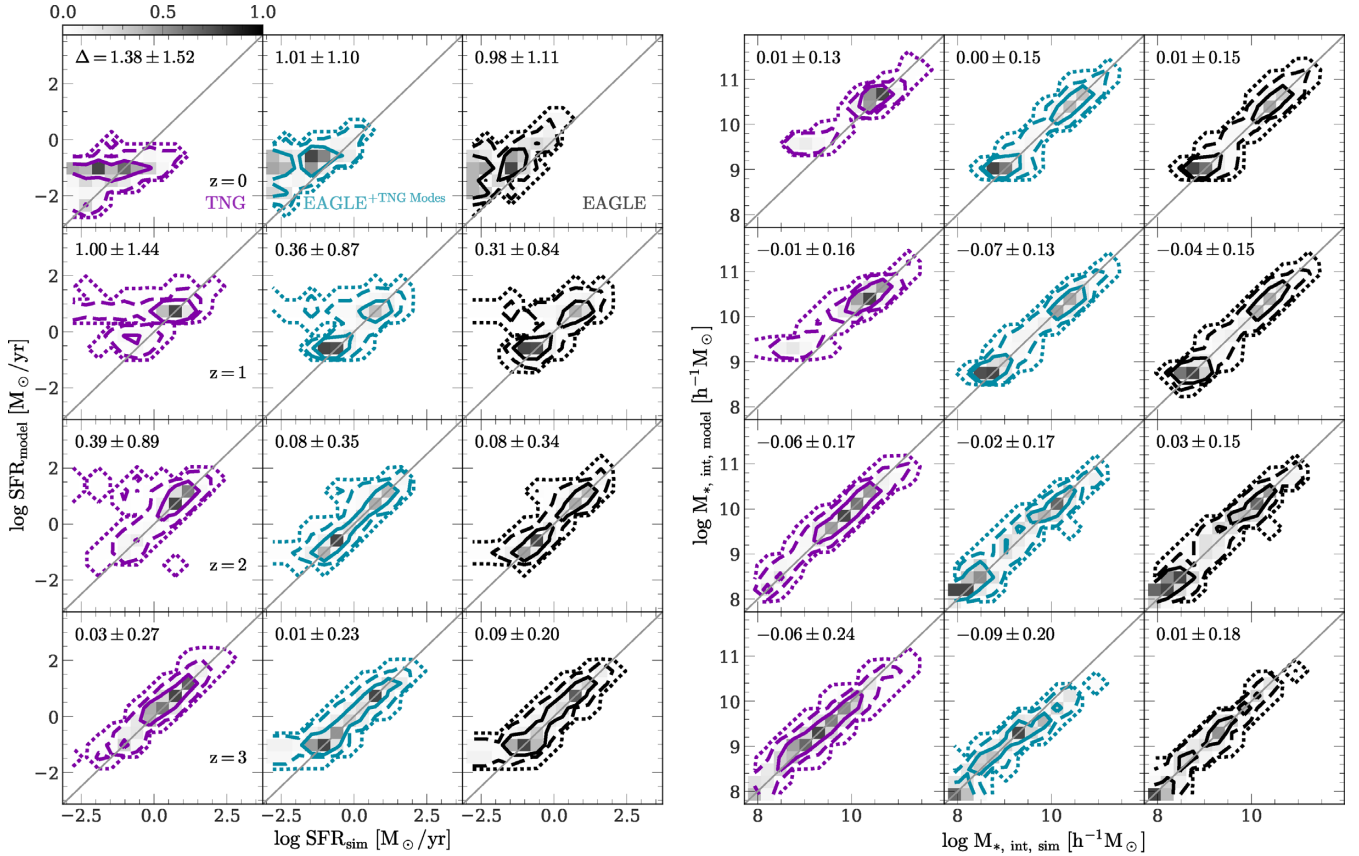


Figure 11. The same as Fig. 9, but for the quenched galaxy sample S_Q .

T_{EAGLE} , indicating that the model can reproduce the SFH even for quenched galaxies.

Based on these test results from T_{TNG} , $T_{EAGLE+TNG-modes}$, and T_{EAGLE} , we conclude that our model can describe accurately the SFH of both star-forming and quenched galaxies except for the current SFR of quenched galaxies. Thus, if we can find a way to separate the star-forming and quenched populations, a model can be constructed for both populations. In the following, we show that a statistically correct model can be constructed even if a clean separation between the two population is not feasible (because of the reason discussed in Section 3.2).

In the remaining two test cases, we need to first classify a galaxy as star forming or quenched, and then model it by the trained model appropriate for its class. Fig. 12 shows the results based on T_{join} , where the distributions of model galaxies in the $(\log M_{*,int}, \log sSFR)$ plane at four different redshifts are compared with the simulated results. At $z = 0$, model galaxies show a bimodal distribution, consistent with the simulation results. At higher redshifts, the simulation shows some weak sign of bimodality, which is not well captured by the model. In the simulation, the mean value of sSFR of the main sequence increases slowly with redshift, a trend that is well reproduced by the model. Consistent with the simulation, the scatter in the modelled main sequence decreases with redshift. However, the predicted amount of scatter at $z = 0$ is smaller than that in the simulation, which is due to the limited degrees of freedom of the random component used in the post-processing. Fig. 12 also shows the galaxy distribution using T'_{join} and EAGLE's own templates to model EAGLE galaxies. It is clear that the use of TNG templates in T_{join} is as good as using EAGLE's own template in T'_{join} .

In Fig. 13, we show the $SFR-M_{halo}$ and M_*-M_{halo} relations at four redshifts from $z = 0$ to $z = 3$ for the test case T_{join} , in comparison with the simulation results. For comparison, we also show the results from the test case T'_{join} to test the effect of borrowing external template. As one can see, the M_*-M_{halo} relations predicted by the model for T_{join} match well the simulation results. Only at $z > 2$ is the modelled M_*-M_{halo} relation slightly lower. Compared with the results for T'_{join} , which match the simulation results almost perfectly, this small difference is clearly produced by the use of the imperfect template in T_{join} .

