Modeling assembly bias with machine learning and symbolic regression

Digvijay Wadekar[a,1], Francisco Villaescusa-Navarro[b,c], Shirley Ho[b,c,d], and Laurence Perreault-Levasseur[c,e,f]

[a]Center for Cosmology and Particle Physics, Department of Physics, New York University, New York, NY 10003; [b]Department of Astrophysical Sciences, Princeton University, Peyton Hall, Princeton NJ 08544-0010; [c]Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, 10010, New York, NY; [d]Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15217; [e]Department of Physics, Université de Montréal, Montréal, Canada; [f]Mila - Quebec Artificial Intelligence Institute, Montréal, Canada

Upcoming 21cm surveys will map the spatial distribution of cosmic neutral hydrogen (HI) over unprecedented volumes. Mock catalogues are needed to fully exploit the potential of these surveys. Standard techniques employed to create these mock catalogs, like Halo Occupation Distribution (HOD), rely on assumptions such as the baryonic properties of dark matter halos only depend on their masses. In this work, we use the state-of-the-art magneto-hydrodynamic simulation IllustrisTNG to show that the HI content of halos exhibits a strong dependence on their local environment. We then use machine learning techniques to show that this effect can be 1) modeled by these algorithms and 2) parametrized in the form of novel analytic equations. We provide physical explanations for this environmental effect and show that ignoring it leads to underprediction of the real-space 21-cm power spectrum at $k \gtrsim 0.05\,h\,\mathrm{Mpc}^{-1}$ by $\gtrsim$10%, which is larger than the expected precision from upcoming surveys on such large scales. Our methodology of combining numerical simulations with machine learning techniques is general, and opens a new direction at modeling and parametrizing the complex physics of assembly bias needed to generate accurate mocks for galaxy and line intensity mapping surveys.

cosmology | machine learning | hydrodynamic simulation |

I n the coming decade, numerous astronomical surveys will gather vast amounts of data about the Universe. To understand all that can be learned about the contents and evolution of the Universe from this data, scientists make predictions using different astrophysical and cosmological parameters. They organize these forecasts into catalogs of mock data, which can later be compared to the astronomical survey observations to infer the actual parameters describing the Universe. Astrophysicists use computer simulations of various types and levels of physical detail to build these catalogs. In particular, there has been a lot of recent progress in developing hydrodynamic simulations, which include the effects of star formation, gas cooling, magnetic fields, and energetic feedback due to supernovae and supermassive black holes. That is, these simulations include detailed descriptions of both dark matter (DM), which has played a vital role in shaping the Universe's large-scale structure, and baryonic matter, which makes up the visible material in galaxies and galaxy clusters caught in the web of that structure. The upcoming surveys, however, will span volumes of 10–100 $(\mathrm{Gpc/h})^3$, and the hydrodynamic simulations typically cost on the order of 10 million CPU hours for a mere 10^{-3} $(\mathrm{Gpc/h})^3$ (1), making it impossible to use them directly for generating the required catalogs of large-scale mock data. Less computationally expensive mock simulations will be needed to fill this gap.

One of the most popular theoretical techniques used to cheaply emulate the expensive hydrodynamic simulations and create large-scale mock baryonic data are halo models (also referred to as halo occupation distribution (HOD) models). HOD was first used to probabilistically model the number of galaxies residing in a host DM halo (2–5) and it typically assumes a simple parametric relation between the halo's mass and its baryonic properties (e.g., the number of galaxies, stellar mass, or neutral hydrogen content of the halo). To calibrate the parameters in this relation, one typically uses semi-analytic models or hydrodynamical simulations (and observations, if available). The best-fit parametric relation is then applied to halos generated by gravity-only N-body simulations to make mock baryonic simulations.

HOD has frequently been used to make large-volume mock simulations, both for galaxy surveys (6–20) and for intensity mapping surveys (21–26). Such mocks are used for 1) determining which summary statistics are the most appropriate to constrain different cosmological parameters; 2) studying the effect of various observational systematics on summary statistics and testing the range up to which perturbative models are robust for parameter inference; 3) constructing covariance matrices; and 4) simulation-based inference (SBI) analyses for parameter inference.

The standard HOD technique assumes that the properties of various baryonic structures inside a halo are governed *solely by the halo mass*, and ignores all other ("secondary") halo properties. Other techniques used to make mocks rely on similar assumptions. For example, sub-halo abundance matching (SHAM; see e.g. (27, 28)) assumes the existence of a scatter-free monotonic relation between a halo's mass and the numbers or masses of the baryonic tracers in it. However, numerous studies using hydrodynamical simulations and semi-analytic models have found that the clustering of galaxies is in fact affected by secondary properties other than halo mass, such as the halo environment, halo assembly history, concentration, spin, velocity anisotropy, and many others (29–36). This phenomenon is referred to as *galaxy assembly bias*[*].

