THE ASTROPHYSICAL JOURNAL, 899:81 (16pp), 2020 August 10 https://doi.org/10.3847/1538-4357/aba597

Relating the Structure of Dark Matter Halos to Their Assembly and Environment

Yangyao Chen[1,2], H. J. Mo[2], Cheng Li[1], Huiyuan Wang[3,4], Xiaohu Yang[5,6], Youcai Zhang[7], and Kai Wang[1,2]

[1] Department of Astronomy, Tsinghua University, Beijing 100084, People's Republic of China; yangyao-17@mails.tsinghua.edu.cn

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

[3] Key Laboratory for Research in Galaxies and Cosmology, Department of Astronomy, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China

[4] School of Astronomy and Space Science, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China

[5] Department of Astronomy, School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai, 200240, People's Republic of China

[6] Tsung-Dao Lee Institute, and Shanghai Key Laboratory for Particle Physics and Cosmology, Shanghai Jiao Tong University, Shanghai, 200240, People's Republic of China

[7] Key Laboratory for Research in Galaxies and Cosmology, Shanghai Astronomical Observatory, Shanghai 200030, People's Republic of China

Received 2020 March 11; revised 2020 June 29; accepted 2020 July 12; published 2020 August 14

Abstract

We use a large *N*-body simulation to study the relation of the structural properties of dark matter halos to their assembly history and environment. The complexity of individual halo assembly histories can be well described by a small number of principal components (PCs), which, compared to formation times, provide a more complete description of halo assembly histories and have a stronger correlation with halo structural properties. Using decision trees built with the random ensemble method, we find that about 60%, 10%, and 20% of the variances in halo concentration, axis ratio, and spin, respectively, can be explained by combining four dominating predictors: the first PC of the assembly history, halo mass, and two environment parameters. Halo concentration is dominated by halo assembly. The local environment is found to be important for the axis ratio and spin but is degenerate with halo assembly. The small percentages of the variance in the axis ratio and spin that are explained by known assembly and environmental factors suggest that the variance is produced by many nuanced factors and should be modeled as such. The relations between halo intrinsic properties and environment are weak compared to their variances, with the anisotropy of the local tidal field having the strongest correlation with halo properties. Our method of dimension reduction and regression can help simplify the characterization of the halo population and clarify the degeneracy among halo properties.

Unified Astronomy Thesaurus concepts: Galaxy dark matter halos (1880)

1. Introduction

In the concordant Λ cold dark matter (ΛCDM) cosmology, dark matter halos, the dense clumps formed through gravitational collapse of the initial density perturbations, are the basic building blocks of the cosmic web. The formation history of a halo not only depends on the properties of the local density field, but may also be affected by the environment within which it forms. Since galaxies are believed to form in the gravitational potential wells of dark matter halos, the halo population provides a link between the dark and luminous sectors of the universe. Consequently, understanding the formation, structure, and environment of dark matter halos and their relations to each other has long been considered one of the most important parts of galaxy formation (e.g., Mo et al. 2010).

Dark matter halos are diverse in their structure, mass assembly history (MAH), and interaction with the large-scale environment. Among the structural properties of dark matter halos, the most important ones are the concentration parameter (Navarro et al. 1997), the spin parameter (Bett et al. 2007; Gao & White 2007; MacCiò et al. 2007), and the shape parameter (Jing & Suto 2002; Bett et al. 2007; Hahn et al. 2007; MacCiò et al. 2007). In *N*-body simulations, halo concentration is found to be correlated with halo mass and MAH (Navarro et al. 1997; Jing 2000; Wechsler et al. 2002; Zhao et al. 2003a, 2003b, 2009; MacCiò et al. 2007, 2008; Ludlow et al. 2014, 2016). The spin and shape parameters are also found to be related to other properties, such as halo mass and large-scale environment (MacCiò et al. 2007, 2008; Wang et al. 2011). However, these relations have a large variance and remain poorly quantified.

The mass assembly histories of halos in general are complex. In the literature, a common practice is to focus on the main-branch assembly histories, ignoring other branches (e.g., van den Bosch 2002; Wechsler et al. 2002; Zhao et al. 2003a, 2003b; Ludlow et al. 2013, 2014, 2016). Attempts have been made to describe the histories of individual halos with simple parametric forms (van den Bosch 2002; Wechsler et al. 2002; Zhao et al. 2003a, 2003b, 2009; Tasitsiomi et al. 2004; McBride et al. 2009; Correa et al. 2015). These simple models are useful in providing some crude description of halo assembly histories, but are not meant to give a complete characterization. Because of this, a variety of formation times have also been defined to characterize different aspects of the assembly histories of dark matter halos (see, e.g., Li et al. 2008 for a review). These formation times, usually degenerate among themselves, again only provide an incomplete set of information about the full assembly history.

The environment of a halo is also complex. To the lowest order, the average mass density around individual halos in a population can be used to characterize the distribution of the population relative to the underlying mass density field. In general, the spatial distribution of halos (halo environment) can depend on the intrinsic properties of the halos. The mass dependence, often called the halo bias (Mo & White 1996), is a natural outcome of the formation of halos in a Gaussian density field (Sheth et al. 2001; Zentner 2007). Furthermore, correlations have also been found between halo bias and halo assembly history. Halos, particularly low-mass ones, that formed earlier tend to be more strongly clustered (Gao et al. 2005; Gao & White 2007; Li et al. 2008). This phenomenon is now referred to as the halo assembly bias, and various studies have been carried out to understand its origin (Sandvik et al. 2007; Wang et al. 2007, 2009; Zentner 2007; Dalal et al. 2008; Desjacques 2008; Lazeyras et al. 2017). In addition to assembly history, halo bias has also been analyzed for its dependencies on other halo properties, such as halo concentration (Wechsler et al. 2006; Jing et al. 2007), substructure occupation (Wechsler et al. 2006; Gao & White 2007), halo spin (Bett et al. 2007; Gao & White 2007; Hahn et al. 2007; Wang et al. 2011), and halo shape (Hahn et al. 2007; Faltenbacher & White 2010; Wang et al. 2011.) These dependencies are collectively referred to as the "secondary bias" and sometimes also as the "assembly bias," presumably because these intrinsic properties may be related to halo formation.

Since halo properties are intrinsically correlated, it is necessary to investigate the joint distribution of different properties. Along this line, Jeeson-Daniel et al. (2011) used rank-based correlation coefficients to quantify the correlation between pairs of halo properties. Lazeyras et al. (2017) investigated halo bias as a function of two halo properties. They found that in all combinations of halo properties considered, halo bias can change with the second parameter when the first is fixed, and that the maximum of the halo bias occurs for halos with special combinations of the halo properties. For the environment, new parameters have also been introduced in addition to the local mass density. For example, Wang et al. (2011) used local tidal fields to represent the halo environment, and Salcedo et al. (2018) used the closest distance to a neighbor halo more massive than the halo in question. However, it is still unclear which quantity is the driving force of halo bias and which combination of bias sources can provide a more complete representation. Answering these questions requires more advanced statistical tools to measure the capability of models to fit the data and to identify hidden degeneracy between model parameters.

Some statistical tools are available for such investigations. For unsupervised learning tasks, the `Principal Component Analysis` (PCA) is a powerful tool to determine the main sources that contribute to the sample scatter and decompose the sample scatter along principal axes. For example, Wong & Taylor (2012) used this technique to reduce the dimension of halo MAH, while Cohn & van de Voort (2015) and Cohn (2018) applied the PCA to model the star formation history of galaxies. Jeeson-Daniel et al. (2011) used PCA to study the correlation of dark matter halo properties. For supervised learning tasks, the `Ensemble of Decision Trees` (EDT) or `Random Forest` (RF) is capable of both classification and regression. This method can also capture the nonlinear patterns, effectively reducing model complexity and discovering the dominant factors in target variables. RF has been widely used recently, for example, in identifying galaxy merger systems (de los Rios et al. 2016), in galaxy morphology classification (Dobrycheva et al. 2017; Sreejith et al. 2018), in predicting neutral hydrogen contents of galaxies (Rafieferantsoa et al. 2018), in determining structure formation in *N*-body simulations (Lucie-Smith et al. 2018; see also Lucie-Smith et al. 2019 for boosted trees), in measuring galaxy redshifts (Stivaktakis et al. 2018), in classifying star-forming versus quenched populations (Bluck et al. 2020), in identifying the best halo mass proxy in observation (Man et al. 2019), and in estimating the star formation rate and stellar mass of galaxies (Bonjean et al. 2019).

In this paper, we use both the PCA and the RF regressor to investigate the dependence of halo structural properties on halo assembly history and environment. The paper is organized as follows. In Section 2 we describe the simulation and the halo quantities to be analyzed. In Section 3 we demonstrate how to use PCA to extract information about halo assembly history. In Section 4 we relate halo structure properties to assembly history and environment, identifying the dominating factors that determine halo properties. We also investigate the dependence of halo structure and assembly history on halo environment and halo mass. We summarize our main results in Section 5.

2. Simulation and Halo Quantities

2.1. The Simulation

The *N*-body simulation used here is the ELUCID simulation carried out by Wang et al. (2016) using L-GADGET code, a memory-optimized version of GADGET-2 (Springel 2005). The simulation uses 3072^3 dark matter particles, each with a mass $3.08 \times 10^8\, h^{-1} M_\odot$, in a periodic cubic box of 500 comoving h^{-1}Mpc on a side. The cosmology parameters used are those based on WMAP5 (Dunkley et al. 2009), a ΛCDM universe with density parameters $\Omega_{\rm K,0} = 0$, $\Omega_{\rm M,0} = 0.258$, $\Omega_{\rm B,0} = 0.044$, and $\Omega_{\Lambda,0} = 0.742$, a Hubble constant $H_0 = 100\, h\ {\rm km\ s^{-1}\ Mpc^{-1}}$ with $h = 0.72$, and a Gaussian initial density field with power spectrum $P(k) \propto k^n$ with $n = 0.96$ and an amplitude specified by $\sigma_8 = 0.80$. A total of 100 snapshots, uniformly spaced in $\log(1 + z)$ between $z = 18.4$ and $z = 0$, are taken and stored.

Halos and subhalos with more than 20 particles are identified by the friends-of-friends (FoF; see, e.g., Davis et al. 1985) and SUBFIND (Springel 2005) algorithms. Halos and subhalos among different snapshots are linked to build halo merger trees.[8] Halo virial radius $R_{\rm vir}$ is related to halo mass, $M_{\rm halo}$, through

$$M_{\rm halo} = \frac{4\pi R_{\rm vir}^3}{3} \Delta_{\rm vir} \bar{\rho}, \tag{1}$$

where $\bar{\rho}$ is the mean density of the universe, and $\Delta_{\rm vir}$ is an overdensity obtained from the spherical collapse model (Bryan & Norman 1998). The center of a halo is assumed to be the position of the most bound particle of the main subhalo. Halo mass $M_{\rm halo}$ is computed by summing over all particles enclosed within $R_{\rm vir}$. The virial velocity, $V_{\rm vir}$, is defined as $V_{\rm vir} = \sqrt{G M_{\rm halo}/R_{\rm vir}}$, where G is the gravitational constant. We use halos with masses $\geqslant 10^{10}\, h^{-1} M_\odot$ directly from the simulation, and we use `Monte Carlo`-based merger trees to extend the mass resolution of trees down to $10^9\, h^{-1} M_\odot$ (see Chen et al. 2019).

From the halo merger tree catalog constructed above, we form four samples according to halo mass: S_1 contains 2000 halos with $10^{11} \leqslant M_{\rm halo}/\, h^{-1} M_\odot \leqslant 10^{11.2}$; S_2 contains 1000 halos with $10^{12} \leqslant M_{\rm halo}/\, h^{-1} M_\odot \leqslant 10^{12.2}$; S_3 contains 500

[8] We use only halos and halo merger trees in this paper; subhalos are not included in the samples.

Table 1
Samples Used in Our Analysis

Sample	$N_{\rm halo}$	$M_{\rm halo}/(h^{-1}M_\odot)$	Usage
S_1	2000	$10^{[11,\ 11.2]}$	Samples with constrained halo masses. Used in Section 3.1.
S_2	1000	$10^{[12,\ 12.2]}$	
S_3	500	$10^{[13,\ 13.2]}$	
S_4	500	$10^{[14,\ 14.5]}$	
S_c	10000	$\geqslant 10^{11}$	Mass-limited sample. Used in Section 3.1.
S_c'	2335	$\geqslant 5\times 10^{11}$	Subsample of S_c, with all halo properties well defined. Used in Sections 3.2, 4.1, 4.2.
S_L	94524	$10^{[11,\ 14.6]}$	The "larger" sample for binning statistics. Used in Sections 4.3, 4.4.

Note. The four columns are sample identifier, number of halos, halo mass range, and usage, respectively. The exact sample definitions can be found in Section 2.1.

halos with $10^{13} \leqslant M_{\rm halo}/h^{-1}M_\odot \leqslant 10^{13.2}$; and S_4 contains 500 halos with $10^{14} \leqslant M_{\rm halo}/h^{-1}M_\odot \leqslant 10^{14.5}$. We also construct a complete sample, S_c, which contains 10,000 halos with $M_{\rm halo}/h^{-1}M_\odot \geqslant 10^{11}$ to represent the total halo population. All halos in samples S_1, S_2, S_3, S_4, and S_c are randomly selected from simulated halos at $z = 0$. When halos need to be divided into subsamples according to some properties, the sample size of S_c may be insufficient. In this case, we construct a larger sample by the following steps. Starting from all simulated halos at $z = 0$ with $M_{\rm halo}/h^{-1}M_\odot = 10^{11}$ –$10^{14.6}$, we divide them into mass bins of width 0.3 dex. If the number of halos in a bin exceeds 10,000, we randomly choose 10,000 from them; otherwise all halos in this mass bin are kept. This gives a sample of 94,524 halos, which is referred to as sample S_L. Due to the mass resolution of the ELUCID simulation, some properties of small halos cannot be derived reliably. Whenever these properties are needed, we use another halo sample, S_c', which contains all halos with $M_{\rm halo} \geqslant 5 \times 10^{11}\,h^{-1}M_\odot$ in sample S_c.