The modelled $SFR-M_{halo}$ relation in T_{join} is also similar to that in the simulation, with moderate discrepancy at $M_{halo} > 10^{12} h^{-1} M_{\odot}$. Comparing this to the predictions of T'_{join} , which match the simulation results better but not perfectly, we infer that this discrepancy is partly due to the use of imperfect template in T_{join} and partly due to the imperfect classification of star-forming and quenched galaxies in both cases. As discussed in Section 3.3, the decision boundary is ambiguous for high-mass galaxies, which are hosted by massive haloes, and a slightly offset in the decision is likely to produce a significantly different result.

Since we have already included PCs of v_{max} as features in the regressors, our model is expected to reproduce the dependency of galaxy properties on halo MAH, a phenomenon usually referred to as the 'assembly bias'. To demonstrate this, we plot the relation between the halo half-mass formation time, $z_{mb, 1/2}$, and galaxy sSFR for galaxies at $z = 0$ for case T_{join} . The results for four different stellar mass bins are shown in Fig. 14, where we also include the results from the simulation and from T'_{join} for comparison. In all of the mass bins, galaxies in haloes of earlier assembly on average have smaller sSFR. Both T_{join} and T'_{join} can reproduce this trend.

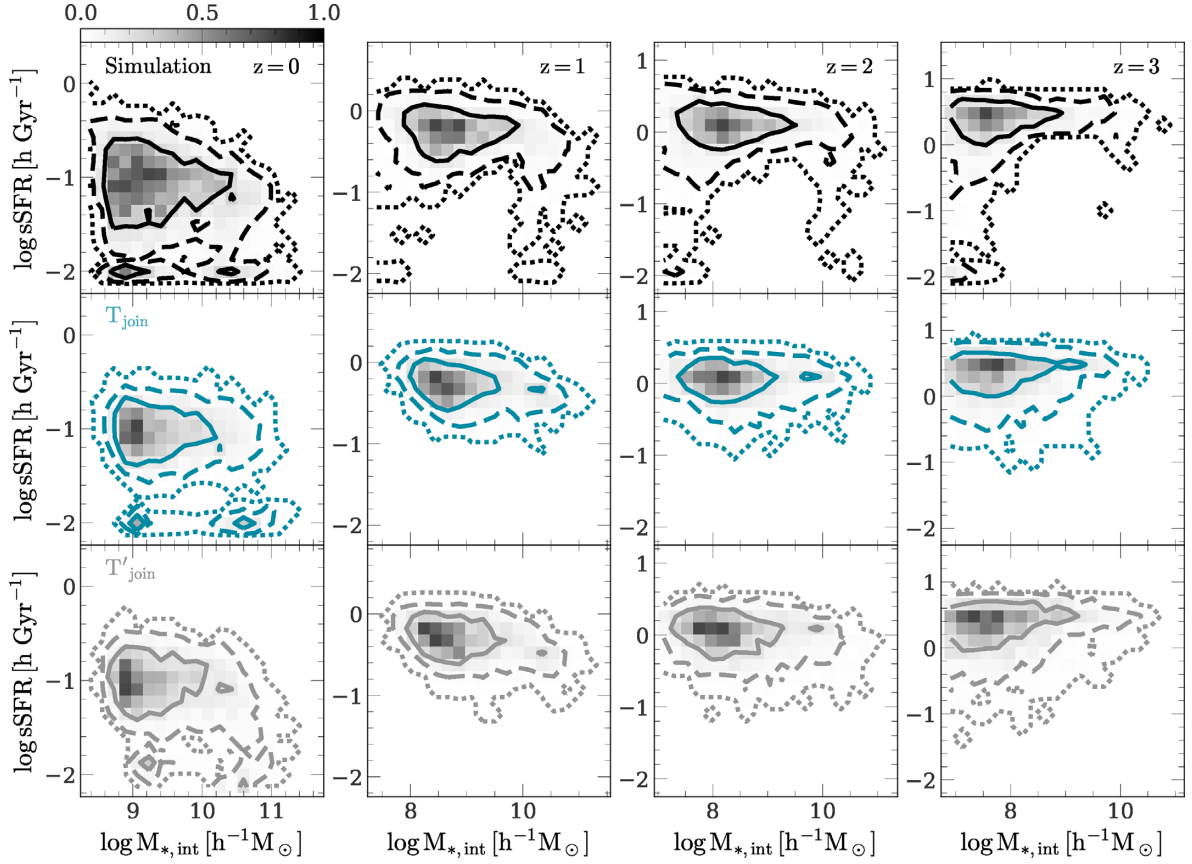


Figure 12. The relation between the smoothed sSFR and $M_{*,\text{int}}$ in the histories of all the test galaxies in T_{join} and T'_{join} , in comparison with the EAGLE simulation. Panels from top to bottom are for the EAGLE simulation, cases T_{join} and T'_{join} , respectively. Each column shows the relation at a specific redshift as indicated in the first row. The grey shades are normalized histograms.