Studying galaxy assembly bias has been an important task as inaccurate galaxy mocks can lead to biases in the inferred cosmological parameters and galaxy formation properties. There has also been some recent interest in this topic because of discrepancies in results inferred using the standard HOD model with the Planck cosmology on a simultaneous comparison to galaxy-galaxy lensing and projected galaxy clus-

---

[*]Note that galaxy assembly bias is different from halo assembly bias, which refers to the dependence of clustering of the DM halos themselves on secondary properties other than their mass (37–40). Halo assembly bias is automatically accounted for in an HOD analysis when halos from an N-body simulation are used. Note also that the term "assembly bias" was first used to characterize the effect of a particular secondary property—the halo assembly history (41, 42)—but the term is now used more generally for the effect of all secondary properties of halos.

[1]Corresponding author. E-mail: jay.wadekar@nyu.edu

tering data from the BOSS CMASS and BOSS LOWZ samples ([15]–[18], [43], [44]). Several studies have tried to incorporate secondary halo parameters into an HOD framework to include the effect of assembly bias. Halo concentration has been a traditionally popular secondary parameter ([32]–[35], [45]), although numerous recent studies have shown that the environment of the halos plays a significant role for modeling the galaxy distribution ([16], [36], [46]–[50]).

All the works mentioned above have focused on understanding the galaxy-halo connection. However, with numerous upcoming surveys recording at the 21-cm wavelength (CHIME, HIRAX, HERA, TIANLAI, FAST, ASKAP, MeerKAT and SKA), which will probe the spatial distribution of neutral hydrogen (HI) in the Universe, it now becomes imperative to also understand the connection between the properties of DM halos and their HI content.

In this paper we quantify, for the first time, the effects of secondary properties of halos (i.e., properties other than their mass) on the clustering of HI (i.e., HI assembly bias) using the state-of-the-art IllustrisTNG magneto-hydrodynamic simulation. Naively, one might assume that baryons trace dark matter and therefore the HI content of a halo should only depend on its total mass. However, we will show that this assumption is invalid and other halo properties such as the halo environment also have a crucial effect. Furthermore, also for the first time, we model HI assembly bias using machine learning and symbolic regression in order to enable the creation of the more-accurate HI catalogs needed to analyze data from the upcoming 21-cm surveys.

The paper is organized as follows. In sections [1] and [2], we describe the hydrodynamic simulation and the details of the HOD model that we use. In Sec. [3], we quantify the effect of various halo secondary properties on halos' HI masses and in its subsection [A] we discuss the physical reasons underlying these effects. In Sec. [4], we model the HI-halo connection using symbolic regression. In Sec. [5], we present the clustering of HI from the different models. Finally, we discuss the relevance of our results, compare our approach to others in the literature, and conclude in Sec. [6]. Before we begin our analysis, let us first discuss the motivation for using machine learning and symbolic regression to model the HI assembly bias in the following three subsections.

A. Motivation for halo environment from saliency maps of neural networks. Apart from theoretical techniques like HOD and SHAM which work on halo catalogs, machine learning tools like deep neural networks (DNNs) can be used to emulate expensive cosmological simulations directly at the field level ([51]–[62]). A number of studies have shown that neural networks outperform traditional tools like HOD in emulating expensive simulations when various statistical properties of the emulated field are compared ([51]–[56]). One question that arises here is whether there are any particular features in the input DM maps that the DNNs are using for their emulation and if such features can be used to augment standard HOD models. However, one of the challenges to answer this question is that the DNNs are notoriously difficult to interpret; this is due to the large number of fitted parameters (weights and biases) and also the depth of the many layers in a deep network. There are however a few methods like saliency maps which can be used for getting insights into deep learning models ([63], [64]) and we will discuss them below.
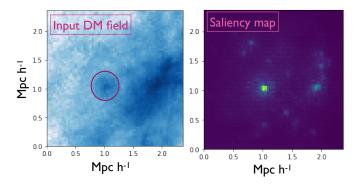

Fig. 1. Left: Dark matter (DM) density field in a particular sub-cube of the TNG100 simulation where we have identified a $10^{12} \, h^{-1} \, M_\odot$ halo (red circle). This field was provided as input to the U-Net trained in W20 to predict the HI field. **Right:** Saliency map where brightness roughly corresponds to the importance of the input-field voxels used by the DNN to predict the HI inside the circled halo. The bright regions well outside the halo indicate that the DNN used not only the local halo information, but also information in the halo's environment to make its prediction. We find that the predicted HI content in the circled halo is ∼15% lower than if the same halo would be in an isolated environment.