Note that halos with recent major mergers may have structural properties that are very different from virialized halos, and including them in our sample will significantly increase the variance of halo properties, thereby affecting the statistics derived from the sample. We use the criteria described in Appendix C to exclude those "unrelaxed" halos.

All of the samples used in this paper are summarized in Table 1. Note that we use only a fraction of all halos in a given mass range available in the simulation to save computational time. We have made tests using larger samples to confirm that the samples we use are sufficiently large to obtain robust results.

2.2. Halo Assembly History

Following the literature (e.g., van den Bosch 2002; Wechsler et al. 2002; Zhao et al. 2003a, 2003b), we define the MAH of a halo as the main-branch mass M_z as a function of redshift z in the halo merger tree rooted in that halo. Based on a theoretical consideration of halo formation (e.g., Zhao et al. 2009), we use the following quantity as the mass variable:

$$s(z) = \sigma(M_0)/\sigma(M_z), \tag{2}$$

where $\sigma(M)$ is the rms of the $z = 0$ linear density field at the mass scale M. Similarly, we use

$$\delta_c(z) = \delta_{c,0}/D(z) \tag{3}$$

as the time variable, 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 $D(z)$ from Carroll et al. (1992). These definitions for the mass and time variables are well motivated by the self-similar behavior of halo formation expected in the Press–Schechter formalism (e.g., Press & Schechter 1974; Mo et al. 2010).

Thus, in our definition, the MAH of a halo is a vector $\boldsymbol{s} = (s(z_1), s(z_2),..., s(z_M))^{\rm T}$, with each of its elements being the main-branch mass at a snapshot in the merger tree.[9] Such a high-dimensional vector is obviously too complex to be useful in characterizing the formation of a halo. To overcome this problem, a common practice is to characterize the full MAH by a set of formation times (e.g., Li et al. 2008). In our analysis, we will use both the formation times and the principal components (PCs; see Section 3) to reduce the dimension of the MAH.

The MAH introduced above only uses the main branch of a halo merger tree, and thus it may potentially lose important information about the formation of a halo. However, our tests including the side branches (e.g., by modeling the assembly history with the progenitor mass distribution, as used in Parkinson et al. 2008) showed that it does not provide important information about halo structural properties. We therefore only use the main-branch assembly history for our analysis.

2.3. Halo Concentration

Dark matter halo profiles are usually modeled by a universal two-parameter form—the Navarro–Frenk–White (NFW) profile (Navarro et al. 1997),

$$\rho_{\rm NFW}(r) = \frac{\delta\rho_{\rm crit}}{(r/r_s)(1 + r/r_s)^2}, \tag{4}$$

where $\rho_{\rm crit} = 3H^2/(8\pi G)$ is the critical density of the universe. This profile is specified by a dimensionless amplitude, δ, and a scale radius, r_s. The scale radius is usually expressed in terms of the virial radius, $R_{\rm vir}$, through a concentration parameter, $c \equiv R_{\rm vir}/r_s$. Since both $\bar{\rho}$ and $\Delta_{\rm vir}$ in Equation (1) are known for a given cosmology, the profile of a halo can be specified by the parameter pair ($M_{\rm halo}$, c), where the halo mass $M_{\rm halo}$ is defined in Section 2.1.

We obtain the concentration parameter of a halo through the following steps (see Bhattacharya et al. 2013):

[9] To avoid confusion, we use log to denote the base ten logarithm; ln to denote the base e logarithm; bold, roman lowercase characters to denote vectors; bold, roman uppercase characters to denote the matrix; and $||\cdot||$ to denote the 2-norm of a vector or matrix.

1. We divide the volume centered on a halo into $N_{\rm r} = 20$ radial bins equally spaced between 0 and $R_{\rm vir}$ (see Section 2.1 for definitions of halo center and radius), and we calculate the mass within each bin i, M_i, using the number of particles in the bin.
2. We compute the mass expected from the NFW profile in this bin, $M_{i,{\rm NFW}}(c)$, assuming a concentration parameter c.
3. We define an objective function, $\chi^2(c)$, to be minimized as

$$\chi^2(c) = \sum_{i=1}^{N_{\rm r}} \frac{[M_i - M_{i,{\rm NFW}}(c)]^2}{M_i^2/n_i}. \tag{5}$$

4. We minimize the objective function to find the best-fit concentration parameter $c_{\rm fit} = {\rm argmin}_c\, \chi^2(c)$.

In what follows, we will drop the subscript "fit" and use c to denote the concentration parameter obtained this way. As tested by Bhattacharya et al. (2013), changing the radius range and binning scheme in the fitting only introduces a negligible difference in c. According to our tests, our results presented in the following sections are also insensitive to such changes.

2.4. Halo Axis Ratio

We model the mass distribution in a dark matter halo with an ellipsoid and use its axis ratio to characterize its shape. Our modeling consists of the following steps (see Macciò et al. 2007):

1. We start from a FoF halo and calculate the spatial position $d\boldsymbol{r}_i$ relative to the center of mass for each particle linked to the halo. The inertia tensor $\mathcal{M}$ of the halo is given by the dyadic of $d\boldsymbol{r}_i$ summing over all of the $N_{\rm p}$ particles in that halo:

$$\mathcal{M} = \sum_{i=1}^{N_{\rm p}} m_i d\boldsymbol{r}_i d\boldsymbol{r}_i^{\rm T}, \tag{6}$$

where m_i is the particle mass.
2. We calculate the eigenvalues $\lambda_{\mathcal{M},i}$ and eigenvectors $v_{\mathcal{M},i} (i = 1, 2, 3)$ of $\mathcal{M}$ and rank the eigenvalues in descending order: $\lambda_{\mathcal{M},1} \geqslant \lambda_{\mathcal{M},2} \geqslant \lambda_{\mathcal{M},3}$. So defined, $v_{\mathcal{M},i}$ gives the axis direction of the inertia ellipsoid, and $a_{\mathcal{M},i} = \sqrt{\lambda_{\mathcal{M},i}}$ gives the length of the corresponding axis.
3. We define the axis ratio $q_{\rm axis}$ of the halo as

$$q_{\rm axis} = \frac{a_{\mathcal{M},2} + a_{\mathcal{M},3}}{2a_{\mathcal{M},1}}. \tag{7}$$

So defined, $q_{\rm axis} = 1$ for a spherical halo, and close to zero if the halo is very elongated along the major axis.

2.5. Halo Spin

Following Bullock et al. (2001), we define the spin parameter of a halo as

$$\lambda_{\rm s} \equiv \frac{\|\boldsymbol{j}\|}{\sqrt{2}M_{\rm halo}R_{\rm vir}V_{\rm vir}}, \tag{8}$$

where $M_{\rm halo}$, $R_{\rm vir}$, and $V_{\rm vir}$ are, respectively, the halo mass, virial radius, and virial velocity defined in Section 2.1. The total angular momentum $\boldsymbol{j}$ is defined as

$$\boldsymbol{j} = \sum_{i=1}^{N_{\rm p}} m_i d\boldsymbol{r}_i \times d\boldsymbol{v}_i, \tag{9}$$

where m_i is the particle mass, and $d\boldsymbol{r}_i$ and $d\boldsymbol{v}_i$ are particle position and velocity vectors relative to the center of mass, respectively. The summation is over all of the $N_{\rm p}$ particles linked in the FoF halo.

2.6. Environmental Quantities

Many definitions can be found in the literature to characterize the environment of a halo at different scales, traced by different objects, and including or excluding the halo itself (see Haas et al. 2012 for a review). Here we define two quantities to describe the environment in which a halo resides: the density contrast as specified by the bias factor, and the tidal tensor. These quantities are computed directly from the N-body simulation.

First we follow Mo & White (1996) to define the halo bias b as the ratio between the halo–matter cross-correlation function and the matter–matter autocorrelation function. For each simulated halo i, we calculate its bias b_i by

$$b_i = \frac{\xi_{{\rm hm},i}(R)}{\xi_{\rm mm}(R)}, \tag{10}$$

where $\xi_{{\rm hm},i}(R)$ is the overdensity centered at halo i at a radius R, and $\xi_{\rm mm}(R)$ is the matter–matter autocorrelation function at the same radius in the simulation at the redshift in question. On linear scales, the bias factor is expected to depend only on halo mass, independent of R. We have checked the values of b on different scales and found that the bias factor is almost constant at $R > 5\ h^{-1}$Mpc. In the following, we compute both $\xi_{{\rm hm},i}$ and $\xi_{\rm mm}$ for R between 5 and 15 h^{-1}Mpc at $z = 0$, and we obtain the corresponding local linear bias b_i for each halo using the above equation.

For the tidal field, we follow Ramakrishnan et al. (2019) and define the halo environment in the following steps: we first divide the simulation box into a sufficiently fine grid (of $N_{\rm cell}^3$ grid points) and compute the density field $\rho(\boldsymbol{x})$ on each grid point using the clouds-in-cell method (Hockney & Eastwood 1988). To describe the environment of a given halo, the density field is smoothed on some scale $R_{\rm sm}$ with a Gaussian kernel. We then compute the potential field $\Phi(\boldsymbol{x})$ by solving the Poisson equation

$$\nabla^2\Phi(\boldsymbol{x}) = 4\pi G\bar{\rho}\delta(\boldsymbol{x}), \tag{11}$$

where $\delta(\boldsymbol{x}) = \rho(\boldsymbol{x})/\bar{\rho} - 1$ is the overdensity and $\bar{\rho}$ the mean density of the universe. Next we obtain the tidal field $\mathcal{T}(\boldsymbol{x})$ through

$$\mathcal{T}(\boldsymbol{x}) = \nabla\nabla\Phi(\boldsymbol{x}). \tag{12}$$

Finally, we solve the eigenvalue problem of the tidal tensor at each grid point to find the eigenvalues $\lambda_1(\boldsymbol{x})$, $\lambda_2(\boldsymbol{x})$, $\lambda_3(\boldsymbol{x})$ (ranked in descending order).

With all these, we obtain four environmental quantities for each halo: the local bias factor and the three eigenvalues of the tidal tensor: λ_i $(i = 1, 2, 3)$. We follow the definition in Ramakrishnan et al. (2019) to define the local tidal anisotropy

Table 2
Summary of Halo Properties Studied in This Paper

Property Type	Notation	Meaning
Assembly History	$\mathrm{PC}_{\mathrm{MAH},i}$	ith PC of MAH
Structure	c	Concentration parameter of the NFW profile
	$\lambda_{\rm s}$	Dimensionless spin parameter
	$q_{\rm axis}$	Axis ratio of inertia momentum
Environment	b	Halo bias factor
	$\alpha_{\mathcal{T}}$	Tidal anisotropy

Note. Detailed descriptions of halo MAH PCs and structural and environmental quantities can be found in Sections 2 and 3.1.

at each halo's position as

$$\alpha_{\mathcal{T}} = \sqrt{q^2}/(1+\delta), \qquad (13)$$

where $q^2 = \frac{1}{2}[(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_1 - \lambda_3)^2]$ is the halocentric tidal shear, and $\delta = \lambda_1 + \lambda_2 + \lambda_3$ is the local overdensity. For our applications, we choose $R_{\rm sm} = 4R_{\rm vir}$, although we have checked that our conclusion does not change significantly by using other values of $R_{\rm sm}$. We use $N_{\rm cell} = 2560$, which is sufficiently fine for computing $\alpha_{\mathcal{T}}$ for the smallest halo in our sample.

In Table 2, we summarize the halo assembly, structure, and environmental properties used in our analysis.

3. Characterizing Halo Assembly History with PCs

Because many parameters can be defined to characterize various aspects of the halo assembly history (e.g., different formation times), it is not feasible to use all of them directly to study the relationship between halo assembly and other halo properties. In this section, we describe a method that can be used to effectively reduce the dimension of the halo assembly history, which has some advantages over using formation times.

We use the PCA described in Appendix A.1 to reduce the dimension of the MAH. In this section we will show the advantages of using such a method in the following two aspects. (1) The PCs capture the most variance among all linear, low-dimension representations and are the best in reducing the reconstruction error. We will demonstrate their power in representing the halo MAH. (2) The PCs are tightly correlated with some formation times widely used in the literature, which helps clarify the types of information PCs contain. In the next section (Section 4), we will show that PCs are strongly correlated with halo structural properties, which helps us understand the origin of such properties. All these indicate that PCs are not only mathematically optimized approximations to the MAH but also have clear physical meanings.