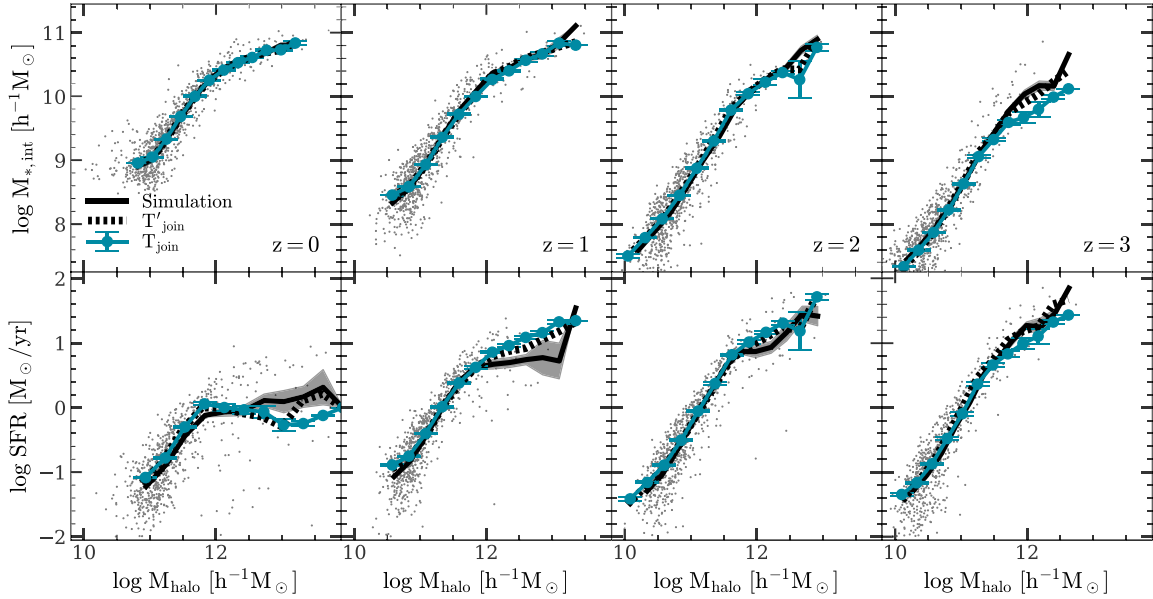


Figure 13. The relation between halo mass M_{halo} and stellar properties (upper row: $M_{*,\text{int}}$; lower row: the smoothed SFR) at different redshifts (as indicated) in the histories of all the test galaxies in T_{join} (the green line with error bars indicating the standard deviation) and T'_{join} (the black-dashed line). In each panel, the grey dots are from the EAGLE simulation; the solid black line and the grey shade indicate the mean and standard deviation, respectively. Galaxies with $\text{sSFR} > 10^{-2} h \text{ Gyr}^{-1}$ are used for the SFR results.

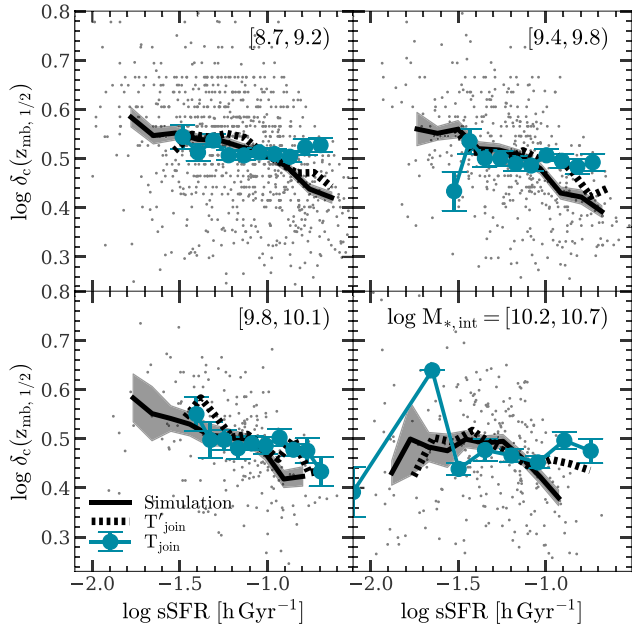


Figure 14. The relation between halo half-mass formation time, $z_{\text{mb}, 1/2}$ and galaxy sSFR for $z = 0$ galaxies with different $M_{*, \text{int}}$, as indicated in each panel. In each panel, grey dots are from the EAGLE simulation, with the solid black line and shade indicating the mean and standard deviation of the mean, respectively. Test results using T_{join} are shown by the green line with error bars, while those using T'_{join} are shown by the black-dashed line. Only galaxies with $\text{sSFR} > 10^{-2} h \text{ Gyr}^{-1}$ are used.

The results of T_{join} have a small bias relative to the simulation. Because the effects of halo PCs on galaxy SFH are much smaller than the total scatter of the star-forming main sequence (see Section 3.1), the regressor that maps halo PCs to galaxy properties tends to reduce the model variance at the cost of increasing bias. When the SFH template adopted in the model is imperfect, the bias is larger, as is seen in the results of T_{join} in comparison with those of T'_{join} . Overall, our model reproduce correctly the assembly bias in the data, especially when the PC template can account for the variance in the SFH of galaxies.

To conclude, the tests using T_{join} and T'_{join} demonstrate that our empirical model can describe the galaxy–halo relation correctly in a statistical way, even though the classification between star-forming and quenched galaxies is not accurate for individual galaxies.

5 SUMMARY

In this paper, we use the TNG and EAGLE simulation data to infer the galaxy–halo relations that are needed to build an empirical model for central galaxies in dark matter haloes. Our analysis is based on PCA for dimension reduction and GBDT for regression and classification. Our main results and their implications are summarized as follows.

(i) The star-forming main sequence is a well-defined population driven by v_{max} of host haloes. The M_*-v_{max} and $\text{SFR}-v_{\text{max}}$ relations for this population at $z = 0$ are both tight, with $R^2 \geq 0.9$ and 0.7 , respectively, and they are even tighter at higher z . Other halo properties are secondary and provide only small improvements in the predictions of M_* and SFR.

(ii) The residual of the $\text{SFR}-M_*$ relation for the main sequence, represented by $\Delta \log \text{sSFR}$, is not dominated by any halo property tested in this paper. Using a combination of a large set of halo

properties, the value of R^2 in the prediction of $\Delta \log \text{sSFR}$ is still < 0.5 at both low and high z . These indicate that modelling the SFR based on halo properties with the use of deterministic relation between the two can lead to spurious and biased results. A random component is needed in order to model SFR in a statistically unbiased way.

(iii) The quenching of a low-mass central galaxy is tightly correlated with the infall-ejection process of the host halo. In contrast, the quenching of a high-mass central galaxy is related closely to the formation of a massive progenitor in its host halo at high z , as indicated by the core formation redshift, $z_{\text{mb, core}}$. For both low-mass and high-mass galaxies, it is difficult to train classifiers that can separate the star-forming from the quenched population because of the sample imbalance and overlapped distribution between these two populations.

