Explaining Dark Matter Halo Density Profiles with Neural Networks

Luisa Lucie-Smith,[1,*] Hiranya V. Peiris,[2,3] and Andrew Pontzen[2]
[1]*Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, 85748 Garching, Germany*
[2]*Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, United Kingdom*
[3]*The Oskar Klein Centre for Cosmoparticle Physics, Stockholm University, AlbaNova, Stockholm SE-106 91, Sweden*

(Received 12 May 2023; revised 4 September 2023; accepted 20 December 2023; published 19 January 2024)

We use explainable neural networks to connect the evolutionary history of dark matter halos with their density profiles. The network captures independent factors of variation in the density profiles within a low-dimensional representation, which we physically interpret using mutual information. Without any prior knowledge of the halos' evolution, the network recovers the known relation between the early time assembly and the inner profile and discovers that the profile beyond the virial radius is described by a single parameter capturing the most recent mass accretion rate. The results illustrate the potential for machine-assisted scientific discovery in complicated astrophysical datasets.

DOI: 10.1103/PhysRevLett.132.031001

Introduction.—In the modern picture of structure formation, galaxies form at the center of extended, overdense "halos" of dark matter, which originate from small fluctuations in the density of matter in the early Universe and undergo highly nonlinear dynamical processes throughout their evolution [1–5]. The history of a halo determines its final structure, commonly parametrized by the spherically averaged radial density profile. Halo density profiles are not only key ingredients of the galaxy-halo connection in cosmological analyses and of direct and indirect dark matter searches, they are also powerful observational test beds of fundamental physics. This is because their shape, from the inner core to the outskirts in the proximity of the splashback radius, is sensitive to the nature of dark matter and modifications to gravity [6,7].

Observationally, it has recently become possible to measure weak lensing and 3D density profiles through a combination of multiwavelength data; upcoming data from *Euclid*, Rubin, and DESI will provide even more detailed measurements of the density profiles of halos from clusters to dwarf galaxies [8–10]. Achieving the potential impact of these measurements requires determining the physical effects that control the shape of the density profiles. However, current theoretical models are limited to empirical fitting functions such as the Navarro-Frenk-White (NFW) [11] and Einasto [12] profiles; these do not explain the physical origin of the profiles' universal shape seen in numerical simulations [13,14]. Understanding the connection between the formation history and the density profile also offers the possibility of using observational constraints on the latter to estimate the halos' mass accretion rate. This will yield valuable constraints on galaxy formation [15–18] as well as extensions to the cosmological model.

In this Letter, we use an "explainable AI" framework to connect the formation histories of dark matter halos to their density profiles. Our goal differs from typical uses of machine learning in cosmology, such as emulating the output of computationally expensive simulations [19,20] or accelerating the estimation of cosmological parameters from data [21–24]. In Ref. [25], we used an interpretable deep learning framework to build a new model for the spherically averaged halo density profiles, generalizing over existing empirical fitting functions. The framework, which we denoted as an "interpretable variational encoder" (IVE), was trained to capture all the information used by the neural network to predict the profile, given the 3D density field around the halo center, within a compact, low-dimensional latent representation. We require the representation to be disentangled, i.e., each latent component captures different, independent factors of variation in the profiles; the latent representation is equivalent to the profiles' degrees of freedom. We found that three components are required (and sufficient) for modeling the profiles out to the halo outskirts: these three components describe, respectively, the normalization of the profile and its shape within and beyond the virial radius.

In this Letter, we turn to the physical interpretation of the learned IVE latent representation to investigate how halo density profiles are determined from the halos' formation histories. Although the network was trained only on the present-day density field, we explore whether the latent parameters carry memory of the evolution history of the halos. We measure the information encoded within each latent about the halos' evolution history using the information-theoretic measure of "mutual information" (MI). By this metric, the IVE representation and the NFW parametrization similarly highlight a dependence of the profile on physical accretion history. However, the IVE additionally allows us to measure the connection between a halo's recent evolution history and the density in its far outskirts, something that the NFW profile does not capture.

Background.—We begin by briefly reviewing the current understanding of the physics of halo density profiles. The NFW profile is the most widely used fitting function for the halo density profile. It is given by

$$\rho(r) = \frac{\rho_s}{r/r_s(1+r/r_s)^2}, \tag{1}$$

where r_s and ρ_s are the scale radius and characteristic density, respectively. The scale radius is often rewritten in terms of a concentration parameter $c \equiv r_{200\ \mathrm{m}}/r_s$, so that the NFW profile depends on the virial radius $r_{200\ \mathrm{m}}$ and concentration c. The virial radius $r_{200\ \mathrm{m}}$ is typically adopted as a proxy for the halo boundary and defined as the radius that contains a mean density that is 200 times the mean density of the Universe. High-resolution simulations have revealed this functional form to be "universal": it provides a good fit to stacked profiles of halos for a large range of halo masses [14,26], for several different cosmological models [27–31], and even in the absence of hierarchical growth [32–34]. This suggests that universal density profiles are a generic feature that arises from collisionless gravitational collapse.