3.1. PCs of Halo Assembly Histories

For a given sample of N halos, the MAHs are represented by $\boldsymbol{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_{\rm N})^{\rm T}$, where $\boldsymbol{s}_i$ is the MAH of the ith halo, as defined in Section 2.2. In simulation, the MAH of a halo is not traced below the mass resolution of the simulation. Halos of different mass therefore may be traced down to different snapshots, resulting in different lengths of the vector $\boldsymbol{s}_i$. For a given sample, we choose a snapshot where 90% of the MAHs can be traced back to this time. MAHs extending beyond this snapshot are truncated, and MAHs that are terminated before this snapshot are padded with 0. In this way, all of the $\boldsymbol{s}_i$ will have the same length, M, suitable for PCA. After applying the PCA to $\boldsymbol{S}$, the $\boldsymbol{s}$ for each halo is transformed into a new coordinate system, producing a new vector that consists of a series of PCs, $\mathbf{pc}_{\rm MAH} = (\mathrm{PC}_{\mathrm{MAH},1}, \mathrm{PC}_{\mathrm{MAH},2}, \ldots, \mathrm{PC}_{\mathrm{MAH,M}})$. In terms of the capability of capturing sample variance and reducing reconstruction errors, PCA is theoretically the optimal linear method. Thus, if the assembly histories of halos have some dominating modes, they are expected to be captured by the PCA.

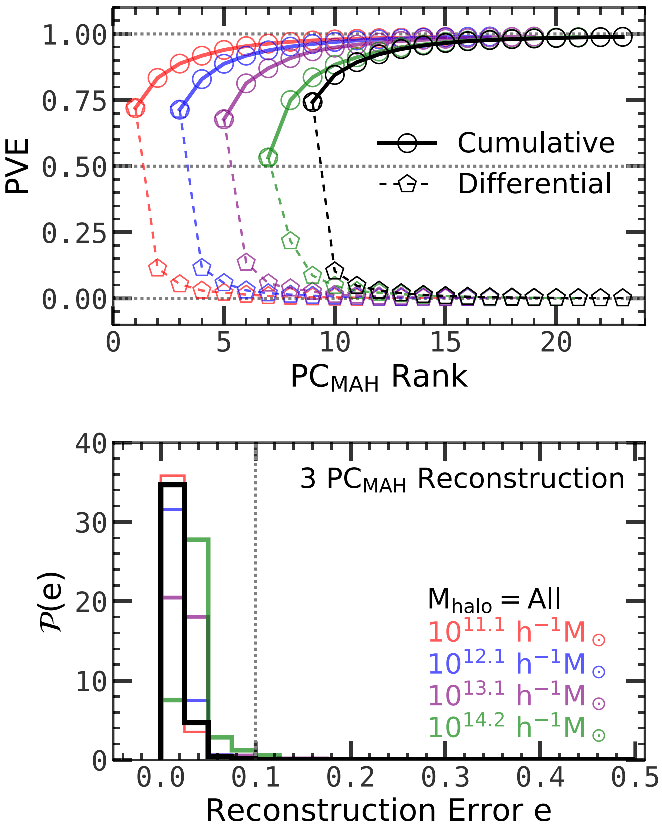

Figure 1. Performance of PCA on a halo MAH. Top: PVEs and their cumulative version (see Appendix A.1 for definitions). Each curve is horizontally offset increasingly by 2 for clarity. Bottom: distribution of reconstruction error e using the first three MAH PCs. In both panels, five samples, S_1, S_2, S_3, S_4, $\mathrm{S}_{\rm c}$, with different halo mass selections (see Section 2.1 and Table 1) indicated in the lower panel, are presented.

The upper panel of Figure 1 shows the PVE and CPVE curves (see Appendix A.1 for definitions) for the PCs of the five samples, S_1, S_2, S_3, S_4, and $\mathrm{S}_{\rm c}$, defined in Section 2.1 (see also Table 1). Since the proportional explained variance (PVE) measures the fraction of sample variance explained by a PC, it is clear that the first several PCs, among all cases, can capture most of the variance in the halo MAH (>80% by using the first three PCs). This demonstrates that a strong degeneracy exists in the MAH of individual halos, suggesting that the assembly history can be described by using only a few eigenmodes. As shown by the CPVE curve, using a single parameter is insufficient to describe the MAH. It can at most explain as much variance as PC_1, which is 55% for the most massive halos ($\approx 10^{14}\,h^{-1}M_\odot$) and 70% for the smallest halos ($\approx 10^{11}\,h^{-1}M_\odot$). In principle, we can add more PCs, which typically leads to better capture of the subtle structures in the MAHs. The lower panel of Figure 1

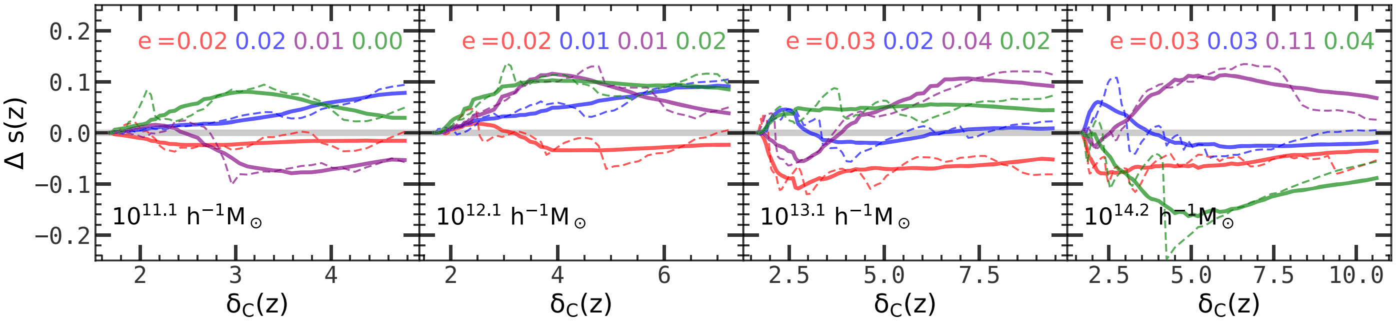

Figure 2. Reconstructed MAHs of example halos (solid) compared with the real ones (dashed). Four panels present halos with different masses indicated in each panel. In each case, the MAHs are reconstructed from the first three PCs. The reconstruction error e for each halo is also presented.

shows the distribution of the error e in each sample when MAHs are reconstructed with the first three PCs (see Appendix A.1 for the reconstruction algorithm). The reconstruction errors are almost all below 10%, demonstrating that the MAHs of halos can be represented well by a small number of PCs.

Figure 2 shows some examples of the reconstructed MAHs using the first three PCs. Here $\Delta s(z) = s(z) - \bar{s}(z)$ is shown for several halos in each halo-mass-constrained sample (see Section 2.1 and Table 1), where $\bar{s}(z)$ is the mean MAH in the sample. The overall shape of the MAH is well captured by the reconstruction, although the MAHs of individual halos are quite diverse. Some fine structures in the MAH, caused by violent changes in the formation history due to merger events, are missed in the reconstruction. They can, in principle, be captured by including more PCs.

The rapidly converged PVE, the sharply peaked distribution of the reconstruction, and the well-reconstructed MAHs of individual halos all indicate that PCs are effective in reducing the dimension of the halo MAH. In the following subsection, we will show the relation between PCs and some widely used halo formation times to gain more physical insights into different PCs.

3.2. PCs versus Halo Formation Times

The diversity of the assembly history shown in Figure 2 indicates that no single parameter can provide a complete description of the MAH. To reflect different aspects of the assembly history, different assembly indicators, for example, formation times, have been defined in the literature. These formation times are physically more intuitive compared with the more abstract PCs, although each of them only provides partial information about the MAH. Here we examine the relations between PCs and a number of halo formation times to gain some physical understanding of the PCs we obtain.

In Appendix B we list the formation times we use and their detailed definitions. Because of the large number of formation times and the potential nonlinear pattern in their relations with PCs, RF is an ideal tool for this task. In Appendix A.2 we describe the RF algorithm in detail. The two important outputs from the RF analysis are (1) the fraction of explained variance, R^2, and (2) the feature importance, $\mathcal{I}(x)$, for any predictor variable x. These two quantities measure the performance of the regression and the contribution from each predictor variable in explaining the target variable, respectively. Figure 3 shows how different formation times contribute to the diversity of the halo MAH. Here we use sample S_c' in which all of the formation times are well defined, as described in Section 2.1 (see also Table 1), to regress the first three MAH PCs on all of the formation times. The performance, R^2, and the contribution $\mathcal{I}(x)$ of each formation time x to each PC are plotted. For all PCs, the contribution from $z_{\mathrm{mb},0.04}$ is the most dominant, while $z_{\mathrm{mb},1/2}$ is also significant. However, the importance of both decreases in higher order PCs. Interestingly, the last major merger redshift, z_{lmm}, which contributes little to $\mathrm{PC}_{\mathrm{MAH},1}$, is increasingly important as the PC order increases. For $\mathrm{PC}_{\mathrm{MAH},3}$, the contribution from z_{lmm} is comparable to those from the other two formation times, $z_{\mathrm{mb},0.04}$ and $z_{\mathrm{mb},1/2}$. Since mathematically higher order PCs are capable of capturing more subtle patterns in the feature space, this behavior of the importance curves means that assembly variables, such as $z_{\mathrm{mb},0.04}$ and $z_{\mathrm{mb},1/2}$, mainly describe the low-order, overall patterns of the halo assembly history, while major mergers are an important factor in producing the fine structure in MAHs.

The contribution curves in Figure 3 automatically suppress variable competition where multiple degenerate feature variables compete in the prediction contribution to the target variable. In the RF, if two variables compete but one of them is slightly better, then the split algorithm prefers the better one and gives it a higher importance value $\mathcal{I}(x)$. To show this more clearly, Figure 4 plots the correlation between $z_{\mathrm{mb},0.04}$, $z_{0.02,1/2}$, and the first two MAH PCs. Strong and nearly linear correlations between $\mathrm{PC}_{\mathrm{MAH},1}$ and $z_{\mathrm{mb},0.04}$ and between $\mathrm{PC}_{\mathrm{MAH},1}$ and $z_{0.02,1/2}$ are seen, indicating that the variances in both formation times contribute significantly to the variance in the MAH of the halos. Inspecting the contours, one can also see that the correlation between $\mathrm{PC}_{\mathrm{MAH},1}$ and $z_{\mathrm{mb},0.04}$ appears stronger. The larger contribution value to $\mathrm{PC}_{\mathrm{MAH},1}$ from $z_{\mathrm{mb},0.04}$ than from $z_{0.02,1/2}$ shown in Figure 3 validates the strength of the correlation.

Figure 4 also demonstrates that the description provided by a single halo formation time is incomplete. The large scatter seen in the contours of $\mathrm{PC}_{\mathrm{MAH},1}$ versus formation times means that the direction of the largest scatter in the MAH space is not fully aligned with the scatter caused by formation time, and that other parameters must also contribute to the distribution of halos in the MAH space. Moreover, compared to $\mathrm{PC}_{\mathrm{MAH},1}$, $\mathrm{PC}_{\mathrm{MAH},2}$ has a much weaker correlation with the halo formation times. Thus, the information contained in the second PC of MAH, which contributes 10%–20% to the total variance as seen from the upper panel of Figure 1, is almost entirely missed when a single formation time is used to predict the MAH.

The degeneracy and completeness problems can, in principle, be overcome if we use PCs to describe the assembly history, as PCs are linearly independent of each other. Typically, the use of a small number of low-order PCs can solve most of the problems associated with the regression of halo MAHs. If more subtle

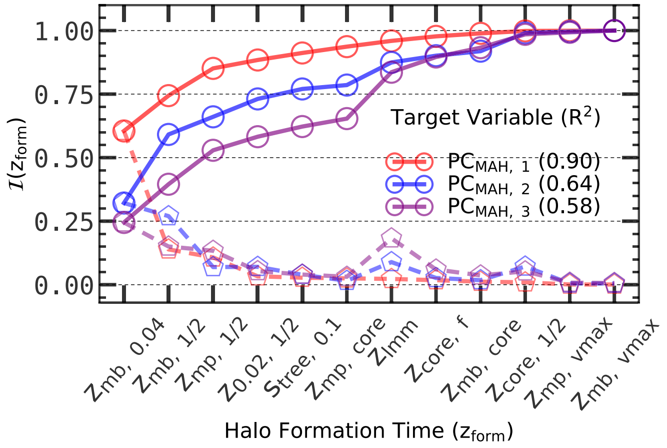

Figure 3. Performances (R^2) and contributions ($\mathcal{I}$) using formation times to predict the first three $\rm PC_{MAH}$s based on RF regressors (see Appendix A.2 for the RF model). Dashed: importance values $\mathcal{I}$ of predictors. Solid: the cumulative of $\mathcal{I}$. The overall performances, R^2, are shown in the parentheses in the legend. Halos are taken from the mass-limited sample S_c' with $M_{\rm halo} \geqslant 5 \times 10^{11}\,h^{-1}M_\odot$ (see Section 2.1 and Table 1).

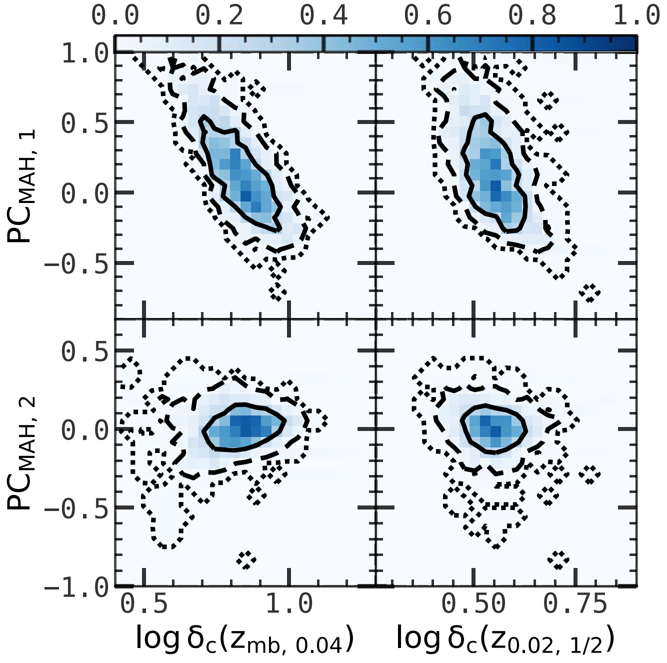

Figure 4. Relation of halo formation time $z_{\rm mb,0.04}$ (left) or $z_{0.02,1/2}$ (right) to the first two MAH PCs. Solid, dashed, and dotted contours cover 1, 2, 3σ regions, respectively. Normalized 2D histograms are color-coded according to the color bar. Halos are taken from the mass-limited sample S_c' with $M_{\rm halo} \geqslant 5 \times 10^{11}\,h^{-1}M_\odot$ (see Section 2.1 and Table 1).

information is needed, one can always add more PCs without introducing too much degeneracy into the problem.