In our previous work, Ref. ([51]) (hereafter W20), we used a convolutional DNN to model the HI field from an input matter field and showed that it outperforms HOD for emulating all summary statistics of the output HI field (∼ 15% improvement for the HI power spectrum upto non-linear scales $k \leq 1 \, h \, \mathrm{Mpc}^{-1}$). As an the input to the DNN, W20 used a high-resolution 3D matter field over a cube with side length $2.34 \, h^{-1}$ Mpc. For modeling HI in a halo in the input field, the DNN therefore has access to information like the local environment of the halo and also the mass distribution inside the halo. We are interested in roughly inferring what information is used by the DNN to make its prediction for HI. To answer this question, we show, as an example, the saliency map corresponding to a particular case when a DM halo is in a tidal environment in Fig. [1]. One can visually see that the DNN models the information in the environment of the halo and uses it when predicting the HI inside the circled halo. Furthermore, it is extremely interesting to see that the DNN lowers the HI content of a halo when the halo is placed in an extremely overdense environment. This is analogous to the astrophysical effect called ram pressure stripping where gas escapes galaxies which are in a dense cluster due to the pressure from the surrounding ionized medium ([65]).

One question that still arises at this point is, what fraction of the network prediction comes by looking at the matter distribution outside halos against the one coming from inside the halo? Answering this question using the DNN is a nontrivial task, given the complex nature of the flow of information within a typical DNN. Furthermore, a common issue with DNNs is their generalizability, i.e their use on datasets with different hyperparameters than the ones they are trained on (for e.g. on an input field with a different resolution than the training sample). There is also a problem of data sparsity when using DNNs directly at the field level: most of the cosmological information comes from halos, which are found rarely in the input matter field ([51], [52], [54], [57]).

B. Random forests for modeling assembly bias. In this paper we perform our analysis directly on the DM halo catalog rather than working on the field level. This drastically reduces the

size of the dataset and therefore enables us to use traditional machine learning techniques like random forest regressors (RF), which are relatively less expensive and more interpretable than DNNs. One other advantage is that we can interface with the huge amount of theoretical work done on halo models and we can quantify and understand the effect of various halo properties on its HI mass (M_{HI}). Our goal is in this paper is to model M_{HI} by approximating the function

$$M_{HI} = f(M_h, \{i_h\}) \quad [1]$$

where M_h is the total mass of the halo, and $\{i_h\}$ corresponds to the set of various secondary halo properties: the overdensity and anisotropy of its environment at various scales, concentration, assembly history like, formation epoch, spin, velocity dispersion...etc. The high dimensionality of the input space makes this a complex and challenging problem. There are also correlations between different input parameters (e.g., the concentration of a halo is related to its assembly history and also its environment), which add to the complexity. Machine learning tools like RFs are well-suited for approximating functions in a high-dimensional input parameter space; the advantage of using these methods over traditional theoretical methods is that there is no need to know the underlying functional form; only samples from that function are needed.

C. Symbolic regression for parameterizing assembly bias. Symbolic regression (SR) is a technique that approximates the relation between an input and an output through analytic mathematical formulae (66–74). The advantage of using SR over other machine learning regression models is that it provides analytic expressions which can be readily generalized and which facilitate the understanding of the underlying physics. One of the downsides of SR, however, is that the dimensionality of the input space needs to be relatively small. In our case, we first use RF to get an indication of which parameters in the set of $\{i_h\}$ in Eq. 1 have the largest effect on M_{HI}. We then compress the $\{i_h\}$ set to include only the five most important parameters. Finally, in Sec. 4, we use SR on the compressed set to obtain an explicit functional form to approximate f from Eq. 1. Having an analytic form with a minimal set of parameters that captures assembly bias is crucial in order to detect this effect through Bayesian analysis of real survey data. In this study, we use the symbolic regressor based on genetic programming implemented in the publicly available PySR package[†] (69, 75). We leave further details on RF and SR to Appendix D.

1. Data

We use data from the TNG300-1 simulation produced by the IllustrisTNG collaboration (1)[‡] throughout this paper. This simulation is a state-of-the-art magneto-hydrodynamic simulation that includes a wide range of relevant physical effects, such as radiative cooling, star formation, metal enrichment, supernova and AGN (active galactic nuclei) feedback, and magnetic fields. We will use two redshifts in our analysis: $z = 5$ and $z = 1$, corresponding to early and late times in the post-reionization era. It is worth mentioning that IllustrisTNG has already been used in multiple studies of galaxy assembly bias (36, 50, 76–79). We show our results for the TNG300-1
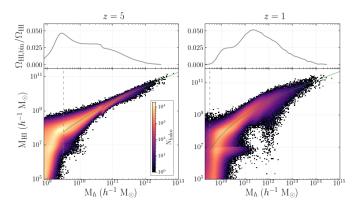
Fig. 2. Bottom: Distribution of HI mass (M_{HI}) versus total mass (M_h) for halos in the TNG300-1 simulation at redshifts $z = 5$ (*left*) and $z = 1$ (*right*). Color coding indicates the number of halos in that region of parameter space. The best-fit for the HI–halo mass function (see Eq. 2) is shown in green and corresponds to the mass-only HOD prediction in the figures below. The dashed gray lines indicate the mass cutoff used in our analysis, that represents the mass of halos with ~50 DM particles. **Top:** The fractional contribution to Ω_{HI} of each halo mass bin.

simulation as it covers the largest volume among the TNG boxes. We also performed the analysis for the TNG100-1 simulation, which has a higher resolution than TNG300-1 but a smaller box size; we have checked that the HI assembly bias results for TNG300-1 are qualitatively similar to the ones from TNG100-1, and therefore our results are robust against volume and resolution effects.