Despite the lack of a first-principles explanation for the self-similarity of halo density profiles, some insights have been gained from studying the correlation between the NFW concentration and summary statistics of the halo evolution process. Mass, concentration, and halo formation time all correlate: on average, low-mass halos assemble earlier and have higher characteristic densities (or concentration), reflecting the larger background density at earlier times [11,26,30,35–38]. This description can explain the qualitative trend of the mean concentration as a function of halo mass, but not the large residual scatter in concentration seen in simulations [36,39]. It is also limited to the simplest summary statistic of the halo evolution history, i.e., the halo formation time. The scatter in concentration at fixed mass has been shown to be at least in part connected to merger events during the halo assembly process [39–41]. Further work has suggested that the self-similarity of halos may be related to the self-similarity of the halo mass assembly history [26], although this has only been validated on stacked profiles of well-behaved,"relaxed" halos.

The situation worsens when modeling profiles beyond the virial radius: the halo outskirts strongly deviate from the NFW form due to the presence of the splashback radius, where particles reach the apocenter of their first orbit. Recent work has focused on modeling the location of the splashback radius, finding that it is sensitive to the late-time mass accretion rate [42–47]. Modeling the full shape of the outer profile remains a difficult task due to its intrinsically nonequilibrium nature, leading to a reliance on multiparameter fitting functions with little physical explainability [48,49].

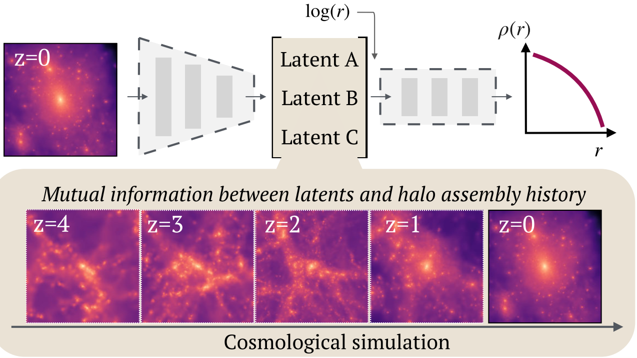

FIG. 1. A neural network is trained to discover the underlying degrees of freedom in halo density profiles in the form of a latent representation, when presented with the full 3D density structure of a halo. We physically interpret the discovered representation by measuring the MI between the latent parameters and the assembly history of the halos.

Deep learning model.—The IVE architecture used in this Letter has two main components: the encoder, mapping the 3D density field to a low-dimensional latent representation, and the decoder, mapping the latent representation and the query radius $\log(r)$ to the output profile $\log[\rho(r)]$. By design, all the information used by the model to predict the density profiles is captured within the latent representation. An illustration of the model is shown in the top half of Fig. 1. The encoder is a 3D convolutional neural network with parameters ϕ that maps the inputs $\boldsymbol{x}$ to a multivariate distribution in the latent space $p_\phi(z|\boldsymbol{x})$. We choose the latent representation to be a set of independent Gaussians, $p_\phi(z|\boldsymbol{x}) = \prod_{i=1}^{L} \mathcal{N}(\mu_i(\boldsymbol{x}), \sigma_i(\boldsymbol{x}))$, where L is the dimensionality of the latent space; under this assumption, the encoder maps the inputs $\boldsymbol{x}$ to the vectors $\mu = \mu_i, \ldots, \mu_L$ and $\sigma = \sigma_i, \ldots, \sigma_L$. The decoder of the IVE consists of another neural network model with parameters θ that maps a sampled latent vector $z \sim p_\phi(z|\boldsymbol{x})$ and a value of the query $\log(r)$ to a single predicted estimate for $\log[\rho_{\mathrm{pred}}(r)]$.

A crucial aspect of the IVE that makes the latent space interpretable is that it is *disentangled*: independent factors of variation in the density profiles are captured by different, independent latents. This is achieved through the design of a loss function that minimizes the mean squared error between predicted and ground-truth profiles, while simultaneously maximizing the degree of independence between the latent variables by encouraging those to be as close as possible to independent Gaussians of mean 0 and variance 1 [50]. More details on the encoder and decoder architectures and the loss function are presented in Ref. [25].