4. Relating Halo Structure to Assembly History and Environment

In this section, we investigate how halo structural properties are related to halo assembly history and environment. First, we will examine which of the assembly history indicators correlates the best with halo structural properties. Second, we will show to what extent the halo structure can be explained by assembly history and environment, and we answer the question whether there is a single dominating predictor for halo structure or if the predictors are degenerate in predicting halo structure. Third, we will show that our conclusion is valid even when halo mass is fixed. Finally, we revisit the problem of assembly bias, aiming to identify environment–assembly pairs of strong correlation.

4.1. Halo Structure versus Assembly History

It is well known that halo concentration is correlated with halo MAH (see e.g., Navarro et al. 1997; Jing 2000; Wechsler et al. 2002; MacCiò et al. 2008; Zhao et al. 2003a, 2003b, 2009). However, it is still unclear which single assembly parameter best predicts the concentration, and whether combinations of multiple parameters can improve the prediction precision. The same problem exists when we consider other halo structural properties, for example, the axis ratio $q_{\rm axis}$ and the spin parameter λ_s.

The difficulties involved here arise from the high dimension of the feature space, the degeneracy or correlation among predictors, a possible nonlinear effect from predictors to target, and the "bias–variance" trade-off in choosing model complexity. Again, `Random Forest` can be used to tackle these problems (see Appendix A.2). To this end, we build the regressor $y = {\rm RF}(\boldsymbol{x})$, where y is one of the three structural properties: c, $q_{\rm axis}$ or λ_s, and $\boldsymbol{x}$ are halo assembly indicators, either the first 10 MAH PCs, or all formation times. We also include halo mass in the predictor variables, because it is treated as one of the major parameters in many halo-related problems. All these regressors are built based on sample S_c' (see Section 2.1 and Table 1) in which all formation times are well defined for all halos.

The outputs of RF regressions, including the performance, R^2, and the contribution $\mathcal{I}(x)$ from each predictor variable, are shown in Figure 5. As one can see from the red curves, when a large number of predictors are used, the upper limit in the prediction of c using assembly history is about 65%. This indicates that the concentration parameter c of a halo is largely determined by its assembly history. About 35% of the variance is still missing if one uses only the mean relation to predict c from assembly history. Furthermore, as seen from the left panel, the first MAH PC is by far the most important, accounting for about 67% of the total information provided by the MAH. As a comparison, in the right panel, the combination of the three formation times, $z_{\rm mb,1/2}$, $z_{\rm mp,1/2}$, and $z_{\rm mb,0.04}$, contains about the same amount of information as $\rm PC_{MAH,1}$. Thus, if a single parameter is to be adopted as the predictor of c, $\rm PC_{MAH,1}$ is the preferred choice. We have also tried to combine halo formation times and PCs of the MAH as predictors, and we found that the overall performance changes little, indicating that the MAH PCs dominate the information content about halo concentration.

For $q_{\rm axis}$ and λ_s, the upper-limit performances R^2, achieved by either PCs or formation times, are about 15% and 20%, respectively, much worse than c. No single variable seems to dominate the contribution, as indicated by the long and low tail in the $\mathcal{I}(x)$ plot. These suggest that the axis ratio and spin can be affected by many factors related to halo mass assembly, but the effects are all small. The similarity in behavior between λ_s and $q_{\rm axis}$ suggests that these two quantities may share parts of their origins. Indeed, as we will show below (Section 4.2), λ_s and $q_{\rm axis}$ are strongly correlated, and both show a strong correlation with the anisotropy of the local tidal field.

Figure 6 shows the distribution of halos in the $\rm PC_{MAH,1}$–structural parameter space. There is a strong trend that halos assembled late (large $\rm PC_{MAH,1}$) tend to be less concentrated,

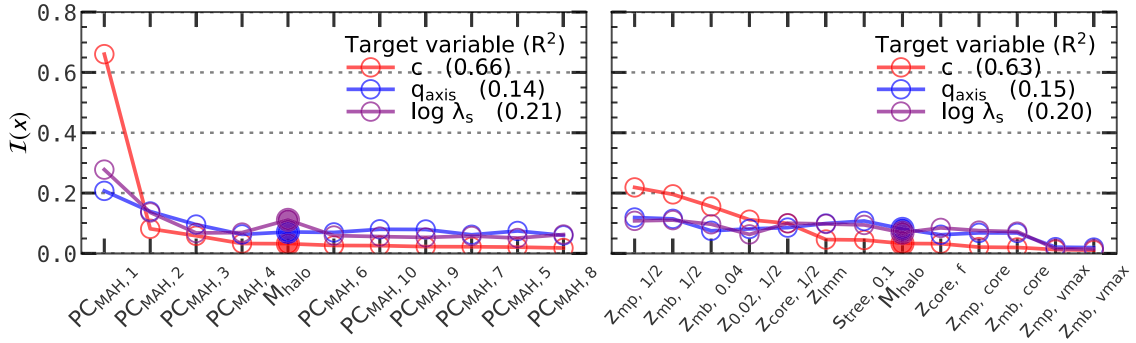

Figure 5. Contributions $\mathcal{I}(x)$ in regressing halo concentration c (red), shape parameter q_{axis} (blue), or spin parameter $\log \lambda_{\text{s}}$ (purple) on halo assembly history variables: the first 10 PCs of MAH (left panel) or formation times (right panel). Halo mass is also included as a predictor variable and is represented by filled symbols. For each case, R^2 indicates the overall performance of the regression. Halos are taken from the mass-limited sample S_{c}' with $M_{\text{halo}} \geqslant 5 \times 10^{11}\, h^{-1}M_{\odot}$ (see Section 2.1 and Table 1).

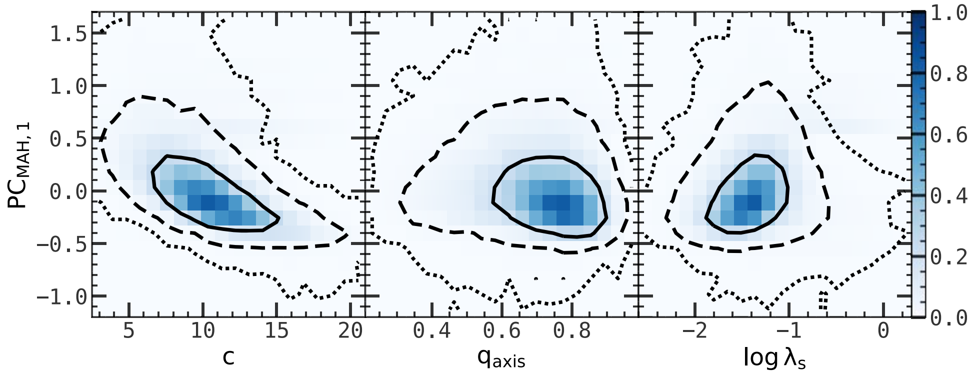

Figure 6. Relation of halo concentration c (left), shape parameter q_{axis} (central), or the spin parameter $\log \lambda_{\text{s}}$ (right) and the first PC of the halo MAH. In each panel, the solid, dashed, and dotted contours cover the 1, 2, and 3σ regions, respectively. The normalized 2D histograms are color-coded according to the color bar. Halos are taken from the mass-limited sample S_{c}' with $M_{\text{halo}} \geqslant 5 \times 10^{11}\, h^{-1}M_{\odot}$ (see Section 2.1 and Table 1).

and a weak trend that such halos tend to be more elongated and spin faster. These are all consistent with the output from the RF regressors and verify that the small contributions from assembly indicators to λ_{s} and q_{axis} are produced by the large scatter, rather than by variable competitions.

4.2. Environmental Effect and Spin–Shape Interaction

We now add environmental predictors to the regression of the structural properties. To quantify the effects of variable competition, we adopt a commonly used approach called "growing," where a series of regressors is built up with an increasing number of predictors. (The approach is called "pruning" if the series runs reversely). Whenever there is a tight correlation between an added predictor and the predictors already used, competition will show up as changes in their importance values, $\mathcal{I}(x)$, but the overall performance, R^2, will not be changed significantly by the addition.

As demonstrated in Section 4.1, among all halo assembly indicators, the first PC of the halo MAH is the dominating indicator for c. Even for q_{axis} and λ_{s}, it is still the most important although less dominating. We therefore start by building a `Random Forest` regressor $y = \mathrm{RF}_1(\mathrm{PC}_{\mathrm{MAH},1}, M_{\text{halo}})$, where y is one of the three structural quantities. Again, the inclusion of M_{halo} is motivated by the fact that halo mass is traditionally considered as one of the most important quantities distinguishing halos. We also calculate the performance R_1^2 as well as the contributions $\mathcal{I}_1(x)$ of the regressor. We then add the two environmental quantities, the bias factor b and the tidal anisotropy $\alpha_{\mathcal{T}}$, to build a second regressor, $y = \mathrm{RF}_2(\mathrm{PC}_{\mathrm{MAH},1}, M_{\text{halo}}, b, \alpha_{\mathcal{T}})$, and obtain the corresponding R_2^2 and $\mathcal{I}_2(x)$.

The left panel of Figure 7 shows the contribution, $\mathcal{I}(x)$, of these two regressors for each of the three structure properties described above, with the values of R^2 indicated, using sample S_{c}' (see Section 2.1 and Table 1). To estimate the uncertainty in the results, we generate 100 random subsamples, each consisting of half of the halos randomly selected from the original sample S_{c}' without replacement. The errors of R^2 and $\mathcal{I}(x)$ are then estimated as the standard deviations among these subsamples. The reason we do not use the standard bootstrap technique is that the sampling with replacement will lead to an artificially large R^2 in the `Random Forest` regressor, because the repeated data points may appear in both the training set that shapes the decision trees and the out-of-bag (OOB) set (see Appendix A.2) that is used to test the performance, making the performance overestimated.

In the case of c, the inclusion of environment only increases R^2 from 0.56 to 0.57, and the importance value $\mathcal{I}(x_{\text{env}})$ is below 0.1. These two results indicate that environment has little impact on halo concentration c, that the concentration c is well determined by the first PC of the MAH, and that the weak dependence of c on environment is mainly through the dependence of halo assembly on environment (assembly bias). These are consistent with the finding of Lu et al. (2006) that the density profiles of individual halos can be modeled accurately from their MAHs. These are also consistent with the result obtained with the Gaussian process regression by Han et al. (2019), who found that the dependence of halo bias on halo

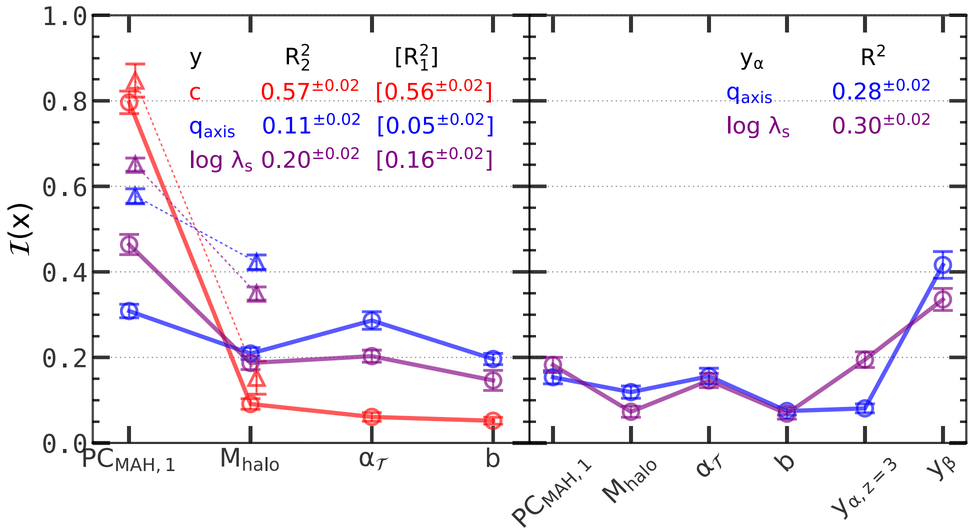

Figure 7. Left: contributions $\mathcal{I}(x)$ from various predictors x to halo structural quantities y: concentration c (red), shape parameter $q_{\rm axis}$ (blue), or the spin parameter $\log\lambda_{\rm s}$ (purple). Solid lines connecting circles show the results when using the first MAH PC $\rm PC_{MAH,1}$, halo mass $M_{\rm halo}$, tidal anisotropy $\alpha_{\mathcal{T}}$, and bias factor b as predictors. Dashed lines connecting triangles show the results when using only $\rm PC_{MAH,1}$ and $M_{\rm halo}$. The overall performances R_2^2 (with environment) and R_1^2 (without environment) are also indicated in the panel. Right: similar to the left panel, except that we add the initial condition of each structural property at redshift $z = 3$ and a β-structure parameter y_β into the predictors. The target variable, now denoted as y_α, is either the shape parameter $q_{\rm axis}$ (blue) or the spin parameter $\log\lambda_{\rm s}$ (purple). For $q_{\rm axis}$, the y_β is $\log\lambda_{\rm s}$, and for $\log\lambda_{\rm s}$, the y_β is $q_{\rm axis}$. The overall performance R^2 is also indicated in the panel. In both panels, halos are taken from the mass-limited sample $\rm S_c'$ with $M_{\rm halo} \geqslant 5 \times 10^{11}\,h^{-1}M_\odot$ (see Section 2.1 and Table 1). The error bars are calculated using 100 half-size resamplings without replacement.

concentration is mainly through the dependence of the bias on formation time.