(iv) For the quenched population, $M_{*, \text{int}}$ is tightly correlated with halo v_{max} . The $M_{*, \text{int}}$ at $z = 0$ depends predominantly on v_{max} , while PCs of the $M_{*, \text{int}}$ history are correlated with the PCs of the v_{max} history. In general, the higher order PCs of $M_{*, \text{int}}$ are less well recovered by the regressors.

Based on the inferred galaxy–halo relations, we propose an empirical model for star formation in central galaxies of dark matter haloes. The main procedures can be summarized as follows:

(i) The empirical model consists of three procedures, which reduce the dimension of halo MAH by the PCA, map the halo properties into stellar properties by the GBDTs, and recover the dimension of the SFH by the inverse of the PCA.

(ii) For both star-forming and quenched galaxies, the empirical model shows good performances in all of our test cases. The reconstructed SFHs of individual galaxies follow the correct trends in comparison with the simulated results. The SFR and $M_{*, \text{int}}$ at all redshifts are reconstructed with small bias and small residuals. The only exception occurs for some quenched galaxies where the SFRs in the simulations decrease too rapidly to capture by the model.

(iii) Central galaxies can be classified into star-forming and quenched populations on the basis of halo properties, and can be modelled separately according to their classes. Although the classification is imperfect and has contamination between the two classes, the predicted statistical properties of the galaxies match well with the simulation inputs. These include the bimodal distribution of galaxies in the SFR–stellar mass diagram, the stellar mass–halo mass, and SFR–halo mass relations of galaxies at different z , and the assembly bias of galaxies.

The results presented here provide a framework of using hydrodynamic simulations to discover ingredients that can be included in empirical models of galaxy formation and to build templates that can be used to reduce the model complexity. In the future, we will extend our analysis by including satellite galaxies. The results obtained in this paper can be used as the initial conditions before a galaxy becomes a satellite, and the subsequent evolution of the satellite population is to be modelled again on the basis of halo properties, such as halo masses and merging orbits. With these, we will build a full empirical model based on the architecture provided by numerical simulations.

ACKNOWLEDGEMENTS

This work is supported by the National Key Research and Development Program of China (grant No. 2018YFA0404502), and the National Science Foundation of China (grant Nos. 11821303,

11973030, 11673015, 11733004, 11761131004, 11761141012). We acknowledge the Virgo Consortium for making their EAGLE simulation data available. We acknowledge Dandan Xu, Yuning Zhang and Jingjing Shi for accessing the TNG simulation data. YC and KW gratefully acknowledge the financial support from China Scholarship Council.

DATA AVAILABILITY

The data and software underlying this article will be shared on reasonable request to the corresponding author. They are available at <https://www.chenyangyao.com/publication/20/empirical-model-satellite/>. The computation in this work is supported by the HPC toolkit HIPPI at <https://github.com/ChenYangyao/hipp>.