Methods.—We generated the training data from four dark-matter-only N-body simulations produced with GADGET-4 [51], each containing 512^3 particles in a $(50\ \mathrm{Mpc}\,h^{-1})^3$ box. We trained two IVE models for different tasks: one (IVE$_{\mathrm{virial}}$) was trained to model the density profile up to the halo virial radius $r_{200\ \mathrm{m}}$, and the second (IVE$_{\mathrm{infall}}$) was trained to model profiles beyond the halo boundary out to $2r_{200\ \mathrm{m}}$. The former is used for direct comparison with the NFW profile, which is also designed to model the profile out to the virial radius, and the latter is used to investigate the less studied halo outer profile. The innermost radius of the profiles we consider is $r_{\mathrm{min}} = 3\epsilon$, where ϵ is the gravitational softening of the simulation; this choice ensures that we can robustly trust the inner profile. The inputs are given by the 3D density field within a $N = 131^3$ subbox of size $L_{\mathrm{subbox}} = 0.4\ \mathrm{Mpc}\,h^{-1}$ for the IVE$_{\mathrm{virial}}$ model and of size $L_{\mathrm{subbox}} = 0.6\ \mathrm{Mpc}\,h^{-1}$ for the IVE$_{\mathrm{infall}}$ one. We considered halos with $\log_{10}(M/M_{\odot}) \in [11, 13]$, but for the IVE$_{\mathrm{infall}}$ model, we further restricted our analysis to halos with $r_{200\ \mathrm{m}} \leq 150\ \mathrm{kpc}\,h^{-1}$. These cuts yielded ~17 000 (13 000) halos for training the IVE$_{\mathrm{virial}}$ (IVE$_{\mathrm{infall}}$) model. Further discussion on the training data of the IVE$_{\mathrm{virial}}$ and IVE$_{\mathrm{infall}}$ models is presented in Ref. [25]. To compare the IVE$_{\mathrm{virial}}$ results with the NFW profile, we fitted the NFW formula in Eq. (1) to each halo's density profile using least-squares minimization and recovered the best-fitting parameters r_s and ρ_s. The concentration was then derived using $c = r_{200\ \mathrm{m}}/r_s$. A description of the simulations used for training and testing the IVE models can be found in the Supplemental Material [52].

The first step of the analysis was to verify that the IVE models learn to predict the density profiles at ~5% accuracy, comparable to the accuracy of NFW fits (see Supplemental Material [52]). Crucially, only the $z = 0$ snapshots were used for training the IVE models to construct the latent representations mapping the 3D density field to dark matter halo profiles—i.e., the model had no access to the merger histories of the halos during training. The resulting disentangled latent space directly corresponds to the underlying degrees of freedom in the halo density profiles. Following recent works [25,57–60], we then used the MI to (i) quantify the information captured by the latent space about the halo density profiles and (ii) connect the IVE latents to the halo's evolution history, showing how the latter determines the present-day density profile.

The MI was estimated using GMM-MI [60], which performs density estimation using Gaussian mixtures and provides MI uncertainties through bootstrap. Background details on MI are provided in the Supplemental Material [52]. We first measured the MI between each latent and the density profile $\rho(r)$; this allows us to directly link each latent to a degree of freedom in the profile that affects its shape over a certain radial range. We then measured the MI between each latent and the mass assembly history of each halo. This, in turn, allowed us to connect each degree of freedom describing the density profile of the halo directly to characteristics of the halos' evolution that determines that component.

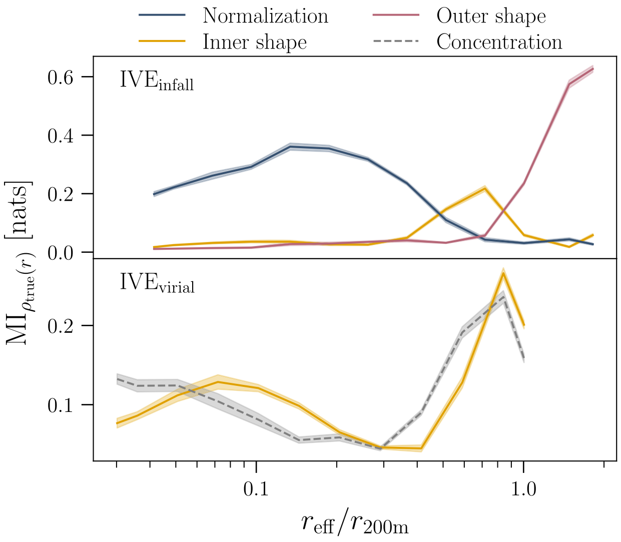

FIG. 2. The MI between the latent parameters and the ground-truth halo profiles $\rho_{\mathrm{true}}(r)$ for the IVE$_{\mathrm{infall}}$ (top) and the IVE$_{\mathrm{virial}}$ (bottom) models. In the IVE$_{\mathrm{virial}}$ case, we also show MI with the NFW concentration. (For clarity we do not show the IVE$_{\mathrm{virial}}$ normalization latent, since it behaves identically to the IVE$_{\mathrm{infall}}$ normalization latent.).

Results.—Figure 2 quantifies the information contained within the latents of the IVE$_{\mathrm{infall}}$ (top panel) and the IVE$_{\mathrm{virial}}$ (bottom panel) models about the ground-truth density profiles. (We verify the conclusions of our previous work in Ref. [25] at higher precision using the new GMM-MI estimator [60].) We show the MI between each latent parameter and the ground-truth profiles, which we denote as $\mathrm{MI}_{\rho_{\mathrm{true}}(r)}$. The three latents discovered by the IVE$_{\mathrm{infall}}$ describe (i) the normalization of the profile, which dominates the variation in the profiles out to $\sim r_{200\ \mathrm{m}}/2$, (ii) the shape of the inner profile, which becomes informative on radial scales approaching $r_{200\ \mathrm{m}}$, and (iii) the shape of the outer profile beyond $r_{200\ \mathrm{m}}$. The first two are analogous to the two NFW parameters, mass and concentration, respectively. A closer comparison between the inner shape latent of the IVE$_{\mathrm{virial}}$ model and concentration (bottom panel of Fig. 2) shows that both parameters carry information about the density in the core and on radial scales close to $r_{200\ \mathrm{m}}$. [The MI between the density profile and the inner latents of the IVE$_{\mathrm{infall}}$ and IVE$_{\mathrm{virial}}$ models (in yellow, Fig. 2 top and bottom panels, respectively) is qualitatively similar, despite the models' different training sets.] This bimodality is due to a compensation effect between the density in the inner region and that close to the virial boundary: at fixed normalization, halos with denser cores become less dense in the outskirts and vice versa. The $\mathrm{MI}_{\rho_{\mathrm{true}}(r)}$ of the inner latent is shifted toward larger radii compared to that of concentration, suggesting that the former is sensitive to variations in the shape of the profile on larger radial scales than the latter; this distinction will become relevant when physically interpreting the latent and comparing it to concentration.