For $q_{\rm axis}$, the R^2 is doubled, from 0.05 to 0.11, when environment factors are included. From the importance curve, $\mathcal{I}(x)$, one can see that these environment factors take away about half of the contribution from the halo mass and $\rm PC_{MAH,1}$. These results together suggest that environment can affect halo shape significantly and is at least as important as halo mass and the PCs of the MAH.

For the spin parameter, the value of R^2 increases from 0.16 to 0.20 when environment factors are included. The increase, 0.04, is about 20%, suggesting that these environment factors do have a sizable effect on halo spin. The contribution, $\mathcal{I}(b) + \mathcal{I}(\alpha_{\mathcal{T}}) = 35\%$, is larger than the 20% they contribute to R^2, suggesting that some of the contribution is actually taken from the halo assembly history and halo mass. Thus, when interpreting the dependence of halo spin on environment, one should remember that part of it may actually come from its degeneracy with halo assembly history.

Another difference between c and the other two structural parameters is in their values of R^2. Both $q_{\rm axis}$ and $\log\lambda_{\rm s}$ have fairly small R^2, much smaller than 50%. This indicates that the major contributors to these two structural parameters are not yet found. In general, factors that can affect halo structural properties can be classified into three categories: the initial conditions of halos, the intrinsic properties of halos (e.g., $M_{\rm halo}$, $\rm PC_{MAH,1}$), and halo environment (e.g., $\alpha_{\mathcal{T}}$ and b). The interaction of halos with their environment depends not only on the environment, but also on halo properties. For example, because a halo is coupled to the local tidal torque only through its quadrupole, we expect that the spin and shape of a halo are correlated. Motivated by this and the similar behavior of the shape and spin parameters revealed in Section 4.1, we add the spin parameter into the predictors of the shape, and vice versa. We denote the target structure parameter as y_α and the added structure parameter y_β. In addition, we also consider the "initial condition" of a halo by tracing all of the halo particles back to redshift $z = 3$ and using these particles to calculate the corresponding shape and spin parameters. This quantity for y_α, denoted by $y_{\alpha,z=3}$, is also added into the set of predictors. The right panel of Figure 7 shows the result when this set of variables are used to predict halo structures. Among all the predictor variables, y_β is the most important for y_α, indicating that $q_{\rm axis}$ and $\log\lambda_{\rm s}$ are strongly correlated. For $\lambda_{\rm s}$, its initial value also matters, ranked as the second important predictor. The performance, R^2, for both $\lambda_{\rm s}$ and $q_{\rm axis}$ is now boosted significantly, to ∼0.3. However, even in this case, R^2 is still less than 50%, meaning that the causes of the main parts of the variances in both $\lambda_{\rm s}$ and $q_{\rm axis}$ are still to be identified. This also indicates that, unlike the concentration parameter c, which is determined largely by halo assembly, $\lambda_{\rm s}$ and $q_{\rm axis}$ may depend on the details of the initial conditions, assembly, and environment. Morinaga & Ishiyama (2020) provided a possible scenario where accretion from filaments may partly account for halo shape and orientation. However, the large scatter and weak trend in their results indicate that the driving factor of halo shape and orientation is still missing. All these suggest that many nuanced factors can contribute to the variances of $\lambda_{\rm s}$ and $q_{\rm axis}$. Models for distributions of $\lambda_{\rm s}$ and $q_{\rm axis}$ have to take into account these nuances by assuming some random processes, such as a normal or log-normal process according to the central-limit theorem.

4.3. Dependence on Halo Mass

The regressors for halo structure (presented in Sections 4.1 and 4.2) are all built using the mass-limited sample. The model training processes and performance measurements are thus dominated by low-mass halos, which are more abundant. However, halos with different masses may have different properties. For example, the halo assembly bias, as reflected by the correlation between halo formation time and the bias factor, is found to be significant only for low-mass halos (e.g., Gao et al. 2005; Gao & White 2007; Li et al. 2008). It is,

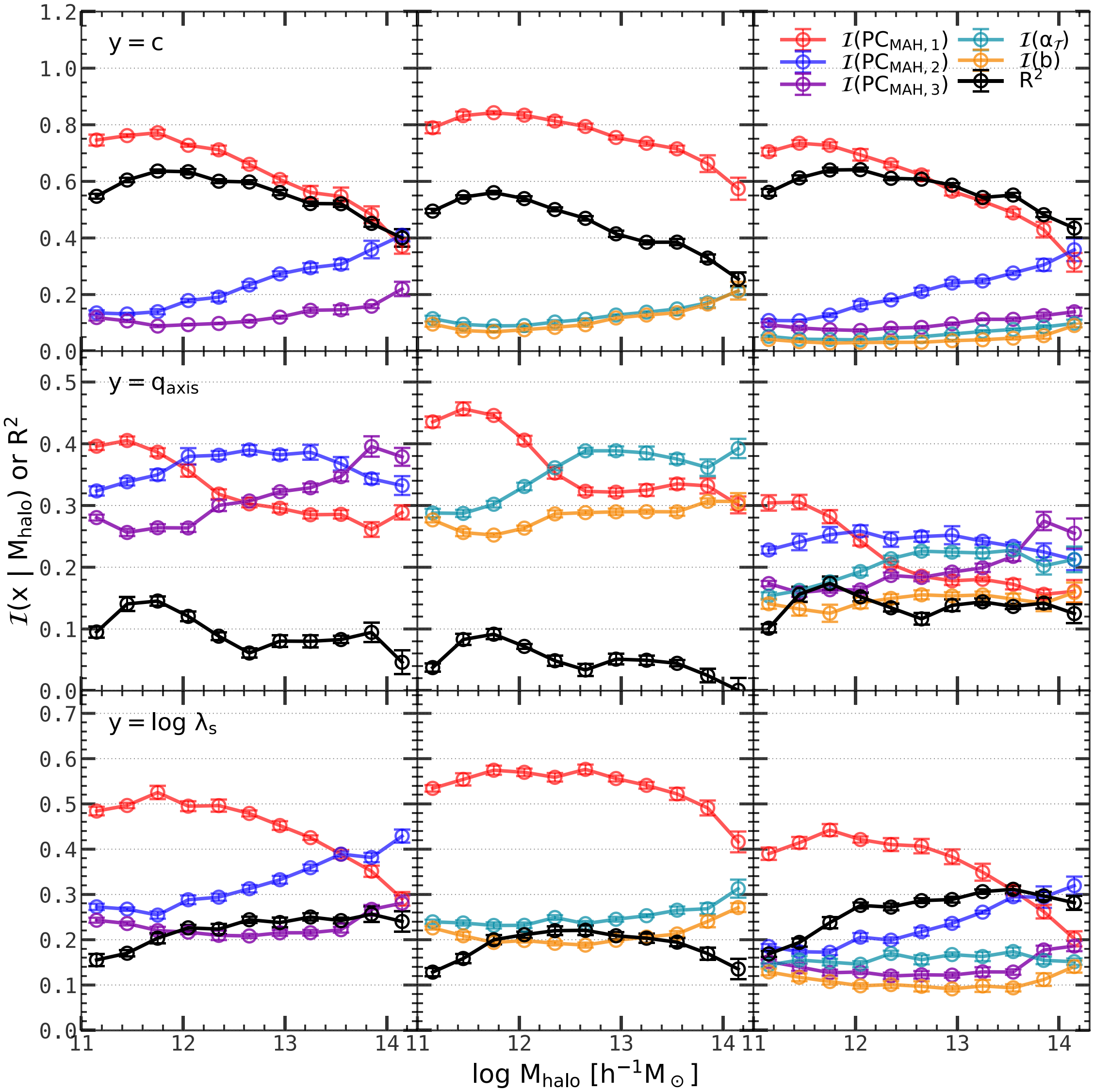

Figure 8. Contributions $\mathcal{I}(x)$ from different predictor variables x to the halo structural properties y for halos of different masses. The upper, central, and bottom rows show the results for structural properties $y = c$, $q_{\rm axis}$, and $\log\lambda_{\rm s}$, respectively. The left, middle, and right columns show the regressors built with predictor variables $\boldsymbol{x} = ({\rm PC}_{\rm MAH,1}, {\rm PC}_{\rm MAH,2}, {\rm PC}_{\rm MAH,3})$, $({\rm PC}_{\rm MAH,1}, \alpha_{\mathcal{T}}, b)$, and$({\rm PC}_{\rm MAH,1}, {\rm PC}_{\rm MAH,2}, {\rm PC}_{\rm MAH,3}, \alpha_{\mathcal{T}}, b)$, respectively. Contributions from different predictor variables x are shown with different colors, as indicated in the upper right panel. The overall performance R^2 for each regressor is represented by the black symbols. The error bars are obtained by 10 half-size resamplings without replacement. Halos are those in the large sample S_L (see Section 2.1 and Table 1).

therefore, interesting to see how the structural properties of massive halos depend on assembly and environment, and how environmental and assembly effects on these halos are related to each other.

Here we quantify such a halo mass dependence by building RF regressors for subsamples of a given halo mass, using the large sample S_L (see Section 2.1 and Table 1). For halos with a given mass, $M_{\rm halo}$, and for each of the three structural properties, $y = c$, $q_{\rm axis}$, and $\log\lambda_{\rm s}$, we build three forest regressors $y = {\rm RF}(\boldsymbol{x}|M_{\rm halo})$ with different sets of predictor variables $\boldsymbol{x}$:

1. $\boldsymbol{x} = ({\rm PC}_{\rm MAH,1}, {\rm PC}_{\rm MAH,2}, {\rm PC}_{\rm MAH,3})$, the first three PCs of the halo MAH;
2. $\boldsymbol{x} = ({\rm PC}_{\rm MAH,1}, \alpha_{\mathcal{T}}, b)$, the first PC of the halo MAH and environmental parameters;
3. $\boldsymbol{x} = ({\rm PC}_{\rm MAH,1}, {\rm PC}_{\rm MAH,2}, {\rm PC}_{\rm MAH,3}, \alpha_{\mathcal{T}}, b)$, the first three PCs of the MAH plus the environmental parameters.

The reason for including ${\rm PC}_{\rm MAH,2}$ and ${\rm PC}_{\rm MAH,3}$ is that the MAHs of massive halos may be more complicated than low-mass ones, and high-order PCs may be needed to capture the more subtle components in their MAHs.

Figure 8 shows the contribution curves and performances of regressors $y = {\rm RF}(\boldsymbol{x}|M_{\rm halo})$ for different halo structural properties, y, using different predictor variables, $\boldsymbol{x}$, and for halos of different masses. In the case where $\boldsymbol{x}$ is the first three MAH PCs (panels in the left column), the ${\rm PC}_{\rm MAH,1}$ is always the most important for the structures of low-mass halos. However, as the halo mass increases, the importance of ${\rm PC}_{\rm MAH,1}$ decreases and eventually is taken over by higher order PCs (the second or the third). Massive halos may have more diverse accretion histories (see, e.g., Obreschkow et al. 2020, who found that tree entropy increases with halo mass), so their structural properties may also be more complex. By using higher order PCs, the complex formation history can be captured so that a better prediction for halo structure properties can be achieved.

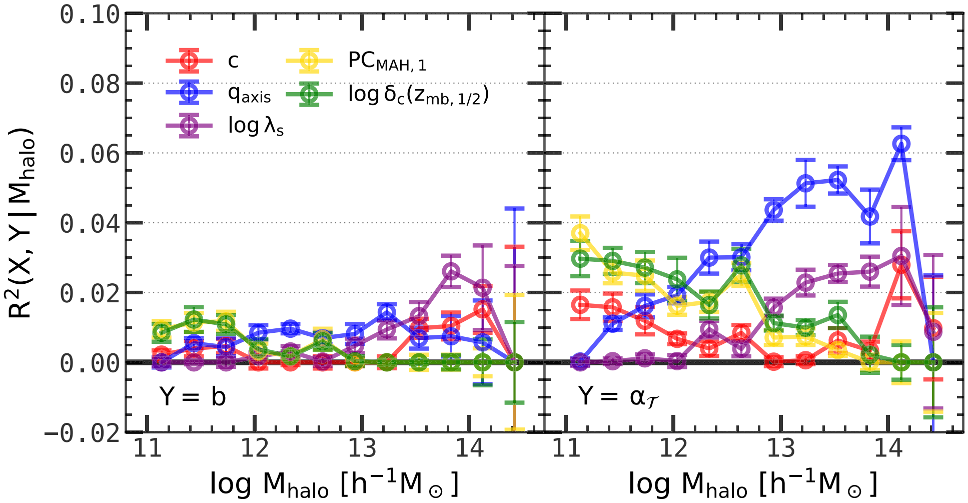

Figure 9. Random Forest regression performance $R^2(X, Y|M_{\rm halo})$ of pairs of variables (X, Y) for halos with a given mass, where Y is an environmental quantity (left panel: halo bias factor *b;* right panel: tidal anisotropy parameter $\alpha_{\mathcal{T}}$), and X is either a structural quantity or an assembly history quantity, as represented with different colors. The error bars of R^2 are obtained by 10 half-size resamplings without replacement. Halos are taken from the large sample S_L (see Section 2.1 and Table 1).