REFERENCES

- Behroozi P., Wechsler R. H., Hearin A. P., Conroy C., 2019, *MNRAS*, 488, 3143
- Berlind A. A., Weinberg D. H., 2002, *ApJ*, 575, 587
- Bernardi M., Hyde J. B., Sheth R. K., Miller C. J., Nichol R. C., 2007, *AJ*, 133, 1741
- Bett P., Eke V., Frenk C. S., Jenkins A., Helly J., Navarro J., 2007, *MNRAS*, 376, 215
- Bishop C. M., 1995, *Neural Comput.*, 7, 108
- Bishop C. M., 2006, *Pattern Recognition and Machine Learning*. Springer, Berlin. Available at: <https://www.springer.com/gp/book/9780387310732>
- Breiman L., 2001, *Mach. Learn.*, 45, 5
- Bryan G. L., Norman M. L., 1998, *ApJ*, 495, 80
- Carroll S. M., Press W. H., Turner E. L., 1992, *ARA&A*, 30, 499
- Chaves-Montero J., Hearin A., 2020, *MNRAS*, 495, 2088
- Chen Y., Mo H. J., Li C., Wang H., Yang X., Zhang Y., Wang K., 2020, *ApJ*, 899, 81
- Cormen T. H., Leiserson C. E., Rivest R. L., Stein C., 2009, *Introduction to Algorithms*, 3rd edn. The MIT Press, Cambridge, MA. Available at: <https://mitpress.mit.edu/books/introduction-algorithms-third-edition>
- Crain R. A. et al., 2015, *MNRAS*, 450, 1937
- Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, *ApJ*, 292, 371
- Dolag K., Borgani S., Murrante G., Springel V., 2009, *MNRAS*, 399, 497
- Eisenstein D. J., Hu W., 1998, *ApJ*, 496, 605
- Faltenbacher A., White S. D. M., 2010, *ApJ*, 708, 469
- Freund Y., Schapire R. E., 1997, *J. Comput. Syst. Sci.*, 55, 119
- Friedman J. H., 2001, *Ann. Stat.*, 29, 1189
- Friedman J. H., 2002, *Comput. Stat. Data Anal.*, 38, 367
- Gao L., White S. D. M., 2007, *MNRAS*, 377, L5
- Gao L., Springel V., White S. D. M., 2005, *MNRAS*, 363, L66
- Gao Y., Fan L.-L., 2020, *Res. Astron. Astrophys.*, 20, 106
- Guo H. et al., 2016, *MNRAS*, 459, 3040
- Guo Q., White S., Li C., Boylan-Kolchin M., 2010, *MNRAS*, 404, 1111
- Hahn O., Porciani C., Carollo C. M., Dekel A., 2007, *MNRAS*, 375, 489
- Hastie T., Tibshirani R., Friedman J., 2001, *The Elements of Statistical Learning*, Springer Series in Statistics. Springer, New York, NY. Available at: <https://link.springer.com/book/10.1007/978-0-387-84858-7>
- He K., Zhang X., Ren S., Sun J., 2015, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New York, NY, p. 770
- Hearin A. P., Watson D. F., 2013, *MNRAS*, 435, 1313
- Hearin A. P., Watson D. F., Becker M. R., Reyes R., Berlind A. A., Zentner A. R., 2014, *MNRAS*, 444, 729
- Hotelling H., 1933, *J. Educ. Psychol.*, 24, 417
- Huang G., Liu Z., Van Der Maaten L., Weinberger K. Q., 2017, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New York, NY, p. 2261
- Jeon-Daniel A., Vecchia C. D., Haas M. R., Schaye J., 2011, *MNRAS*, 415, L69
- Jiang L., Helly J. C., Cole S., Frenk C. S., 2014, *MNRAS*, 440, 2115
- Jing Y. P., 2000, *ApJ*, 535, 30
- Jing Y. P., Mo H. J., Borner G., 1998, *ApJ*, 494, 1
- Jing Y. P., Suto Y., Mo H. J., 2007, *ApJ*, 657, 664
- Kauffmann G. et al., 2003, *MNRAS*, 341, 54
- Li Y., Mo H. J., Gao L., 2008, *MNRAS*, 389, 1419
- Lim S. H., Mo H. J., Wang H., Yang X., 2016, *MNRAS*, 455, 499
- Lu Y., Mo H. J., Weinberg M. D., Katz N., 2011, *MNRAS*, 416, 1949
- Lu Z., Mo H. J., Lu Y., Katz N., Weinberg M. D., van den Bosch F. C., Yang X., 2014a, *MNRAS*, 439, 1294
- Lu Y., Mo H. J., Lu Z., Katz N., Weinberg M. D., 2014b, *MNRAS*, 443, 1252
- Lu Z., Mo H. J., Lu Y., Katz N., Weinberg M. D., van den Bosch F. C., Yang X., 2015, *MNRAS*, 450, 1604
- McAlpine S. et al., 2016, *Astron. Comput.*, 15, 72
- MacCìo A. V., Dutton A. A., Van Den Bosch F. C., 2008, *MNRAS*, 391, 1940
- Marinacci F. et al., 2018, *MNRAS*, 480, 5113
- Meng J., Li C., Mo H., Chen Y., Wang K., 2020, preprint ([arXiv:2008.13733](https://arxiv.org/abs/2008.13733))
- Mo H. J., White S. D. M., 1996, *MNRAS*, 282, 347
- Mo H. J., Mao S., White S. D. M., 1999, *MNRAS*, 304, 175
- Mo H., van den Bosch F., White S., 2010, *Galaxy Formation and Evolution*. Cambridge Univ. Press, Cambridge. Available at: <http://ebooks.cambridge.org/ref/id/CBO9780511807244>
- Moster B. P., Naab T., White S. D. M., 2018, *MNRAS*, 477, 1822
- Moster B. P., Naab T., Lindström M., O’Leary J. A., 2020, preprint ([arXiv:2005.12276](https://arxiv.org/abs/2005.12276))
- Naiman J. P. et al., 2018, *MNRAS*, 477, 1206
- Navarro J. F., Frenk C. S., White S. D. M., 1997, *ApJ*, 490, 493
- Nelson D. et al., 2018, *MNRAS*, 475, 624
- Nelson D. et al., 2019, *Comput. Astrophys. Cosmol.*, 6, 2
- O’Donnell C., Behroozi P., More S., 2021, *MNRAS*, 501, 1253
- Pearson K., 1901, *London Edinburgh Dublin Phil. Mag. J. Sci.*, 2, 559
- Pillepich A. et al., 2018a, *MNRAS*, 473, 4077
- Pillepich A. et al., 2018b, *MNRAS*, 475, 648
- Planck Collaboration I, 2014, *A&A*, 571, A1
- Planck Collaboration XIII, 2015, *A&A*, 594, A13
- Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, *ApJ*, 771, 30
- Rodriguez-Gomez V. et al., 2015, *MNRAS*, 449, 49
- Schaye J. et al., 2014, *MNRAS*, 446, 521
- Sedgewick R., Wayne K., 2011, *Algorithms*, 4th edn. Addison-Wesley Professional, Boston, MA. Available at: <https://www.pearson.com/us/higher-education/program/Sedgewick-Algorithms-4th-Edition/PGM100869.html>
- Shen S., Mo H. J., White S. D. M., Blanton M. R., Kauffmann G., Voges W., Brinkmann J., Csabai I., 2003, *MNRAS*, 343, 978
- Sheth R. K., Mo H. J., Tormen G., 2001, *MNRAS*, 323, 1
- Shi J., Wang H., Mo H. J., Xie L., Wang X., Lapi A., Sheth R. K., 2018, *ApJ*, 857, 127
- Shi J. et al., 2020, *ApJ*, 893, 139
- Springel V., 2005, *MNRAS*, 364, 1105
- Springel V., 2010, *MNRAS*, 401, 791
- Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, *MNRAS*, 328, 726
- Springel V. et al., 2018, *MNRAS*, 475, 676
- The EAGLE team, 2017, preprint ([arXiv:1706.09899](https://arxiv.org/abs/1706.09899))
- Vale A., Ostriker J. P., 2004, *MNRAS*, 353, 189
- van den Bosch F. C., 2002, *MNRAS*, 331, 98
- Wang H., Mo H. J., Jing Y. P., Yang X., Wang Y., 2011, *MNRAS*, 413, 1973
- Wang H. et al., 2018, *ApJ*, 852, 31
- Wechsler R. H., Tinker J. L., 2018, *ARA&A*, 56, 435
- Wechsler R. H., Bullock J. S., Primack J. R., Kravtsov A. V., Dekel A., 2002, *ApJ*, 568, 52
- Wechsler R. H., Zentner A. R., Bullock J. S., Kravtsov A. V., Allgood B., 2006, *ApJ*, 652, 71
- Weinberger R. et al., 2017, *MNRAS*, 465, 3291
- Wetzel A. R., Nagai D., 2015, *ApJ*, 808, 40
- Xu H., Zheng Z., Guo H., Zu Y., Zehavi I., Weinberg D. H., 2018, *MNRAS*, 481, 5470
- Yang X., Mo H. J., van den Bosch F. C., 2003, *MNRAS*, 339, 1057
- Yoon Y., Park C., 2020, *ApJ*, 897, 121