We now move on to a physical interpretation of the latents in relation to characteristics of the halos' evolution histories. Recall that the network did not have access to this information during training. The interpretation of the normalization latent is straightforward: it captures the $z = 0$ mass of the halo, $M_{200\,\mathrm{m}}$. Their MI is $\sim 2.07 \pm 0.01$ nats, where the nat is the natural unit of information, implying a strong correlation between the two. This also matches expectations from the literature [11,12], as halo mass also controls the normalization in the NFW and Einasto fitting functions. To physically interpret the inner and outer shape latents, we measure their MI with two quantities that describe the assembly history of the halos over cosmic time. The first is the mass accretion history, $M_{200\,\mathrm{m}}(z)/M_{200\,\mathrm{m}}(z = 0)$, which describes the evolution of the halo mass as a function of time $M_{200\,\mathrm{m}}(z)$ normalized to the present-day halo mass $M_{200\,\mathrm{m}}(z = 0)$. The second is the mass accretion rate $\Gamma(t) \equiv \Delta \ln M_{200\,\mathrm{m}}(a)/\Delta \ln a$ [46], which describes the rate of change in halo mass with respect to the scale factor $a(t)$. The value of the accretion rate depends on the time interval used to compute the change in mass and scale factor; we compute $\Gamma(t)$ by taking the finite difference of the halo masses at each consecutive time step in the simulation.

Figure 3 shows the MI between the latents and mass accretion history ($\mathrm{MI}_{M(z)}$; top row) and that between the latents and the mass accretion rate ($\mathrm{MI}_{\mathrm{d}M(z)/\mathrm{d}z}$; bottom row). We first focus on the inner shape latent, which we compare to the NFW concentration. The $\mathrm{MI}_{M(z)}$ of the inner shape latent increases with time during the early formation period, peaks at $z \sim 1$, and declines rapidly toward $z = 0$; recall that this is the MI with the mass assembly history normalized to the present-day halo mass. This result reveals that the inner shape latent is sensitive to the early assembly history of halos. The $\mathrm{MI}_{dM(z)/dz}$ of the same latent reveals that the latter is also sensitive to the later-time mass accretion rate. This dual dependence explains the bimodal shape of the MI between the inner latent and the profile (Fig. 2, bottom panel): the early assembly phase determines the shape of the profile in the innermost region of the halo, while the later-time mass accretion rate determines the shape of the profile close to the virial radius. We further validate this interpretation in the Supplemental Material [52].

The NFW concentration shows a similar picture to the inner shape latent. However, its $\mathrm{MI}_{M(z)}$ peaks at earlier times ($z \sim 0.55$) compared to the inner shape latent. This implies that the inner shape latent carries information about the build up of mass onto the halo over a longer period of time than concentration, which therefore affects the inner halo structure (and the profile) out to larger scales. The sensitivity of the inner latent to later times/larger scales in the profile explains why the inner latent $\mathrm{MI}_{\rho_{\mathrm{true}}(r)}$ is shifted toward larger radial scales than that of concentration (Fig. 2). Moreover, the absolute magnitude of the concentration $\mathrm{MI}_{M(z)}$ is higher than that of the inner shape latent; this is because the closer to the halo core, the stronger the correlation with the early assembly history due to halos accreting mass "inside out." As a result, concentration, which is sensitive to the profile on smaller r than the latent, has a higher MI with the early assembly history than the latent. Finally, the NFW concentration is related to the later-time mass accretion rate in a similar way to the inner latent.