As discussed in Section 4.2, for the total halo population, environmental effects are different for the three halo structural properties. Similar conclusions can be reached for halos of a given mass. The middle and right columns in Figure 8 show the results for regressors that combine the MAH and environment as predictors. For the halo concentration, the PCs of MAH always outperform environmental quantities, although there is a slight increase in $\mathcal{I}(x)$ for environment quantities at the high-mass end. Compared to the regressor with only MAH PCs (upper left panel), the performance R^2 including the two environmental properties (upper right panel) only increases slightly, indicating again that the environmental effect on halo concentration is mainly through the dependence of halo MAH on environment.

The environmental effect on the shape parameter, $q_{\rm axis}$, is totally different. As seen from the contribution curves, the environment is as important as MAH, and including the environment variables increases R^2 significantly. This implies that the environmental effect on the halo shape parameter is important, and that the effect is not degenerate with that of the MAH. The environmental contribution to the spin parameter, λ_s, is intermediate, larger than that to the concentration parameter but smaller than that to the shape parameter. The value of R^2 after including environment variables increases, but less significantly than in the case of the shape parameter. This suggests that the environmental effect does contribute to halo spins, but part of the contribution is taken from the assembly.

4.4. Halo Assembly Bias

As a final demonstration of the application of the Random Forest regressor, we show how assembly parameters correlate with environment for halos of a given mass. Such a correlation is usually referred to as the halo assembly bias. The purpose here is to identify the best correlated pair of variables (X, Y) at a given halo mass, where X is an assembly property and Y is an environmental property. The method to measure the correlation strength is straightforward. First, we bin halos in sample S_L (see Section 2.1 and Table 1) into subsamples according to the halo mass. Within each subsample, we build a Random Forest regressor for each pair of variables (X, Y). The value of $R^2(X, Y|M_{\rm halo})$ then provides a measurement of the correlation strength between the two quantities. Figure 9 shows the results for different cases, where the environmental quantity Y is either the bias factor b or the tidal anisotropy parameter $\alpha_{\mathcal{T}}$, and X is either $\rm PC_{MAH,1}$ or $\log\delta_c(z_{\rm mb,1/2})$. As one can see, the dependence of $\log\delta_c(z_{\rm mb,1/2})$ on b is present only for halos with $M_{\rm halo} < 10^{13}h^{-1}M_\odot$ and totally absent for more massive ones. The values of R^2 between b and the two assembly properties are both smaller than 2%, indicating that the correlation between b and assembly history is weak. These results are consistent with those obtained previously (e.g., Gao et al. 2005; Gao & White 2007; Li et al. 2008): the assembly bias is significant only for low-mass halos, and one has to average over a large number of halos to detect the weak trend.

For comparison, we also build regressors between structure properties (c, λ_s, and $q_{\rm axis}$) and b for halos of a given mass. The results are shown in the left panel of Figure 9. Clearly, the dependence of these properties on b is also weak. The results are consistent with those obtained previously by Mao et al. (2018), who found that b is better correlated with c and λ_s than with assembly properties for massive halos.

In a recent paper, Ramakrishnan et al. (2019) showed that the tidal anisotropy parameter, $\alpha_{\mathcal{T}}$, is a good variable that correlates well with many structural properties. We present the correlation between $\alpha_{\mathcal{T}}$ and other halo quantities in the right panel of Figure 9. It is clear that $\alpha_{\mathcal{T}}$ shows a better correlation with halo intrinsic properties than the bias factor, as indicated by the larger values of R^2. In particular, the correlations between the assembly properties ($\delta_c(z_{\rm mb,\,1/2})$ and $\rm PC_{MAH,1}$) and $\alpha_{\mathcal{T}}$ are significant, except at the very massive end. In Section 4.2, we demonstrate that part of the contribution from the environment to the structural properties is produced by the degeneracy between the environment and the halo assembly history. The strong relation between $\alpha_{\mathcal{T}}$ and assembly history is a proof of this degeneracy. Note that $q_{\rm axis}$ shows a strong correlation with $\alpha_{\mathcal{T}}$ for massive halos, indicating that the local tidal field plays an important role in determining the shape of the halo.

As mentioned above, the small values of R^2 between assembly and environment imply that assembly bias is a weak

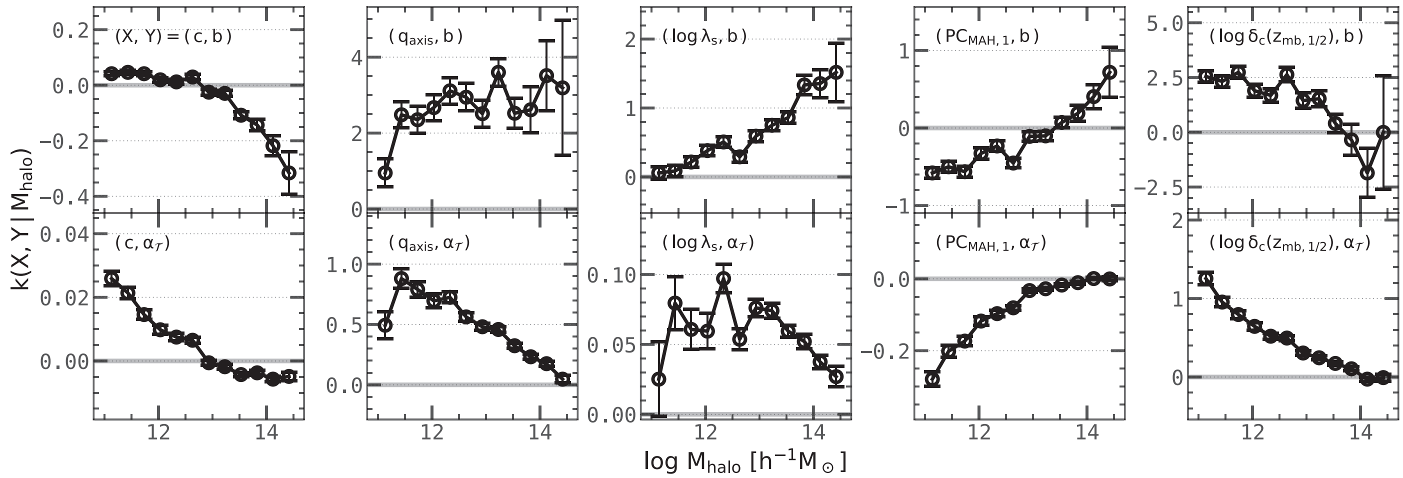

Figure 10. Linear regression slope $k(X,\, Y|M_{\rm halo})$ of pairs of variables $(X,\, Y)$ for halos with a given mass, where Y is an environmental quantity and X is either a structural quantity or an assembly history quantity. Each panel is for a pair of $(X,\, Y)$. The error bars are standard errors estimated from a linear regression model. Halos are taken from the large sample S_L (see Section 2.1 and Table 1).

relation compared to the variance. Detecting such a bias thus needs the average over a large number of halos. To obtain the mean trend in the bias relation, we can build linear regression models between pairs of variables and use the slopes of the regression, k, to represent the mean trend between the two variables in question. Figure 10 shows the slopes of halo properties with b and $\alpha_{\mathcal{T}}$ for halos with different masses. Given any two variables, a larger absolute value of k means a more significant linear correlation of the two variables for halos of a given mass. The relations between halo environment and different halo properties show different trends with halo mass. For example, the slope $k(\log\delta_{\rm c}(z_{\rm mb,1/2}),\, b|M_{\rm halo})$ is significant only at the low-mass end, while $k({\rm PC}_{\rm MAH,1},\, b|M_{\rm halo})$ is significant at the low-mass end and becomes less significant at the high-mass end, suggesting that halo assembly bias is more important for halos of lower mass. In contrast, the slopes $k(c,\, b|M_{\rm halo})$, $k(q_{\rm axis},\, b|M_{\rm halo})$, and $k(\log\lambda_{\rm s},\, b|M_{\rm halo})$ are all more significant at the high-mass end. As discussed in Wang et al. (2011), this halo mass dependence is a result of the competition between the environmental effect and the self-gravity of the halo. For example, a higher density environment not only provides more material for halos to accrete, but also is the location of a stronger tidal field that tends to suppress halo accretion. Depending on the halo mass, one of the two effects dominates. For example, for lower mass halos where self-gravity is weaker, the local tidal field may be more effective in making them more elongated and spin faster, and in preventing them from accreting new mass, so as to make their mass assembly earlier and concentration higher. This is indeed what can be seen from the lower panels of Figure 10. We note, however, that the physical units of the different curves in these panels are not the same, so these curves cannot be compared directly with one another. Note also that a significant value of k does not necessarily imply a significant R^2 in Figure 9, and vice versa, because a linear model may not represent faithfully the data of nonlinear relations. In addition, the value of R^2 depends not only on the value of k, but also on the absolute value of the variance in the regressed sample.

5. Summary and Discussion

In this paper, we have used the ELUCID N-body simulation to relate the structural properties of dark matter halos to their assembly history and environment. Our analysis is based on the PCA and the RF regressor. Our main results and their implications can be summarized as follows.

First, PCA is a simple yet effective tool for reducing the dimension of the halo MAH, and it is preferred over formation times in characterizing the halo MAH. It has the following three major advantages (see Sections 3 and 4.1):

1. PCs are complete and linearly independent. The first three PCs can already explain more than 80% of the variance of the halo MAH, with a reconstruction error <10%.
2. The PCs of the MAH have clear physical meanings. The lower order PCs, such as the first PC, are tightly related to the (halo formation) times when a halo forms fixed fractions of its current mass, such as $z_{\rm mb,1/2}$, $z_{\rm mb,0.04}$, and so on. Higher order PCs, on the other hand, are related to more subtle events, such as the presence of major mergers.
3. PCs are the best, among all assembly indicators, to explain halo structure. The first PC, among all assembly indicators, accounts for about 67%, 20%, and 28% of the variance in halo concentration, shape parameter, and spin parameter, respectively.

Second, the dependence on assembly and environment is quite different for the three halo structural properties (see Section 4.2). About 60%, 10%, and 20% of the variances in c, $q_{\rm axis}$, and $\lambda_{\rm s}$, respectively, can be explained by four predictors: ${\rm PC}_{\rm MAH,1}$, $M_{\rm halo}$, $\alpha_{\mathcal{T}}$, and b. Halo concentration is dominated by the first PC of the MAH, with the contributions from other factors negligible. For $q_{\rm axis}$ and $\lambda_{\rm s}$, there is no single property of assembly that is dominating. The environment has significant effects on these two structural parameters, but its effect on $\lambda_{\rm s}$ is degenerate with the assembly history. The correlation between $q_{\rm axis}$ and $\lambda_{\rm s}$ is strong, indicating that these two quantities share some common origins. The initial condition is also important for $\lambda_{\rm s}$. However, putting all of the factors together, we see that the values of R^2 are still smaller than 0.5 for both $q_{\rm axis}$ and $\lambda_{\rm s}$, indicating that these two halo quantities may be affected by many subtle factors and thus are difficult to model.

Third, the structural properties depend mainly on the first PC of the MAH for low-mass halos, but have a significant dependence on higher order PCs for high-mass halos. The conclusions for the overall population still hold for halos of a given mass: environment has almost no effect on c once the MAH is included; environment is more important for q_{axis} and λ_{s}, although its effect on λ_{s} is partly degenerate with that of the MAH.

Fourth, the tidal anisotropy, $\alpha_{\mathcal{T}}$, has a stronger correlation with halo assembly and structure than the bias factor b does. We see that $\alpha_{\mathcal{T}}$ is correlated with halo assembly history for all halos except the most massive ones, and it also shows a significant correlation with q_{axis}, indicating that the local tidal field plays an important role in shaping a halo. However, all types of assembly bias tested here are weak compared to the variance in the relation, and averaging with a large sample is needed to detect them reliably.

For dimension-reduction tasks such as those for the halo MAH, one may also use nonlinear algorithms, such as the locally linear embedding (LLE; Roweis 2000) and the spectral embedding (Belkin & Niyogi 2003). The degrees of freedom of these manifold-learning techniques are higher than those of the linear algorithms, so they are not stable for noise data. Indeed, we have tried implementing LLE in halo MAH, but we found that it does not outperform the PCA in terms of both the reconstruction error and the correlation with halo structural properties.

The correlation of structural properties with halo assembly and environment revealed by the RF analysis provides insights into the halo population in the cosmic density field. Since galaxies form and evolve in dark matter halos, understanding the formation, structure, and environment of dark matter halos is a crucial step in establishing the link between galaxies and halos. Empirical approaches, such as (sub)halo abundance matching (Mo et al. 1999; Vale & Ostriker 2004; Guo et al. 2010), clustering matching (Guo et al. 2016), age matching (Hearin & Watson 2013), conditional color–magnitude diagrams (Xu et al. 2018), halo occupation distributions (Jing et al. 1998; Berlind & Weinberg 2002), conditional luminosity functions (Yang et al. 2003), and those based on star formation rate (Lu et al. 2014; Moster et al. 2018; Behroozi et al. 2019) all use halo properties to make predictions for the galaxy population. A key question in all of these models is which halo quantities should be used as the predictors of galaxies. Using too little information about the halo population will make the model too simple to capture the real effects of halos on galaxy formation; using too many halo properties may be unnecessary because of the degeneracy between them. Our results, therefore, provide a foundation for building galaxy formation models such as those listed above.