Zhao D. H., Mo H. J., Jing Y. P., Börner G., 2003a, *MNRAS*, 339, 12
 Zhao D. H., Jing Y. P., Mo H. J., Börner G., 2003b, *ApJ*, 597, L9
 Zhao D. H., Jing Y. P., Mo H. J., Börner G., 2009, *ApJ*, 707, 354

APPENDIX A: PCA OF GALAXY AND HALO FORMATION HISTORIES

The PCA is an unsupervised, reduced linear Gaussian dimension-reduction method (Pearson 1901; Hotelling 1933). As demonstrated in Chen et al. (2020), the halo MAH, which is a vector in high-dimensional space, can be effectively reduced to several PCs that still capture most of the sample variance. Here, we briefly describe how we apply the PCA to galaxy and halo formation histories. A modern and detailed theoretical description of the PCA can be found in Bishop (2006).

The various ‘history’ quantities considered in this paper are also vectors in high dimension space, and we use the PCA to reduce their dimensions so that each history can be described by a set of PCs. For each of the histories, \mathbf{h} ($\mathbf{h} = \mathbf{v}_{\max}$, \mathbf{SFR} , or $\mathbf{M}_{*,\text{int}}$), we apply the PCA according to the following steps:

(i) Because of the resolution limit of the simulations, the history of a galaxy cannot be traced back to an arbitrarily high redshift. For a galaxy sample S , we trim \mathbf{h} of each galaxy above a chosen redshift, so that 90 per cent of the galaxies have history measurements for the remaining redshifts. Galaxies that do not have history measurements at some of the remaining redshifts are padded with a small value to ensure numerical stability.

(ii) We make a proper transformation of \mathbf{h} according to the description given in Section 4.1 to make it suitable for PCA. The transformed history is denoted by $\tilde{\mathbf{h}}$.

(iii) We apply the PCA to $\tilde{\mathbf{h}}$ of all galaxies in S . The PCA gives a mean offset \mathbf{o} , and a set of new base vectors \mathbf{e}_i ($i = 1, 2, 3, \dots$) whose eigen-values, λ_i ($i = 1, 2, 3, \dots$), are ranked in a descending order. The history is then transformed into the new frame by

$$\mathbf{PC} = (\mathbf{e}_1, \mathbf{e}_2, \dots)^T (\tilde{\mathbf{h}} - \mathbf{o}). \quad (\text{A1})$$

To reduce the dimension of $\tilde{\mathbf{h}}$, we can keep a set of m important PCs, $\mathbf{PC}_m = (\text{PC}_1, \text{PC}_2, \dots, \text{PC}_m)^T$. We can reconstruct $\tilde{\mathbf{h}}$ from \mathbf{PC}_m using

$$\tilde{\mathbf{h}}_{\text{recon},m} = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m) \mathbf{PC}_m + \mathbf{o}. \quad (\text{A2})$$

This inevitably causes some loss of information. The information retained by \mathbf{PC}_m is by described the cumulative proportional variance ratio (cPVE), defined as

$$\text{cPVE}_m = \frac{\text{Var}[\tilde{\mathbf{h}}_{\text{recon},m}]}{\text{Var}[\tilde{\mathbf{h}}]}. \quad (\text{A3})$$

In Section 4.1, we consider a case mimicking real applications, in which the dimension-reduction templates from the TNG simulation are applied to reduce the dimension of the SFHs of the EAGLE simulation. To this end, we first apply the PCA to both TNG and EAGLE. We then keep only the EAGLE offset vector \mathbf{o} , and replace all EAGLE base vectors \mathbf{e}_i with the TNG base vectors interpolated to the redshifts of the EAGLE snapshots. Using this new frame, we can compute the PCs for each EAGLE SFH, and measure the performance of the reconstruction by the corresponding cPVE.

The cPVE as a function of m is shown in Fig. A1. As one can see, when using the templates obtained from a simulation itself, the v_{\max} and M_* histories in both TNG and EAGLE converge quickly to 1, indicating that the first several PCs take most of the variance. This shows that the main structures of the halo MAH and the stellar

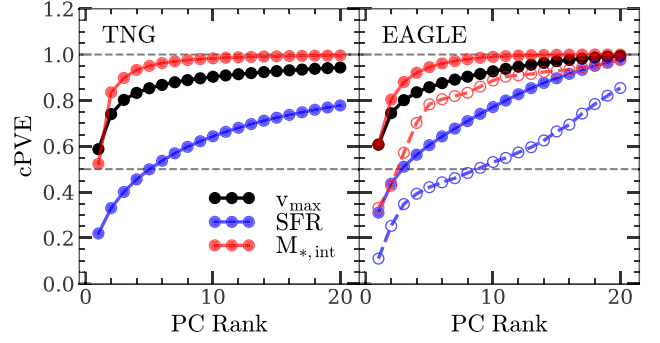


Figure A1. cPVE curves for v_{\max} , SFR, and $M_{*,\text{int}}$ histories of haloes and galaxies in TNG (left) and EAGLE (right). In the right-hand panel, the open circles joined by dashed lines are cPVE curves for EAGLE SFR (blue) and $M_{*,\text{int}}$ (red) histories using PCA templates from TNG.