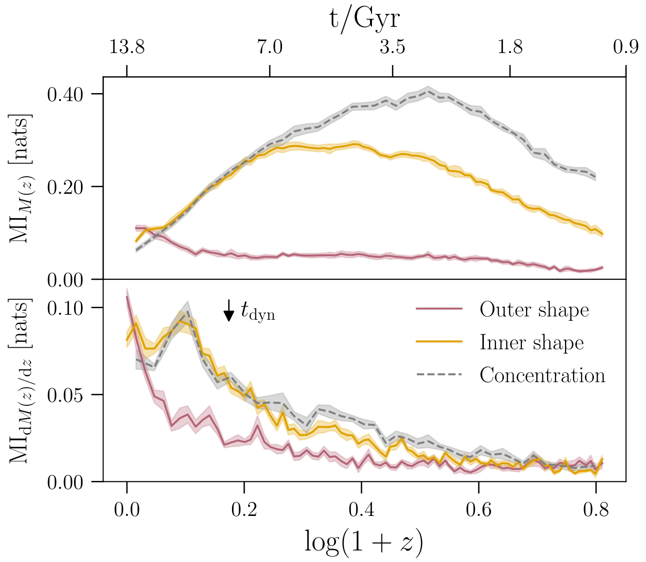

FIG. 3. The MI between the latent parameters and the mass accretion histories (denoted $\mathrm{MI}_{M(z)}$; top) and that between the latent parameters and the mass accretion rate (denoted $\mathrm{MI}_{\mathrm{d}M(z)/\mathrm{d}z}$; bottom). The inner shape latent and the NFW concentration carry memory of the early-time mass assembly history, as well as the later-time mass accretion rate. The outer shape latent carries information about the halos' most recent mass accretion rate over the past dynamical time (indicated by the arrow).

Figure 3 shows that the outer shape latent shares information about the mass accretion rate over the past ~5 Gyr, since this is the period over which the MI roughly doubles (see bottom panel). This timescale corresponds to the halo dynamical time, $t_{\mathrm{dyn}} \equiv 2 \times r_{200\,\mathrm{m}}/v_{200\,\mathrm{m}}$, defined as the time it takes for material to cross the halo at a typical virial velocity $v_{200\,\mathrm{m}} = \sqrt{GM_{200\,\mathrm{m}}/r_{200\,\mathrm{m}}}$. This suggests that the outer profile is primarily determined by the infall of dynamically unrelaxed material within the last dynamical time, which has not yet virialized within the halo.

Discussion.—Our results show that the IVE framework has extracted a direct connection between the assembly history of cold dark matter halos and their density profiles, without having access to explicit information about the time evolution of the halos during training. This has deep implications for understanding the origin of universality in dark matter halos; the universality in the profiles, captured by 3 degrees of freedom alone, may originate from a universality in the halo assembly histories themselves, since the latents contain comparable amounts of information about both quantities.

Previous work [26] found a resemblance between the shape of the average mass accretion history, expressed in terms of the critical density of the Universe, and the average enclosed mass profile, expressed in terms of its enclosed density, for a selected set of "well-behaved" halos of similar mass. In the halo outskirts, the profile has been linked to the dynamical accretion history of the halos primarily through the relation between the splashback radius and the mass accretion rate [43,48]; existing models make use of multiparameter fitting functions to capture the dynamical impact on the outer profile [49].

By contrast, within the IVE framework, the connection between the density profiles and the *entire* mass accretion history or mass accretion rate is clearly elucidated through MI. This result was obtained using all halos in the simulations, without requiring a curated sample of well-behaved halos. The IVE rediscovers the known correlation between the inner profile and halo formation time [35,36]; it then additionally demonstrates that the complexity of the dynamical, infalling material is encoded in only a single degree of freedom that captures the recent mass accretion rate. In future work, we will use the connection between assembly history, latents, and density profile captured by the IVE framework to build a model that can determine mass accretion histories from density profiles.

In future work, we will explore extensions to this work using hydrodynamical simulations. We expect the same IVE framework to successfully disentangle the relevant factors in the baryonic case. Previous work found that baryons primarily impact the inner profile [61–63] and that the results can still be encoded with minimal modifications to the pure dark matter expectations [64,65]. Conversely, the splashback radius in the halo outskirts remains unchanged when comparing hydrodynamical and dark-matter-only simulations [66]. Thus, an IVE with the same dimensionality or a single additional dimension should suffice to account for the impact of baryonic physics on the halo profiles for the baryonic feedback models included in the training set simulations.

More broadly, our results represent progress toward enabling *new* machine-assisted scientific discoveries, going beyond artificial rediscovery of known physical laws [67–69]. Our IVE approach toward this goal consisted of compressing the information within a dataset into a set of minimal ingredients that disentangles the independent factors of variation in the output (interpretability) and can be explained in terms of the physics it represents through MI (explainability). The approach shows promise for gaining insight into other emergent properties of the cosmic large-scale structure (e.g., void density profiles [70] and the halo mass function [71]), building physical explanations that are more accurate and complete than traditional methods have achieved.

L. L.-S. thanks Benedikt Diemer, Eiichiro Komatsu, and Simon White for useful discussions. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreements No. 818085 GMGalaxies and No. 101018897 CosmicExplorer). This work has been enabled by support from the research project grant "Understanding the Dynamic Universe" funded by the Knut and Alice Wallenberg Foundation under Dnr KAW 2018.0067. The work of H. V. P. was additionally supported by the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine. A. P. was additionally supported by the Royal Society. H. V. P. and L. L.-S. acknowledge the hospitality of the Aspen Center for Physics, which is supported by National Science Foundation Grant No. PHY-1607611. The participation of H. V. P. and L. L.-S. at the Aspen Center for Physics was supported by the Simons Foundation. This work was partially enabled by the University College London (UCL) Cosmoparticle Initiative.