For readers who are interested in generating Monte Carlo samples of halos of different structural properties, including dependencies on assembly and environment properties, we provide both an online calculator and a programming interface at https://www.chenyangyao.com/publication/20/haloprops/.

This work is supported by the National Key R&D Program of China (grant Nos. 2018YFA0404502 and 2018YFA0404503), and the National Science Foundation of China (grant Nos. 11821303, 11973030, 11761131004, 11761141012, 11833005, 11621303, 11733004, 11890693, 11421303). Y.C. gratefully acknowledges the financial support from the China Scholarship Council.

Appendix A
Methods of Analysis

Throughout this work, we use two statistical methods to analyze halo properties. The PCA is used to reduce the complexity of quantities in high-dimensional space, and the EDT, also called the RF, is used to study correlations among different quantities. A brief description of the two methods is given below. For a more detailed description, see *Pattern Recognition and Machine Learning* by Bishop (2006). The programming interfaces and implementation can be found in scikit-learn.[10]

A.1. PCA

PCA is an unsupervised, reduced linear Gaussian dimension-reduction method (Pearson 1901; Hotelling 1933). Consider a set of N vectors $\boldsymbol{X} = (\boldsymbol{x}_1 ,..., \boldsymbol{x}_{\mathrm{N}})^{\mathrm{T}}$, each in an M–d space, V_{M}. The idea of the PCA is to find an M'–d subspace, $V_{\mathrm{M}'}$ ($M' \leqslant M$), in which the projection of $\boldsymbol{X}$,

$$\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{P}, \tag{A1}$$

has maximal variance, where $\boldsymbol{P}$ is the projection operator. It can be shown that the problem to be solved is equivalent to solving the eigenvalue problem for the sample covariance matrix of $\boldsymbol{X}$, defined as

$$\boldsymbol{S} = \sum_{i=1}^{\mathrm{N}} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}}, \tag{A2}$$

where $\bar{\boldsymbol{x}} = \sum_{i=1}^{\mathrm{N}} \boldsymbol{x}_i / N$ is the sample mean. If we rank the eigenvalues in descending order, the first eigenvector of $\boldsymbol{S}$, $\boldsymbol{v}_1$, is the direction along which the sample has the maximum projected variance, $\lambda_1 = \boldsymbol{v}_1^{\mathrm{T}} \boldsymbol{S} \boldsymbol{v}_1$. This variance is exactly the first eigenvalue of $\boldsymbol{S}$. Similarly, the ith eigenvector and the ith eigenvalue are, respectively, the direction and value of the ith largest variance. Consider the space $V_{\mathrm{M}'}$ spanned by the first M' eigenvectors. The linearity of the transformation can be used to prove that the projected variance in M' is $\sigma^2 = \sum_{i=1}^{\mathrm{M}'} \lambda_i$. Thus, one can project each data point $\boldsymbol{x}$ into $V_{\mathrm{M}'}$ by $\boldsymbol{P} = (\boldsymbol{v}_1 ,..., \boldsymbol{v}_{\mathrm{M}'})$, $\boldsymbol{x}' = \boldsymbol{P}^{\mathrm{T}}\boldsymbol{x}$, to find a lower-dimension representation for it. The ith component of $\boldsymbol{x}'$ is called the ith PC of this data point in the sample.

In general, any dimension-reduction algorithm will lose information contained in the original data. In PCA, the proportional variance explained (PVE) by the ith PC, defined as

$$\mathrm{PVE}_i = \frac{\lambda_i}{\mathrm{Tr}(\boldsymbol{S})}, \tag{A3}$$

is used to quantify the importance of the ith PC. The cumulative PVE, defined as

$$\mathrm{CPVE}_{\mathrm{M}'} = \sum_{i=1}^{M'} \mathrm{PVE}_i, \tag{A4}$$

can be used to quantify the performance of using the first M' PCs in the dimension reduction. Typically, if the data in question are generated from an intrinsic process of lower dimension, the CPVE should quickly converge to 1 as M' increases. We will see that halo assembly histories have this property (Section 3.1).

The inverse operation of the projection, $\tilde{\boldsymbol{x}} = \boldsymbol{P}\boldsymbol{x}'$, allows one to reconstruct the original vector, but with information loss. We use the following quantity,

$$e = ||\tilde{\boldsymbol{x}} - \boldsymbol{x}|| / ||\boldsymbol{x}||, \tag{A5}$$

[10] https://scikit-learn.org/stable/modules/ensemble.html

to quantify the reconstruction error of the data point $\boldsymbol{x}$.

A.2. The Random Forest

The RF regressor or classifier is a supervised, decision-tree-based, highly nonlinear, nonparametric model ensemble method in statistical learning (Breiman 2001). Here we first introduce the decision tree algorithm, and then we discuss how the trees are combined into a forest to make a regression or classification.

Given a set of observations $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, each consisting of a vector of predictor variables $\boldsymbol{x}_i \in \mathbb{R}^{\mathrm{M}}$ and a target random variable y_i (continuous in the regression problem, discrete in the classification problem), a decision tree T can be trained to fit the data by minimizing some error functional $\mathcal{E}(\mathrm{T}|D)$. In regression problems, a common choice for the error functional, which we adopt here, is the mean residual sum-of-square,

$$\mathcal{E}(\mathrm{T}|\mathcal{D}) = \frac{1}{N}\sum_{n=1}^{N}[f(\boldsymbol{x}_i) - y_i]^2, \tag{A6}$$

where $f(\boldsymbol{x}_i)$ is the predicted value for the ith observation by the tree T. The tree can then be used to predict the target value for a future test observation, $\boldsymbol{x}$.

A decision tree is built by sequentially bipartitioning the feature space along some axes. At each partitioned region $\mathcal{R}$ in the feature space, the tree fits the observations by a constant function,

$$f(\boldsymbol{x}) = \frac{1}{N_{\mathcal{R}}}\sum_{n=1}^{N_{\mathcal{R}}} y_i, \tag{A7}$$

where the summation is over all observations in the region $\mathcal{R}$, and $N_{\mathcal{R}}$ is the number of training observations in this region. In the first step, the variable x to be bipartitioned and the position of the partition plane are both chosen to minimize the error functional $\mathcal{E}(\mathrm{T}|D)$. After a partition, the feature space is split into two subregions, each of which can be bipartitioned further to minimize the error functional. This recursive process partitions the feature space into a tree-like structure and is continued until some stop criterion (e.g., the maximum tree height, or the maximum number of tree nodes, or the minimum number of data points in individual leaf nodes, defined as the nodes at the top of the tree) is achieved.

Once a tree is built, the amount of error reduced by partitioning variable x can be computed as

$$\mathcal{I}_{\mathrm{T}|D}(x) = C\sum_{n} N_n\left(\mathcal{E}_n - \frac{N_{n_{\mathrm{L}}}}{N_n}\mathcal{E}_{n_{\mathrm{L}}} - \frac{N_{n_{\mathrm{R}}}}{N_n}\mathcal{E}_{n_{\mathrm{R}}}\right). \tag{A8}$$

Here the summation is over all tree nodes partitioned by variable x; $\mathcal{E}_n$ is the error functional computed at observations in the region represented by node n before partition; $\mathcal{E}_{n_{\mathrm{L}}}$ and $\mathcal{E}_{n_{\mathrm{R}}}$ are the errors computed at the left and right child nodes of node n, respectively, after partition; and N_n, $N_{n_{\mathrm{L}}}$, and $N_{n_{\mathrm{R}}}$ are the number of observations in node n and its two child nodes, respectively. The normalization factor C is chosen so that the summation of $\mathcal{I}$ from all feature variables is one. So defined, the quantity $\mathcal{I}_{\mathrm{T}|D}(x)$ is the amount of contribution from variable x in building the regressor and can therefore be viewed as the important value of the variable in explaining target y.

Such a tree model suffers from the overfitting problem: the more complicated the tree is, the less training error it will have. However, the tree will eventually be dominated by noise as its height increases. Many methods have been proposed to control such overfitting, for example, the cross-validation, bootstrap, and jackknife ensembles. Here we adopt the RF, an extension of the bootstrap ensemble, designed specifically to deal with overfitting in tree-like algorithms. The building of an RF involves two levels of randomness. First, one uses n_{re} bootstrap resamplings, and each is used to train a tree T_i $(i = 1, \ldots, n_{\mathrm{re}})$. Second, when training each of the n_{re} trees, only a random subset (size $n_{\mathrm{var}} < M$) of all M predictor variables is used at each partition step. The n_{re} trees are then combined, and the final prediction for a given feature $\boldsymbol{x}$, denoted as $\mathrm{RF}(\boldsymbol{x})$, is then averaged among all trees (arithmetic average in regression, and majority voting in classification). Also, the importance of the predictor, $\mathcal{I}_{\mathrm{T}|D}(x)$, is averaged among trees, which we denote as $\mathcal{I}_{\mathrm{RF}|D}(x)$. The overall performance of the RF is represented by the explained variance fraction, R^2, defined as

$$R^2 = 1 - \frac{\sum_{n=1}^{N_{\mathrm{T}}}[y_n - \mathrm{RF}(\boldsymbol{x}_n)]^2}{\sum_{n=1}^{N_{\mathrm{T}}}[y_n - \bar{y}]^2}, \tag{A9}$$

where in principle the summation should be computed over an independent test sample, T, of size N_{T}. The RF regressor has the advantage that the test performance can be directly estimated with the OOB sample in the bootstrap process, and therefore an extra test sample is not necessary. In addition, RF does not suffer from the issue of scaling or arbitrary transformation of predictors, which exists in many nonlinear approaches, such as K-nearest-neighbors and support vector machines.

The RF method has some free parameters to be specified, and we choose them based on the following considerations: (1) The number of trees in the forest, n_{re}, should be as large as possible to suppress overfitting. But a larger n_{re} is computationally more difficult. In our analysis, we choose $n_{\mathrm{re}} = 100$, which is sufficiently large for most applications of RF. (2) The number of predictors randomly chosen in the partition, n_{var}, also controls the suppression of overfitting. We optimize the value of n_{var} by maximizing the OOB score through grid searching. (3) The tree termination criterion affects the complexity of each tree. We choose to control the number of data points in the leaf nodes, s_{leaf}. This choice makes the tree self-adaptive when more data points are available, and also reduces issues associated with transformations of target variables and the choice of the error functional. The value of s_{leaf} is also optimized by maximizing the OOB score, again through grid searching.

Appendix B
Definitions of Halo Formation Times

Because of the diversity in MAHs, different formation times can be defined to describe different aspects of the assembly. Here we summarize the halo formation times we used in our analysis. Most of the definitions can be found in Li et al. (2008), but we also add some new definitions that have been used by others. Since the halo mass and redshift in different time steps are discrete, the mass–redshift relation is linearly interpolated within adjacent time steps.

1. $z_{\rm mb,1/2}$: the highest redshift at which the main branch has assembled half of its final mass $M_{\rm halo}$(e.g., van den Bosch 2002; Shi et al. 2018).
2. $z_{\rm mb,core}$: the highest redshift at which the main branch has reached a fixed mass $M_{\rm h,core} = 10^{11.5}\,h^{-1}M_\odot$. A halo at such a mass typically has the highest star formation efficiency (van den Bosch et al. 2003; Yang et al. 2003) and therefore is capable of forming a bright galaxy in it.
3. $z_{\rm mb,0.04}$: the highest redshift at which the main branch of a halo assembled 4% of its final mass. This formation time is found to be related to the concentration of the halo in a wide range of cosmological models (see Zhao et al. 2009).
4. $z_{\rm mp,1/2}$: similar to $z_{\rm mb,1/2}$, but using the most massive progenitors (MMPs) in the entire tree rooted from a halo, instead of the main branch. This definition is used in Wang et al. (2011).
5. $z_{\rm mp,core}$: similar to $z_{\rm mb,core}$, but using MMPs.
6. $z_{0.02,1/2}$: the highest redshift at which half of the halo mass has been assembled into its progenitors with masses $\geqslant$2% of the halo mass. This definition is used in Navarro et al. (1997) to study the correlation between halo concentration and formation history (see also Jeeson-Daniel et al. 2011, where a different mass threshold is used).
7. $z_{\rm core,1/2}$: the highest redshift at which half of the halo mass has been assembled into its progenitors with masses $M_{\rm halo} \geqslant M_{\rm h,core} = 10^{11.5}\,h^{-1}M_\odot$. This represents the time when the massive progenitors are capable of forming large amounts of stars.
8. $z_{{\rm core},f}$: the highest redshift at which a fraction $f = \frac{1}{2}(M_{\rm halo}/M_{\rm h,core})^{-\gamma}$ of the halo mass has been assembled into its progenitors with masses $M_{\rm halo} \geqslant M_{\rm h,core}$, where $\gamma = 0.32$. This definition takes into account the dependence of star formation efficiency on halo mass (Yang et al. 2003).
9. $z_{\rm mb,vmax}$: the redshift at which the main branch has achieved its maximum virial velocity. This definition, therefore, reflects the formation of the gravitational potential well.
10. $z_{\rm mp,vmax}$: the same as $z_{\rm mb,vmax}$, but using the maximum virial velocity of MMPs.
11. $z_{\rm lmm}$: the last major merger time of a halo. Here the major merger is defined as a merger event in which the mass ratio $r = m/M$ between the two merger parts is larger than one-third. A major merger is a violent event and may change the halo structure significantly.
12. $s_{{\rm tree},\beta}$: the tree entropy with entropy update efficiency β (see Obreschkow et al. 2020). Different from the formation parameters defined above, this parameter is not associated with any specific event of the halo assembly history, but it describes the complexity of the whole tree. By construction, $s_{{\rm tree},\beta}$ is bound to the range [0, 1]. A close-to-zero $s_{{\rm tree},\beta}$ represents a continuous accretion history, while a close-to-one $s_{{\rm tree},\beta}$ describes a history that is given by the merger of two progenitors of equal mass. The parameter $\beta \in [0,\ 1]$ controls the balance between the entropy inherited from the progenitors and that generated in recent merger events. We choose $\beta = 0.1$ so that the tree entropy can reflect its assembly history at high z.