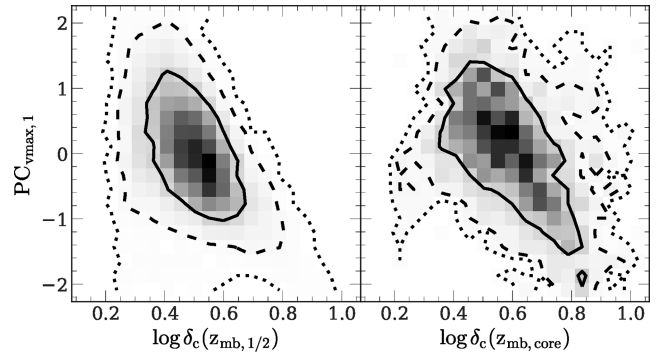


Figure A2. Relation between the first PC of halo v_{\max} history and formation time $z_{\text{mb},1/2}$ (left-hand panel; using all $z = 0$ TNG haloes with $M_{\text{halo}} \geq 10^{8.5} h^{-1} M_{\odot}$) and $z_{\text{mb,core}}$ (right-hand panel; using all $z = 0$ TNG haloes with $M_{\text{halo}} \geq 10^{10} h^{-1} M_{\odot}$ because small haloes do not have $z_{\text{mb,core}}$ measurements).

MAH are fairly simple, and can be effectively described by a small number of parameters. For the SFR history, the first several PCs are still the most important ones, but cPVE increases slowly as m increases, indicating that the SFR history is noisy on small time-scales. This can be seen from the plots of SFR histories of individual galaxies presented in Section 4.3. It is thus only sensible to link the main structure of the SFR history to halo properties, but to treat the small-scale fluctuation as a random (uncorrelated) component to be included in the empirical model. The design of our empirical model in Section 4.1 exploits this idea.

When using the TNG templates to describe EAGLE histories, the reconstruction is poorer, as shown by the open circles connected by dashed lines in Fig. A1. However, the first several PCs are still the most important ones and each of the higher order PCs contributes only a small fraction of the cPVE.

The parametrization using PCs for halo MAH has several advantages over formation times. PCs of MAHs are linearly orthogonal therefore reducing the degeneracy. The first PC of MAH has better correlation with halo concentration than other formation times do, as shown in Chen et al. (2020). PCs also have clear physical meaning. The first PC of MAH is tightly related to the formation time for haloes with mass exceeding a certain value. Fig. A2 demonstrates its relation with two formation times, and

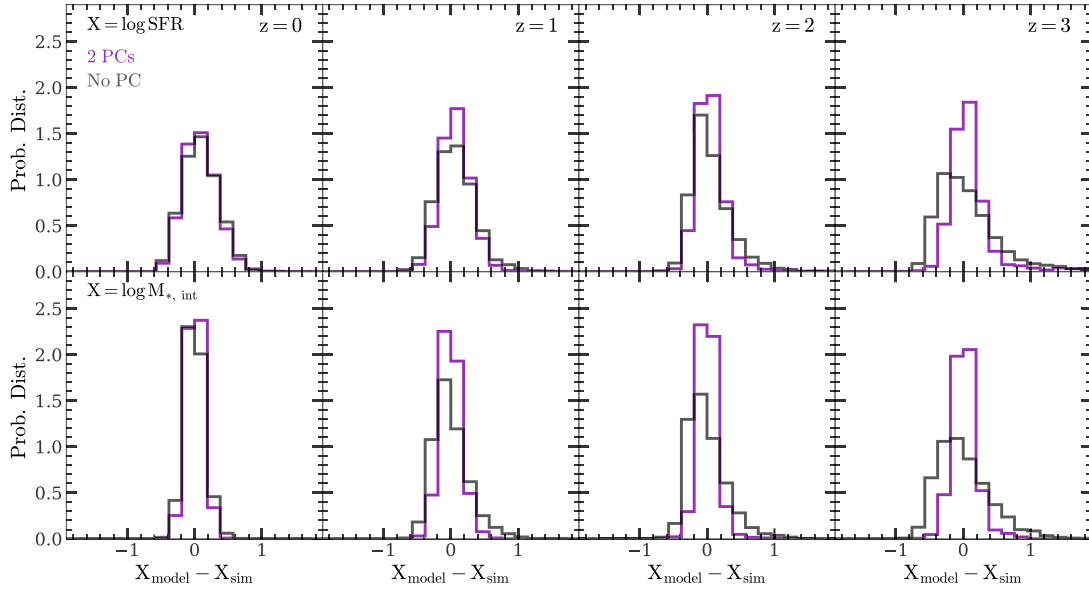


Figure A3. Distributions of the log SFR (upper row) and log $M_{*,\text{int}}$ (lower row) difference between model galaxies and TNG simulated galaxies. Here, we show the test galaxies in the sample S_{SF} in the first test case T_{TNG} . The purple histograms are from the model using the first two PCs for MAH and SFH. The black histograms are from the model without any PC.

we see a tight relation between them. Higher order PCs naturally reflect more subtle properties in the formation histories of haloes, such as major mergers (see Chen et al. 2020, for detailed discussions).

PCs of halo MAHs are also tightly correlated with the SFH of galaxies (see Section 3). Although M_{halo} or v_{max} is the dominant factor in galaxy SFH, more information can be captured with the help of PCs, so that the empirical model is more powerful in describing the details of galaxy SFH. We show an example of this improvement in the Fig. A3, where we use the test case T_{TNG} (see Section 4.2) to demonstrate the difference made by including the PCs of both MAH and SFH.

APPENDIX B: GRADIENT BOOSTED DECISION TREES

Boosting is a large set of model ensemble methods that combine multiple weak learners (regressors or classifiers) to produce a strong learner capable of capturing complex patterns in statistical learning tasks. Compared with other ensemble methods, such as the random forest (Breiman 2001) that starts with strong learners and uses multiple-sourced randomness to suppress model variance, boosting methods are faster in computation and still maintain comparable performance.

A successful example of boosting methods is AdaBoost (Freund & Schapire 1997), which can be viewed as a ‘greedy’ algorithm that optimizes an exponential objective function (see e.g. Bishop 2006). The extensions of this method to arbitrary differentiable objective functions can be made through gradient boosting or GBDT (Friedman 2001), and stochastic optimization strategies (Friedman 2002). The idea behind boosting motivates the developments of some modern deep neural networks, such as those with residual blocks (ResNet, see He et al. 2015) and densely connected blocks (DenseNet, see Huang et al. 2017). In this paper, we use GBDT for both regression and classification.