The contributions from the authors are as follows. L. L.-S.: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, and writing (original draft, review, and editing). H. V. P.: conceptualization, methodology, interpretation, validation, and writing (review and editing). A. P.: methodology and writing (review and editing).

[*]luisals@mpa-garching.mpg.de

[1] J. M. Bardeen, J. R. Bond, N. Kaiser, and A. S. Szalay, The statistics of peaks of Gaussian random fields, Astrophys. J. **304**, 15 (1986).

[2] G. R. Blumenthal, S. M. Faber, J. R. Primack, and M. J. Rees, Formation of galaxies and large-scale structure with cold dark matter, Nature (London) **311**, 517 (1984).

[3] M. Davis, G. Efstathiou, C. S. Frenk, and S. D. M. White, The evolution of large-scale structure in a universe dominated by cold dark matter, Astrophys. J. **292**, 371 (1985).

[4] C. S. Frenk, S. D. M. White, M. Davis, and G. Efstathiou, The formation of dark halos in a universe dominated by cold dark matter, Astrophys. J. **327**, 507 (1988).

[5] S. D. M. White and C. S. Frenk, Galaxy formation through hierarchical clustering, Astrophys. J. **379**, 52 (1991).

[6] K. Bechtol *et al.*, Snowmass2021 cosmic frontier white paper: Dark matter physics from halo measurements, in Snowmass 2021, arXiv:2203.07354.

[7] A. Drlica-Wagner *et al.* (LSST Dark Matter Group), Probing the fundamental nature of dark matter with the Large Synoptic Survey Telescope, arXiv:1902.01055.

[8] A. Leauthaud, S. Singh, Y. Luo, F. Ardila, J. P. Greco, P. Capak, J. E. Greene, and L. Mayer, Deep, wide lensing surveys can measure the dark matter halos of dwarf galaxies, Phys. Dark Universe **30**, 100719 (2020).

[9] v. Ivezić *et al.* (LSST Collaboration), LSST: From science drivers to reference design and anticipated data products, Astrophys. J. **873**, 111 (2019).

[10] R. Mandelbaum *et al.* (LSST Dark Energy Science Collaboration), The LSST Dark Energy Science Collaboration (DESC) science requirements document, arXiv:1809.01669.

[11] J. F. Navarro, C. S. Frenk, and S. D. M. White, A universal density profile from hierarchical clustering, Astrophys. J. **490**, 493 (1997).

[12] J. Einasto, On the construction of a composite model for the galaxy and on the determination of the system of galactic parameters, Tr. Astrofiz. Inst. Alma-Ata **5**, 87 (1965).

[13] J. F. Navarro, E. Hayashi, C. Power, A. R. Jenkins, C. S. Frenk, S. D. M. White, V. Springel, J. Stadel, and T. R. Quinn, The inner structure of ΛCDM haloes—III. Universality and asymptotic slopes, Mon. Not. R. Astron. Soc. **349**, 1039 (2004).

[14] J. Wang, S. Bose, C. S. Frenk, L. Gao, A. Jenkins, V. Springel, and S. D. M. White, Universal structure of dark matter haloes over a mass range of 20 orders of magnitude, Nature (London) **585**, 39 (2020).

[15] B. P. Moster, T. Naab, and S. D. M. White, EMERGE—an empirical model for the formation of galaxies since $z \sim 10$, Mon. Not. R. Astron. Soc. **477**, 1822 (2018).

[16] P. Behroozi, R. H. Wechsler, A. P. Hearin, and C. Conroy, UNIVERSEMACHINE: The correlation between galaxy growth and dark matter halo assembly from $z = 0$–10, Mon. Not. R. Astron. Soc. **488**, 3143 (2019).

[17] R. H. Wechsler and J. L. Tinker, The connection between galaxies and their dark matter halos, Annu. Rev. Astron. Astrophys. **56**, 435 (2018).

[18] A. P. Hearin, J. Chaves-Montero, M. R. Becker, and A. Alarcon, A differentiable model of the assembly of individual and populations of dark matter halos, Open J. Astrophys. **4**, 7 (2021).

[19] T. McClintock, E. Rozo, M. R. Becker, J. DeRose, Y.-Y. Mao, S. McLaughlin, J. L. Tinker, R. H. Wechsler, and Z. Zhai, The Aemulus Project. II. Emulating the halo mass function, Astrophys. J. **872**, 53 (2019).

[20] D. Jamieson, Y. Li, R. A. de Oliveira, F. Villaescusa-Navarro, S. Ho, and D. N. Spergel, Field-level neural network emulator for cosmological N-body simulations, Astrophys. J. **952**, 145 (2023).

[21] T. Kacprzak and J. Fluri, Deeplss: Breaking parameter degeneracies in large-scale structure with deep-learning analysis of combined probes, Phys. Rev. X **12**, 031029 (2022).

[22] C. D. Kreisch, A. Pisani, F. Villaescusa-Navarro, D. N. Spergel, B. D. Wandelt, N. Hamaus, and A. E. Bayer, The GIGANTES data set: Precision cosmology from voids in the machine-learning era, Astrophys. J. **935**, 100 (2022).