2. HOD model for HI

Halo models have been traditionally popular for modeling galaxies and there have been multiple recent attempts at developing a halo model for the abundance and spatial distribution of HI (21–24, 26, 80–85). The main idea behind such models is that most of the HI in the post-reionization era resides inside halos: more than 99% at $z < 0.2$ (the fraction decreases to ~88% at $z = 0.5$) (24). We can use this fact to generate HI fields by populating halos in an N-body simulation with HI and this method is called as HOD.[§] In this study, we will use the HOD model of Ref. (24) (hereafter VN18), which has also been used to make gigaparsec volume HI mocks (25). We briefly describe their model in what follows and refer the reader to VN18 for further details. The first step involves running a DM-only simulation, identifying halos, and saving their positions and masses. A DM halo of mass M_h is then assigned an HI mass (denoted by M_{HOD-HI}) using the HI−halo mass relation given by:

$$M_{HOD-HI}(M_h, z) = M_0 \left(\frac{M_h}{M_{min}}\right)^\alpha \exp[-(M_{min}/M_h)^{0.35}] \quad [2]$$

where M_0 is a normalization factor, α is the power-law slope, and M_{min} is the characteristic minimum mass[¶] of halos that host HI. We calibrate this relation using halos from the TNG300-1 simulation; we only consider halos with masses above $10^{9.5} h^{-1} M_\odot$, which have ~ 50 bound DM particles, to ensure our sample is robust. We get the best-fit values for $z = 1$ ($z = 5$) to be: $M_0 = 0.64 (0.2) \times 10^{10} h^{-1} M_\odot$,

$M_{\min} = 25.41\,(2.36) \times 10^{10}\,h^{-1}\,M_\odot$ and $\alpha = 0.52\,(0.76)$. We show the corresponding fits in Fig. 2. Note that the best-fit parameter values are different from those in VN18, which were calibrated for the TNG100-1 simulation. This is caused by 1) resolution effects, that affect the strength of the astrophysical effects such as AGN and supernova feedback, and 2) the different choice of halo mass cutoff for calibrating the $M_{\text{HOD-HI}}$ relation. One can immediately note from Fig. 2 that there is a large scatter in the H I -halo mass relation at fixed M_h. We will later show that a part of this scatter is due to halo environment and discuss its impact on the clustering of the modeled H I field. As we are interested in analyzing large scales in the paper, we have ignored the one-halo term (which account for distribution of H I within the halo) in our model throughout the paper and assumed the entire H I is located at the center of halo; the one halo term only becomes important on scales $k \gtrsim 1\,h\,\text{Mpc}^{-1}$ that are not relevant for this work.

We emphasize that the best-fit values quoted earlier for $\{M_0, \alpha, M_{\min}\}$ were obtained by using halo parameters from the hydrodynamical (or full physics, FP) IllustrisTNG simulation and therefore these best-fit values *should not be directly used on halos in a N-body (dark matter only, DMO) simulation.* This is because baryonic effects have a significant effect on the halo mass (31, 86). Using the value of the free parameters calibrated from the FP simulation, directly for a N-body simulation, would lead to significant discrepancies: for e.g., Ω_{HI} will change by ∼15%. We have separately calibrated the fitting formula in Eq. 2 using halos from the phase matched DMO version of the TNG300-1 simulation and we present the corresponding best-fit parameters in Appendix C. We leave further discussion of the differences in the halo masses between the FP and DMO simulations for Fig. S3 and Appendix C.

3. Effect of secondary halo properties on its H I mass

VN18 showed that halos with the same mass but different H I content cluster differently. This hints at the fact that the H I content of the halo is affected by secondary properties of the halo rather than M_h alone. We explore this effect in detail in this section. We focus on 1) halo environmental parameters and 2) internal halo properties such as mass fraction in subhalos and concentration.

Let us first outline the procedure that we use to calculate the environmental overdensity (δ_R) within a radius R (h^{-1} Mpc) surrounding the halos. We compute the matter overdensity field δ of the TNG300-1 box on a $N_g = 2048^3$ grid using CIC interpolation and then smooth in Fourier space by a top-hat filter with radius R. We then transform the smoothed field back to real space and interpolate the smoothed field to the locations of the halos. We finally calculate the overdensity after subtracting the contribution from the halo mass, which is equivalent to using

$$1 + \delta_R \equiv \frac{1}{\bar{\rho}} \frac{M_R - M_h}{4/3\,\pi R^3}\,, \qquad [3]$$

where M_R is the total mass within a radius $R^{\|}$. The subtraction is made to make δ_R independent of halo mass. It is worth mentioning that some studies in the literature use a Gaussian kernel instead of a top-hat kernel for smoothing the density

field to calculate the effect of the environment. In that case, however, the environmental variable has some contribution from the halo mass itself and we choose our definition to make δ_R independent of halo mass.