The idea of boosting is to build a sequence of weak learners $f_i(\mathbf{x})$ ($i = 1, 2, \dots, M$), and combine them to form a regression function or classification function,

$$F_M(\mathbf{x}) = \sum_{i=0}^M f_i(\mathbf{x}). \quad (\text{B1})$$

In regression problems, F_M maps the feature variable \mathbf{x} to the target value. In classification problems, F_M maps \mathbf{x} to the class probability, and the final prediction is chosen to be the class with the highest probability. Once we have a training data set $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, the best F_M is the one that minimizes the loss function $l(F_M|D)$.

Without any constraint, the optimization of l is infeasible because the functional space of f_i has infinity dimensions and the possible combinations are also infinite. The GBDT provides a tree-based ‘greedy’ algorithm to solve this problem. Starting from an arbitrary naive learner F_0 (e.g. a constant function), the GBDT algorithm recursively adds new learner f_M into F_{M-1} to give $F_M = F_{M-1} + f_M$, such that $l(F_M|D) < l(F_{M-1}|D)$. To find the best f_M at each iteration, we expand l as a series,

$$l(F_M|D) = l(F_{M-1}|D) + f_M \cdot \nabla_F l(F|D)|_{F=F_{M-1}}. \quad (\text{B2})$$

If f_M is chosen such that $f_M = -\alpha \nabla_F l(F|D)|_{F=F_{M-1}}$, with α being the learning rate, then the loss function is guaranteed to decrease, and the iteration is an example of the gradient descent algorithm. In general, f_M can be any function that is parallel with the gradient.

In our applications, we use only the loss function derived from the exponential family (L2 loss in regression; cross-entropy loss in classification; see Bishop 2006), so that the gradient

$$\nabla_F l(F|D)|_{F=F_{M-1}} = F_{M-1}(\mathbf{x}_i) - \mathbf{y}_i \quad (i = 1, 2, \dots, N), \quad (\text{B3})$$

which is the residual of F_{M-1} relative to the real target values. With this choice, we train a shallow decision tree regressor $t(\mathbf{x})$ with the training set $\{(\mathbf{x}_i, F_{M-1}(\mathbf{x}_i) - \mathbf{y}_i)\}_{i=1}^N$, and finally obtain f_M using $f_M = -\alpha t$. In the iterative process modelled above, trees built earlier mainly handle the large-scale structures in the feature space, while

those built later focus on the local difficulties that have not been captured.

The boosting algorithm defined above may have problems from overfitting. To overcome them, we use the stochastic GBDT (Friedman 2002). At each iteration step, we only use a random subset of the whole data set to train the tree $t(\mathbf{x})$. Such a randomness in the training set can effectively suppress the model variance, and is proved equivalent to ordinary regularization in some cases (see e.g. Bishop 1995).

For our analysis, we use the `scikit-learn` package to perform the GBDT. We choose the maximal depth of each tree to be three, which gives a sufficiently weak learner as required by boosting. A random subset of 75 per cent of the training data is used at each iteration step, which is sufficient to suppress overfitting for most tasks. We adopt a small learning rate, $\alpha = 0.08$, as recommended by Hastie, Tibshirani & Friedman (2001) to avoid overshoot. We use 25 per cent of the training data as the validation set, and terminate the iteration if the validation performance is not improved in 10 consecutive steps.

Once the ensemble of trees is built, the contribution of each variable $x \in \mathbf{x}$ in the prediction of target \mathbf{y} can be described by an importance value, $\mathcal{I}(x)$. This value is defined as the fraction of the decrease of the total loss caused by x in the construction of each tree, satisfying the normalization condition

$$\sum_{x \in \mathbf{x}} \mathcal{I}(x) = 1. \quad (\text{B4})$$

The definition of $\mathcal{I}(x)$ is motivated by the fact that the goal of a regressor or a classifier is to reduce the loss value. A variable x is more important if including it reduces more loss. So defined, a variable x with $\mathcal{I}(x) = 0$ does not contribute to determining the target \mathbf{y} , and can be neglected. In the other extreme where $\mathcal{I}(x) = 1$, the variable x dominates the prediction for \mathbf{y} , and other variables can be neglected.

The final performance of the ensemble is then evaluated at some test data, and is measured by R^2 , defined as the fraction of the variance of the target values explained, in regression problems, and by the correct-classification rate, r , in classification problems (see e.g. Chen et al. 2020, for a detailed description).

The R^2 value satisfies the condition

$$0 \leq R^2 \leq 1. \quad (\text{B5})$$

If a regressor has $R^2 = 1$, the relation between \mathbf{x} and \mathbf{y} is deterministic. On the other hand, $R^2 = 0$ indicates that there is no significant correlation between \mathbf{x} and \mathbf{y} . Thus, R^2 measures the correlation strength between the predictor and target variables.

As a simple demonstration, we consider an example where x and y satisfy a linear relation $y = kx + \epsilon$, with ϵ being a Gaussian random noise of zero mean and a constant variance. In such a case, the Pearson correlation coefficient $\rho_{x,y}$, reflects the correlation strength between the variable pair. If one builds a linear least-squares regression model and calculates the R^2 defined above, one gets $R^2 = \rho_{x,y}^2$, i.e. R^2 is just the square of the correlation coefficient.

If the variable pair have a non-linear relation or they are in high-dimensional space, the Pearson correlation coefficient is not so meaningful. In such cases, R^2 is a natural extension that has an interpretation similar to that in the linear case.

It is a common practice to adopt a threshold value to determine whether a correlation is strong or not. For example, if $R^2 < 0.5$ we may conclude that most of the driving factors are still missing in the model. On the other hand, if $R^2 > 0.5$, we may conclude that the main factors driving the target variable have already been included in the predictor set.

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.