[23] B. Y. Wang, A. Pisani, F. Villaescusa-Navarro, and B. D. Wandelt, Machine-learning cosmology from void properties, Astrophys. J. **955**, 131 (2023).

[24] A. Akhmetzhanova, S. Mishra-Sharma, and C. Dvorkin, Data compression and inference in cosmology with self-supervised machine learning, Mon. Not. R. Astron. Soc. **527**, 7459 (2023).

[25] L. Lucie-Smith, H. V. Peiris, A. Pontzen, B. Nord, J. Thiyagalingam, and D. Piras, Discovering the building blocks of dark matter halo density profiles with neural networks, Phys. Rev. D **105**, 103533 (2022).

[26] A. D. Ludlow, J. F. Navarro, R. E. Angulo, M. Boylan-Kolchin, V. Springel, C. Frenk, and S. D. M. White, The mass-concentration-redshift relation of cold dark matter haloes, Mon. Not. R. Astron. Soc. **441**, 378 (2014).

[27] B. Diemer and A. V. Kravtsov, A universal model for halo concentrations, Astrophys. J. **799**, 108 (2015).

[28] C. A. Correa, J. S. B. Wyithe, J. Schaye, and A. R. Duffy, The accretion history of dark matter haloes—II. The connections with the mass power spectrum and the density profile, Mon. Not. R. Astron. Soc. **450**, 1521 (2015).

[29] A. D. Ludlow, S. Bose, R. E. Angulo, L. Wang, W. A. Hellwing, J. F. Navarro, S. Cole, and C. S. Frenk, The mass-concentration-redshift relation of cold and warm dark matter haloes, Mon. Not. R. Astron. Soc. **460**, 1214 (2016).

[30] F. Prada, A. A. Klypin, A. J. Cuesta, J. E. Betancort-Rijo, and J. Primack, Halo concentrations in the standard Λ cold dark matter cosmology, Mon. Not. R. Astron. Soc. **423**, 3018 (2012).

[31] S. T. Brown, I. G. McCarthy, S. G. Stafford, and A. S. Font, Towards a universal model for the density profiles of dark matter haloes, Mon. Not. R. Astron. Soc. **509**, 5685 (2021).

[32] A. Huss, B. Jain, and M. Steinmetz, How universal are the density profiles of dark halos?, Astrophys. J. **517**, 64 (1999).

[33] J. Wang and S. D. M. White, Are mergers responsible for universal halo properties?, Mon. Not. R. Astron. Soc. **396**, 709 (2009).

[34] B. Moore, S. Ghigna, F. Governato, G. Lake, T. Quinn, J. Stadel, and P. Tozzi, Dark matter substructure within galactic halos, Astrophys. J. Lett. **524**, L19 (1999).

[35] J. S. Bullock, T. S. Kolatt, Y. Sigad, R. S. Somerville, A. V. Kravtsov, A. A. Klypin, J. R. Primack, and A. Dekel, Profiles of dark haloes: Evolution, scatter and environment, Mon. Not. R. Astron. Soc. **321**, 559 (2001).

[36] R. H. Wechsler, J. S. Bullock, J. R. Primack, A. V. Kravtsov, and A. Dekel, Concentrations of dark halos from their assembly histories, Astrophys. J. **568**, 52 (2002).

[37] D. H. Zhao, Y. P. Jing, H. J. Mo, and G. Börner, Accurate universal models for the mass accretion histories and concentrations of dark matter halos, Astrophys. J. **707**, 354 (2009).

[38] N. Dalal, Y. Lithwick, and M. Kuhlen, The origin of dark matter halo profiles, arXiv:1010.2539.

[39] M. P. Rey, A. Pontzen, and A. Saintonge, Sensitivity of dark matter haloes to their accretion histories, Mon. Not. R. Astron. Soc. **485**, 1906 (2019).

[40] N. Roth, A. Pontzen, and H. V. Peiris, Genetically modified haloes: Towards controlled experiments in λCDM galaxy formation, Mon. Not. R. Astron. Soc. **455**, 974 (2015).

[41] K. Wang, Y.-Y. Mao, A. R. Zentner, J. U. Lange, F. C. van den Bosch, and R. H. Wechsler, Concentrations of dark haloes emerge from their merger histories, Mon. Not. R. Astron. Soc. **498**, 4450 (2020).

[42] S. More, B. Diemer, and A. V. Kravtsov, The splashback radius as a physical halo boundary and the growth of halo mass, Astrophys. J. **810**, 36 (2015).

[43] S. Adhikari, N. Dalal, and R. T. Chamberlain, Splashback in accreting dark matter halos, J. Cosmol. Astropart. Phys. 11 (2014) 019.

[44] X. Shi, The outer profile of dark matter haloes: An analytical approach, Mon. Not. R. Astron. Soc. **459**, 3711 (2016).

[45] S. Adhikari, J. Sakstein, B. Jain, N. Dalal, and B. Li, Splashback in galaxy clusters as a probe of cosmic expansion and gravity, J. Cosmol. Astropart. Phys. 11 (2018) 033.