Apart from the environmental overdensity, we also explore the dependence of M_{HI} on the anisotropy of the matter distribution around halos. There is increasing evidence that the tidal anisotropy is a key factor in determining halo assembly bias (39, 87–90), and we therefore investigate whether it affects the H I clustering. In order to quantify the anisotropy, we first calculate a dimensionless version of the tidal tensor as $T_{ij} \equiv \partial^2 \phi_R / \partial x_i \partial x_j$, where ϕ is the dimensionless potential field calculated using Poisson's equation: $\nabla^2 \phi_R = -\rho_R/\bar{\rho}$. We then calculate the tidal shear q_R^2 using[**] (93, 94)

$$q_R^2 \equiv \frac{1}{2}\left[(\lambda_2 - \lambda_1)^2 + (\lambda_3 - \lambda_1)^2 + (\lambda_3 - \lambda_2)^2\right] \qquad [4]$$

where λ_i are the eigenvalues of T_{ij}. However, Ref. (39) showed that tidal shear on small scales is also correlated with the environmental overdensity and in order to isolate the anisotropy effect one needs to appropriately normalize the shear. We adopt the normalization of the shear proposed by Ref. (39) which is

$$\alpha_R \equiv \frac{\bar{\rho}}{\rho_R}\sqrt{q_R^2} \qquad [5]$$

where the scalar parameter α_R is referred to as the tidal anisotropy parameter and efficiently encodes the tidal information. Ref. (39) also recommended to use an adaptive top-hat smoothing scale $4\,R_{200}$ for individual halos; we however use a global smoothing scale for all halos as it is computationally more straightforward to calculate using fast fourier transforms (FFTs). The smallest scale for smoothing that we consider is $0.5\,h^{-1}$ Mpc, as it becomes increasingly expensive to smooth to smaller scales because a larger grid with a finer resolution is needed. We show the relation of M_{HI} with the halo environmental overdensity and anisotropy in Fig. 3. The mean of the trends are shown in solid and we find a large scatter in the relation as seen from the shaded areas representing 1σ deviations (we will later see in Fig. 5 and Sec. 5 that the mean trends have a significant effect on the clustering statistics of H I and the scatter largely gets averaged out). We pick two scales to show the dependence of the results with the smoothing scale: $R = 0.5\,h^{-1}$ Mpc for the small-scale environment, and $R = 5\,h^{-1}$ Mpc for the large-scale environment. We did not show the trends for smoothing scales larger than $5\,h^{-1}$ Mpc because the effect of the halo environment starts to become less significant. We do not show lines corresponding to high-mass halos ($M_h \gtrsim 10^{12.5}\,h^{-1}\,M_\odot$) for $z = 5$ (as such halos are rare at high redshifts) and for $z = 1$ (as such halos dominate the small-scale environment and $\delta_{0.5} \to 0$ as per Eq. 3). We also have not shown the number of halos in a particular region of the parameter space in Fig. 3 (there are typically much fewer halos towards the high end of the environmental parameter values).

Overall, we see that the effect of halo environment is more pronounced for low-mass halos $M_h \lesssim 10^{11.5}\,h^{-1}\,M_\odot$. Note that such low-mass halos do not typically host galaxies that can be observed in galaxy surveys: they would be too faint. On the other hand, these halos, due to the large abundance,

---

[∥] We assign $\delta_R = 0$ when $M_R < M_{\text{halo}}$ which roughly happens when the $R < R_{\text{virial}}$ for large halos.