Appendix C
Effect of Unrelaxed Halos

As demonstrated by Macciò et al. (2007), halos that have undergone recent major mergers may be unrelaxed and have structural properties significantly different from virialized halos (see also Ludlow et al. 2012). Following Macciò et al. (2007), we define two parameters to quantify the dynamic states of halos. The first is the χ_c^2 parameter, defined as the minimized χ^2 in fitting the NFW profile, normalized by halo mass (see Section 2.3). The second is the offset parameter, $x_{\rm off}$, defined as the distance from the center of mass of the halo to the most bound particle, normalized by the virial radius. Figure C1 shows the distributions of $\chi_{\rm c}^2$ and $x_{\rm off}$ for subsamples with different last major merger times. As one can see, if a halo has experienced a recent major merger ($\log\delta_{\rm c}(z_{\rm lmm}) < 0.3$), it is likely that its profile deviates from the NFW profile, and that its most bound particle is far away from the center of mass of the halo.

Including those unrelaxed halos in our sample will significantly increase the variance of halo properties, thereby affecting the statistics derived from the sample. To reduce their effects, we exclude all halos that have undergone a major

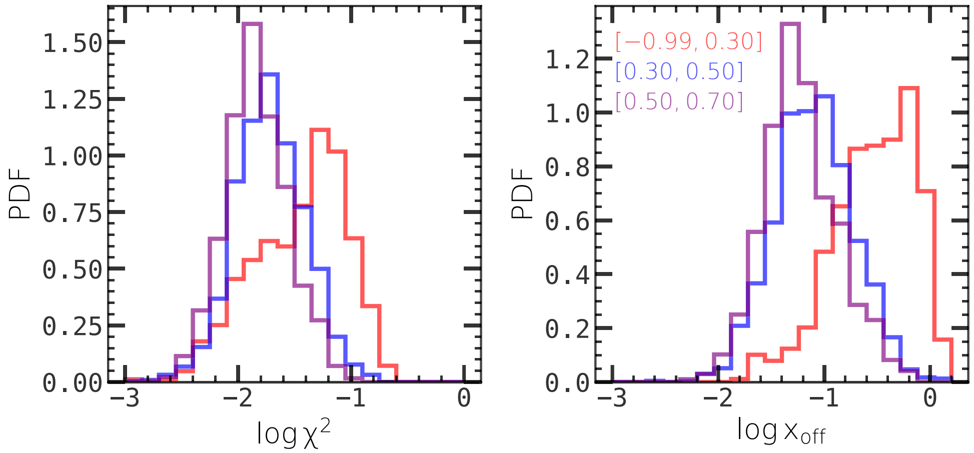

Figure C1. Distributions of halo relaxation-related parameters. Left: relaxation parameter $\chi_{\rm c}^2$, which is the normalized χ-squared value in fitting the NFW profile. Right: offset parameter $x_{\rm off}$, which is the normalized distance from the center of mass to the most bound particle of the halo. In each panel, halos are taken from the mass-limited sample $\mathrm{S}_{\rm c}'$ with $M_{\rm halo} \geqslant 5 \times 10^{11}\,h^{-1}M_\odot$ (see Section 2.1 and Table 1) and are divided into three subsamples according to their last major merger time $\log\delta_{\rm c}(z_{\rm lmm})$, indicated in the right panel with different colors.

merger at $z < 0.3$. This will remove 9.6%, 9.6%, 14.2%, 21.8%, 10.5%, and 13.8% of halos in samples S_1, S_2, S_3, S_4, S_c, and S_L, respectively. To make our conclusion even less dependent on relaxation processes, we compute λ_s and q_{axis} using only simulated particles that are within a radius 2.5 r_s from the most bound particle, where r_s is the scale radius of the fitted NFW profile. According to our test, our results are not sensitive to the radius chosen.

ORCID iDs

Yangyao Chen https://orcid.org/0000-0002-4597-5798
H. J. Mo https://orcid.org/0000-0002-9665-5380
Cheng Li https://orcid.org/0000-0002-8711-8970
Huiyuan Wang https://orcid.org/0000-0002-4911-6990
Xiaohu Yang https://orcid.org/0000-0003-3997-4606
Youcai Zhang https://orcid.org/0000-0003-1967-4091
Kai Wang https://orcid.org/0000-0002-3775-0484

References

Behroozi, P., Wechsler, R. H., Hearin, A. P., & Conroy, C. 2019, MNRAS, 488, 3143
Belkin, M., & Niyogi, P. 2003, Neural Computation, 15, 1373
Berlind, A. A., & Weinberg, D. H. 2002, ApJ, 575, 587
Bett, P., Eke, V., Frenk, C. S., et al. 2007, MNRAS, 376, 215
Bhattacharya, S., Habib, S., Heitmann, K., & Vikhlinin, A. 2013, ApJ, 766, 32
Bishop, C. M. 2006, Pattern Recognition and Machine Learning (Berlin: Springer)
Bluck, A. F. L., Maiolino, R., Sánchez, S. F., et al. 2020, MNRAS, 492, 96
Bonjean, V., Aghanim, N., Salomé, P., et al. 2019, A&A, 622, A137
Breiman, L. 2001, Machine Learning, 45, 5
Bryan, G. L., & Norman, M. L. 1998, ApJ, 495, 80
Bullock, J. S., Dekel, A., Kolatt, T. S., et al. 2001, ApJ, 555, 240
Carroll, S. M., Press, W. H., & Turner, E. L. 1992, ARA&A, 30, 499
Chen, Y., Mo, H. J., Li, C., et al. 2019, ApJ, 872, 180
Cohn, J. D. 2018, MNRAS, 478, 2291
Cohn, J. D., & van de Voort, F. 2015, MNRAS, 446, 3253
Correa, C. A., Wyithe, J. S. B., Schaye, J., & Duffy, A. R. 2015, MNRAS, 450, 1514
Dalal, N., White, M., Bond, J. R., & Shirokov, A. 2008, ApJ, 687, 12
Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371
de los Rios, M., Domínguez, R. M. J., Paz, D., & Merchán, M. 2016, MNRAS, 458, 226
Desjacques, V. 2008, MNRAS, 388, 638
Dobrycheva, D. V., Vavilova, I. B., Melnyk, O. V., & Elyiv, A. A. 2017, arXiv:1712.08955
Dunkley, J., Komatsu, E., Nolta, M. R., et al. 2009, ApJS, 180, 306
Eisenstein, D. J., & Hu, W. 1998, ApJ, 496, 605
Faltenbacher, A., & White, S. D. M. 2010, ApJ, 708, 469
Gao, L., Springel, V., & White, S. D. M. 2005, MNRAS: Letters, 363, L66
Gao, L., & White, S. D. M. 2007, MNRAS: Letters, 377, L5
Guo, H., Zheng, Z., Behroozi, P. S., et al. 2016, MNRAS, 459, 3040
Guo, Q., White, S., Li, C., & Boylan-Kolchin, M. 2010, MNRAS, 404, 1111
Haas, M. R., Schaye, J., & Jeeson-Daniel, A. 2012, MNRAS, 419, 2133
Hahn, O., Porciani, C., Carollo, C. M., & Dekel, A. 2007, MNRAS, 375, 489
Han, J., Li, Y., Jing, Y., et al. 2019, MNRAS, 482, 1900
Hearin, A. P., & Watson, D. F. 2013, MNRAS, 435, 1313
Hockney, R. W., & Eastwood, J. W. 1988, Computer Simulation Using Particles (Bristol, PA: Taylor & Francis, Inc.)
Hotelling, H. 1933, Journal of Educational Psychology, 24, 417
Jeeson-Daniel, A., Vecchia, C. D., Haas, M. R., & Schaye, J. 2011, MNRAS: Letters, 415, L69
Jing, Y. P. 2000, ApJ, 535, 30
Jing, Y. P., Mo, H. J., & Borner, G. 1998, ApJ, 494, 1
Jing, Y. P., & Suto, Y. 2002, ApJ, 574, 538
Jing, Y. P., Suto, Y., & Mo, H. J. 2007, ApJ, 657, 664
Lazeyras, T., Musso, M., & Schmidt, F. 2017, JCAP, 2017, 059
Li, Y., Mo, H. J., & Gao, L. 2008, MNRAS, 389, 1419
Lu, Y., Mo, H. J., Katz, N., & Weinberg, M. D. 2006, MNRAS, 368, 1931
Lu, Z., Mo, H. J., Lu, Y., et al. 2014, MNRAS, 439, 1294
Lucie-Smith, L., Peiris, H. V., & Pontzen, A. 2019, MNRAS, 490, 331
Lucie-Smith, L., Peiris, H. V., Pontzen, A., & Lochner, M. 2018, MNRAS, 479, 3405
Ludlow, A. D., Bose, S., Angulo, R. E., et al. 2016, MNRAS, 460, 1214
Ludlow, A. D., Navarro, J. F., Angulo, R. E., et al. 2014, MNRAS, 441, 378
Ludlow, A. D., Navarro, J. F., Boylan-Kolchin, M., et al. 2013, MNRAS, 432, 1103
Ludlow, A. D., Navarro, J. F., Li, M., et al. 2012, MNRAS, 427, 1322
MacCiò, A. V., Dutton, A. A., van den Bosch, F. C., et al. 2007, MNRAS, 378, 55
MacCiò, A. V., Dutton, A. A., & van den Bosch, F. C. 2008, MNRAS, 391, 1940
Man, Z.-y., Peng, Y.-J., Shi, J.-J., et al. 2019, ApJ, 881, 74
Mao, Y.-Y., Zentner, A. R., & Wechsler, R. H. 2018, MNRAS, 474, 5143
McBride, J., Fakhouri, O., & Ma, C. P. 2009, MNRAS, 398, 1858
Mo, H., van den Bosch, F., & White, S. 2010, Galaxy Formation and Evolution (Cambridge: Cambridge Univ. Press)
Mo, H. J., Mao, S., & White, S. D. M. 1999, MNRAS, 304, 175
Mo, H. J., & White, S. D. M. 1996, MNRAS, 282, 347
Morinaga, Y., & Ishiyama, T. 2020, MNRAS, 495, 502
Moster, B. P., Naab, T., & White, S. D. M. 2018, MNRAS, 477, 1822
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493
Obreschkow, D., Elahi, P. J., Lagos, C. d. P., Poulton, R. J. J., & Ludlow, A. D. 2020, MNRAS, 493, 4551
Parkinson, H., Cole, S., & Helly, J. 2008, MNRAS, 383, 557
Pearson, K. 1901, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2, 559
Press, W. H., & Schechter, P. 1974, ApJ, 187, 425
Rafieferantsoa, M., Andrianomena, S., & Davé, R. 2018, MNRAS, 479, 4509
Ramakrishnan, S., Paranjape, A., Hahn, O., & Sheth, R. K. 2019, MNRAS, 489, 2977
Roweis, S. T. 2000, Sci, 290, 2323
Salcedo, A. N., Maller, A. H., Berlind, A. A., et al. 2018, MNRAS, 475, 4411
Sandvik, H. B., Moller, O., Lee, J., & White, S. D. M. 2007, MNRAS, 377, 234
Sheth, R. K., Mo, H. J., & Tormen, G. 2001, MNRAS, 323, 1
Shi, J., Wang, H., Mo, H. J., et al. 2018, ApJ, 857, 127
Springel, V. 2005, MNRAS, 364, 1105
Sreejith, S., Pereverzyev, S., Kelvin, L. S., et al. 2018, MNRAS, 474, 5232
Stivaktakis, R., Tsagkatakis, G., Moraes, B., et al. 2018, 1, arXiv:1809.09622
Tasitsiomi, A., Kravtsov, A. V., Gottlober, S., & Klypin, A. A. 2004, ApJ, 607, 125
Vale, A., & Ostriker, J. P. 2004, MNRAS, 353, 189
van den Bosch, F. C. 2002, MNRAS, 331, 98
van den Bosch, F. C., Yang, X., & Mo, H. J. 2003, MNRAS, 340, 771
Wang, H., Mo, H. J., & Jing, Y. P. 2009, MNRAS, 396, 2249
Wang, H., Mo, H. J., Jing, Y. P., Yang, X., & Wang, Y. 2011, MNRAS, 413, 1973
Wang, H., Mo, H. J., Yang, X., et al. 2016, ApJ, 831, 164
Wang, H. Y., Mo, H. J., & Jing, Y. P. 2007, MNRAS, 375, 633
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
Wong, A. W. C., & Taylor, J. E. 2012, ApJ, 757, 102
Xu, H., Zheng, Z., Guo, H., et al. 2018, MNRAS, 481, 5470
Yang, X., Mo, H. J., & van den Bosch, F. C. 2003, MNRAS, 339, 1057
Zentner, A. R. 2007, IJMPD, 16, 763
Zhao, D. H., Jing, Y. P., Mo, H. J., & Börner, G. 2009, ApJ, 707, 354
Zhao, D. H., Jing, Y. P., Mo, H. J., & Brner, G. 2003a, ApJL, 597, L9
Zhao, D. H., Mo, H. J., Jing, Y. P., & Borner, G. 2003b, MNRAS, 339, 12