[46] B. Diemer, The splashback radius of halos from particle dynamics. I. The SPARTA algorithm, Astrophys. J. Suppl. Ser. **231**, 5 (2017).

[47] T. Shin and B. Diemer, What sets the splashback radius of dark matter haloes: Accretion history or other properties?, Mon. Not. R. Astron. Soc. **521**, 5570 (2023).

[48] B. Diemer and A. V. Kravtsov, Dependence of the outer density profiles of halos on their mass accretion rate, Astrophys. J. **789**, 1 (2014).

[49] B. Diemer, A dynamics-based density profile for dark haloes—II. Fitting function, Mon. Not. R. Astron. Soc. **519**, 3292 (2022).

[50] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, beta-VAE: Learning basic visual concepts with a constrained variational framework, in *Proceedings of the ICLR* (2017).

[51] V. Springel, R. Pakmor, O. Zier, and M. Reinecke, Simulating cosmic structure formation with the Gadget-4 code, Mon. Not. R. Astron. Soc. **506**, 2871 (2021).

[52] See Supplemental Material [52] at http://link.aps.org/supplemental/10.1103/PhysRevLett.132.031001 for more details on the numerical simulations, which includes Refs. [53–56].

[53] Planck Collaboration *et al.*, Planck 2018 results. VI. Cosmological parameters, Astron. Astrophys. **641**, A6 (2020).

[54] S. Stopyra, A. Pontzen, H. Peiris, N. Roth, and M. P. Rey, Genetic—a new initial conditions generator to support genetically modified zoom simulations, Astrophys. J. Suppl. Ser. **252**, 28 (2021).

[55] V. Springel, The cosmological simulation code GADGET-2, Mon. Not. R. Astron. Soc. **364**, 1105 (2005).

[56] A. Pontzen and M. Tremmel, Tangos: The agile numerical galaxy organization system, Astrophys. J. Suppl. Ser. **237**, 23 (2018).

[57] B. Pandey and S. Sarkar, How much a galaxy knows about its large-scale environment?: An information theoretic perspective, Mon. Not. R. Astron. Soc. **467**, L6 (2017).

[58] S. Sarkar and B. Pandey, A study on the statistical significance of mutual information between morphology of a galaxy and its large-scale environment, Mon. Not. R. Astron. Soc. **497**, 4077 (2020).

[59] N. Sedaghat, M. Romaniello, J. E. Carrick, and F.-X. Pineau, Machines learn to infer stellar parameters just by looking at a large number of spectra, Mon. Not. R. Astron. Soc. **501**, 6026 (2021).

[60] D. Piras, H. V. Peiris, A. Pontzen, L. Lucie-Smith, N. Guo, and B. Nord, A robust estimator of mutual information for deep learning interpretability, Mach. Learn. **4**, 025006 (2023).

[61] J. F. Navarro, V. R. Eke, and C. S. Frenk, The cores of dwarf galaxy haloes, Mon. Not. R. Astron. Soc. **283**, L72 (1996).

[62] A. R. Duffy, J. Schaye, S. T. Kay, C. Dalla Vecchia, R. A. Battye, and C. M. Booth, Impact of baryon physics on dark matter structures: A detailed simulation study of halo density profiles, Mon. Not. R. Astron. Soc. **405**, 2161 (2010).

[63] A. Pontzen and F. Governato, Cold dark matter heats up, Nature (London) **506**, 171 (2014).

[64] A. Di Cintio, C. B. Brook, A. A. Dutton, A. V. Macciò, G. S. Stinson, and A. Knebe, A mass-dependent density profile for dark matter haloes including the influence of galaxy formation, Mon. Not. R. Astron. Soc. **441**, 2986 (2014).

[65] M. A. Henson, D. J. Barnes, S. T. Kay, I. G. McCarthy, and J. Schaye, The impact of baryons on massive galaxy clusters: Halo structure and cluster mass estimates, Mon. Not. R. Astron. Soc. **465**, 3361 (2017).

[66] H. Aung, D. Nagai, and E. T. Lau, Shock and splash: Gas and dark matter halo boundaries around ΛCDM galaxy clusters, Mon. Not. R. Astron. Soc. **508**, 2071 (2021).

[67] R. Iten, T. Metger, H. Wilming, L. del Rio, and R. Renner, Discovering physical concepts with neural networks, Phys. Rev. Lett. **124**, 010508 (2020).

[68] A. Seif, M. Hafezi, and C. Jarzynski, Machine learning the thermodynamic arrow of time, Nat. Phys. **17**, 105 (2021).

[69] S.-M. Udrescu and M. Tegmark, AI Feynman: A physics-inspired method for symbolic regression, Sci. Adv. **6**, eaay2631 (2020).

[70] N. Hamaus, P. M. Sutter, and B. D. Wandelt, Universal density profile for cosmic voids, Phys. Rev. Lett. **112**, 251302 (2014).

[71] J. Tinker, A. V. Kravtsov, A. Klypin, K. Abazajian, M. Warren, G. Yepes, S. Gottlöber, and D. E. Holz, Toward a halo mass function for precision cosmology: The limits of universality, Astrophys. J. **688**, 709 (2008).