[**] Note that perturbation theory based models use a closely related variable $s^2 \equiv 2q^2/3$ for studying the non-local bias (91, 92).
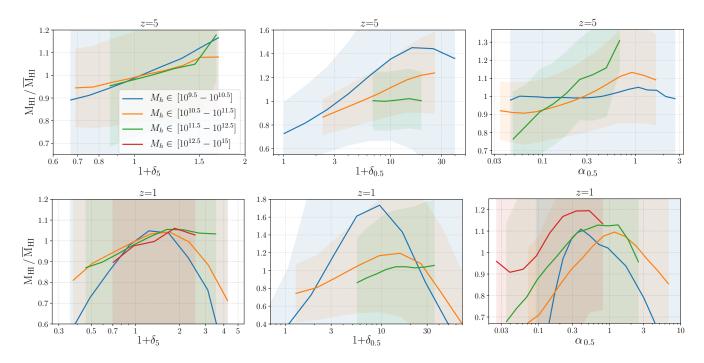
Fig. 3. This figure shows the effect of various halo environmental properties on their HI masses (M$_{HI}$). δ_R (α_R) is the environmental overdensity (anisotropy) within a radius $R/(h^{-1}\,\mathrm{Mpc})$ surrounding the halo as defined in Eq. 3 (Eq. 5). We split the halos in the TNG300-1 volume into four mass bins and show the ratio of their HI mass to the average HI mass of halos in the corresponding mass-bin; solid lines are the mean relations and shaded areas enclose 68% of the data. If M$_{HI}$ would only depend on M_h, the result will be constant value of one on the y-axis; we instead see strong mean trends for effect of the environment (especially for the low mass halos). There is also a large scatter, which is a result of the highly stochastic nature of gas accumulation in a halo. See Sec. 3A for a physical explanation of the mean trends. All masses are in units of $h^{-1}\,M_\odot$.
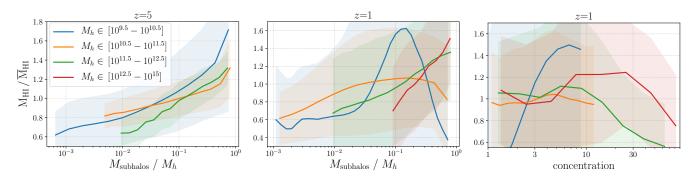

Fig. 4. Same as Fig. 3 but for the effect of internal halo properties like the fractional mass in subhalos (left and middle) and concentration on the halo HI mass (right). These trends are shown to complement the physical explanations outlined in Sec. 3A for the effect of the environment seen on the halo HI mass in Fig. 3.

host a very large fraction of the total HI mass, as seen in the top panel of Fig. 2.

A. Physical explanations for the environmental trends. Let us now provide a physical interpretation to some of the trends of M_{HI} with environmental variables in Fig. 3. A key physical effect for understanding the HI content of a halo is that, in the presence of ionizing radiation, HI can only form in places where the gas has sufficiently high density to self-shield itself. Let us, for clarity, discuss explanations of the environmental trends for the two redshift cases separately:

High-z case ($z = 5$): Increasing the environmental over-density typically results in more mergers and therefore more substructure in halos (see Fig. S6). We show in the left panel of Fig. 4 that, at fixed halo mass, increasing the fraction of mass in subhalos increases the HI content of the halo. This is

likely because the gas in subhalos is more dense and can more efficiently self-shield itself as compared to gas in the CGM (circumgalactic medium) of the halo.

Low-z case ($z = 1$): The main difference here, as compared to the high-z case, is that the ionizing metagalactic background and the feedback due to AGN and supernovae are both much stronger, and lead to ionization of HI in low density regions. Let us first discuss the high-mass end: $M_h > 10^{12}\,h^{-1}\,M_\odot$. Due to strong AGN feedback in the central galaxy, a significant fraction of HI is located in the subhalos of these halos (see Fig. 7 of VN18). This explains why we find a correlation between the HI mass and the total mass in subhalos (see central panel of Fig. 4). This explanation breaks down for low-mass halos, where we see a turnover in the trends (especially for smaller halos) when the environment becomes dense or anisotropic

beyond a certain threshold. The turnover likely occurs in cases when a halo is close enough to large objects like galaxy clusters; those halos can lose their gas content due to ram pressure stripping. It is worth noting that this effect is similar to the one in Fig. 1, where the neural network lowers the H I inside a halo when its is located in a highly overdense and anisotropic environment. This hints at the fact that neural networks can learn to model complex astrophysical effects directly from the output data of hydrodynamic simulations.

There is another effect of halo concentration which comes into play for the lowest mass halos $10^{9.5} < M_h < 10^{10.5}\,h^{-1}\,M_\odot$ at $z = 1$. For these halos, the ionizing feedback from the central galaxy is much lower and most of the H I is concentrated in the central galaxy. Having a higher concentration therefore makes it easier to accumulate a large density of gas in the center of the halo (which is required for H I to self-shield itself from the metagalactic radiation), and hence we see the steep rising trend in the right panel of Fig. 4. The low-mass halos in more dense and anisotropic environments typically have higher concentration (37, 39), which could be the reason behind the initial steep rising trend with the environment for such halos in Fig. 3. Among the internal properties of the halo, we have only considered the effect of the subhalo mass-fraction and concentration in this work. We leave further discussion on the halo internal properties to Appendix B, where we also show additional plots corresponding to these parameters.

4. Modeling the halo HI mass with symbolic regression

Our goal in this section is to model the trends observed in Fig. 3 with compact analytic expressions using symbolic regression (SR). As discussed earlier in Sec. C, SR allows us to obtain a functional form that can capture the structure in a high-dimensional dataset, and is therefore an ideal tool for our purposes.

Before we input the various environmental parameters into the symbolic regressor, we rescale the parameters by using logarithms to shorten the range over which these parameters vary: $m_{10} \equiv \log[M_h/(10^{10}\,h^{-1}\,M_\odot)]$ for the halo mass, $\delta'_R \equiv \log(2 + \delta_R)$ for the environmental overdensity and $\alpha'_R \equiv \log(1 + \alpha_R)$ for the environmental anisotropy[††]. We use the SR for modeling the ratio of $M_{\rm HI}$ to the output $M_{\rm HOD\text{-}HI}$ from the mass-only HOD model from Eq. 2. We train the SR separately at two redshifts and get

$$\frac{M_{\rm HI}}{M_{\rm HOD\text{-}HI}} = 0.95 + \alpha'_{0.5}\,\delta'_{0.5}\,(\alpha'_{0.5} + \delta'_{0.5}) \qquad [z = 5] \qquad [6a]$$

$$\frac{M_{\rm HI}}{M_{\rm HOD\text{-}HI}} = 0.81 + 1.44\,\alpha'_{0.5}\,m_{10}$$
$$\qquad\qquad - 0.57\,(\alpha'^2_{0.5}\,m^2_{10} + \alpha'_{0.5}\,\delta'_5) \qquad [z = 1] \qquad [6b]$$

We have presented the most concise expressions that include the effect of environment on the H I mass over the full range of halo masses. It is important to note that the expressions in Eq. 6 are not unique, i.e we have found expressions which fit the TNG data better than the ones in Eq. 6, however their form is relatively much more complex. Furthermore, due to the presence of a large scatter in the environmental relations as seen in Fig. 3, the risk of overfitting goes up as the equations get more complex. A question which arises

_____

[††]Note that the particular constants are added before using the logarithm in order to prevent the rescaled parameters from diverging when $\delta \to -1$ or $\alpha \to 0$.
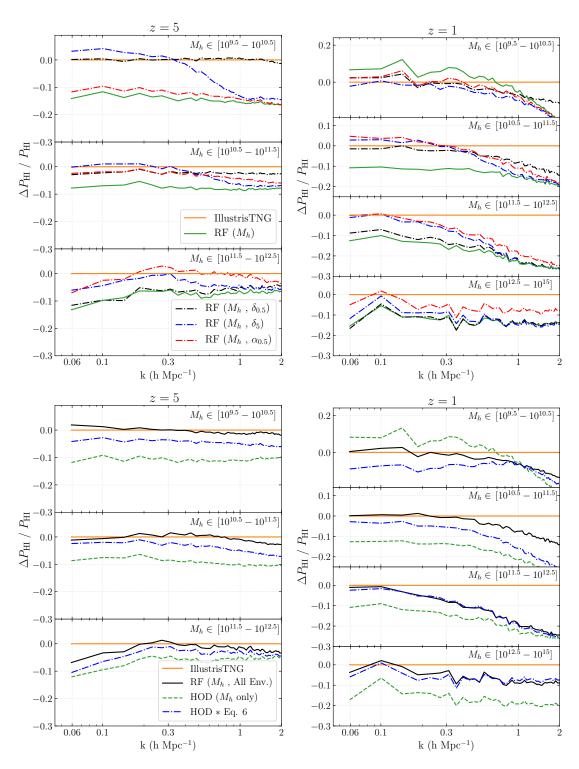
at this point is whether the forms of Eq. 6 are robust when the astrophysical feedback and cosmology parameters in the hydrodynamic simulations are changed. We plan to answer this in a future study using the CAMELS simulations suite (72), which contains multiple hydrodynamic simulations run with different feedback parameters.

Note that the expressions in Eq. 6 do not contain all the environmental parameters seen in Fig. 3 (for e.g., Eq. 6a does not involve the large-scale term δ_5); this is likely because the environment at different scales is correlated and sometimes the information gained from different environmental parameters is degenerate. We also show in Fig. S5 the performance of these expressions. Let us discuss the connections of these equations to some of the trends seen in Fig. 3. For $z = 1$, apart from a constant, there is a linear term with respect to the environmental variables $(1.44\,\alpha'_{0.5}\,m_{10})$ and a corresponding quadratic order term with a negative sign. The negative quadratic order term arises due to ionization of H I because of baryonic feedback. Note that the negative terms are only present in the $z = 1$ case as feedback becomes stronger at low-z. For $z = 5$, one can see that the response of $M_{\rm HI}$ is stronger for $\delta_{0.5}$ ($\alpha_{0.5}$) for the low (high) mass halos. Therefore, a combination of them will give a fairly monotonic trend, which is what we find in Eq. 6a.

It is also worth mentioning that our expressions in Eq. 6 do not capture all the scatter in the H I-halo mass relation seen in Fig. 2. We have only modeled the part of the scatter connected to the environment and, as we will see in the next section, this is sufficient for improving the accuracy of the clustering of H I by a significant amount. See Ref. (71) for a more comprehensive modeling of H I-halo mass scatter and its correlation with various baryonic properties of the halo.

5. Results for clustering of HI

Till now we have discussed the effects of halo environment on its H I mass. In this section, we investigate how this dependence propagates into the clustering of H I (we refer to this phenomenon as H I assembly bias in this paper). We will now focus on the real-space H I power spectrum $P_{\rm HI}(k)$ as it is directly related to the 21-cm power spectrum (which will be measured from upcoming surveys via $P_{\rm 21cm}(k) = \bar{T}_b^2\,P_{\rm HI}(k)$, where $\bar{T}_b$ is the mean brightness temperature of the 21-cm line). For results corresponding to other summary statistics like the bispectrum and cross-correlation coefficient, see Appendix A.

Let us first discuss an intuitive explanation of how the environment of the halos impacts the overall clustering of H I. It is well-known from excursion set theory that the environment of halos has a very strong impact on the clustering of halos themselves (95) (for e.g., it is easier to form halos in a denser environment as the collapse threshold is more frequently reached and such halos therefore are significantly more clustered; this is indeed also the case in the TNG300-1 simulation as shown in Fig. S2). If halos in denser regions have even slightly more H I than those in sparser regions, the overall clustering of H I would increase; this is because the calculation of $P_{\rm HI}$ involves weighting halos by their H I mass and the halos in denser regions would get upweighted. We thus conclude that halo environment not only can affect the H I content of a halo, but also the overall clustering signal of H I.

In order to gauge the effect of various environmental variables on $P_{\rm HI}(k)$, we use a machine learning model called ran-

Fig. 5. This figure illustrates the effects of halo environment on the power spectrum of the modeled H I field. We train random forest (RF) regressors to predict the H I mass of halos as a function of different environmental properties (see legend). The line legends are the same for the left and right panels. **Top:** The RF is trained separately on three different environmental variables and the results show slightly different improvements in each case. **Bottom:** The RF is trained using the environmental information (both the overdensity and tidal information) at various scales. We also show the result from the mass-only HOD model in dashed green which uses Eq. 2 and the dot-dashed blue line shows the results from modeling the halo environment using Eq. 6, which was derived using symbolic regression. Overall, we see that the predictions improve significantly for all halo masses when the environmental information is included in modeling of the halo H I mass.
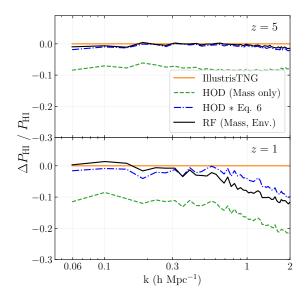
Fig. 6. Same as bottom panels of Fig. 5 but for a combined set of halos of all masses. Note that modeling the effect of the halo environment using either random forests (RF) or symbolic regression significantly improves the accuracy of the predicted HI field at both high and low redshifts; we observe a $\sim 10\%$ improvement in both cases. See Fig. S1 for a comparison of summary statistics other than P_{HI}.

dom forest regressor (RF). We first split the TNG300-1 box into training and test sets: the full box has a side-length $205\,h^{-1}\,\mathrm{Mpc}$, from which we use a box with a side-length $150\,h^{-1}\,\mathrm{Mpc}$ (comprising $\sim 40\%$ of the total volume) for testing the RF and the rest for training the RF. Using the halos in the training volume, we train the RF to predict the halo HI mass as a function of various different halo environmental parameters, and then use it to predict the HI mass of halos in the test set.

We show the power spectrum of the generated HI field from the RF in Fig. 5. For the top panels, we train the RF separately with three environmental parameters $\{\delta_{0.5}, \alpha_{0.5}, \delta_5\}$. As discussed above, we see that including the environmental information causes the HI clustering to increase in general (as halos in denser or more anisotropic environments typically have more HI than average). The only exception to this trend is for $M_h \in [10^{9.5} - 10^{10.5}]\,h^{-1}\,M_\odot$ at $z = 1$. To understand this, we can look at this case in the lower panels of Fig. 3 (blue line): as the environment becomes very dense, the HI mass in the halos significantly decreases and therefore, on average, halos in less dense environments have more HI. We also see that a lot of the information from the environment at small-scales $(\delta_{0.5}, \alpha_{0.5})$ is degenerate with the information from large scales (δ_5). We again see that for $z = 5$, $\delta_{0.5}$ and $\alpha_{0.5}$ give complementary trends, similar to our discussion on Eq. 6b earlier. Another trend in Fig. 5 worth noting is that including the variable δ_5 does not improve clustering at small-scales $(k \gtrsim 1\,h\,\mathrm{Mpc}^{-1})$, which is expected since halos at smaller separations are in the same large-scale environment. To compare with the environmental trends, we also show a case similar to the HOD, where the RF is only trained with the halo mass (M_h) (the RF is just predicting the mean of the scatter in Fig. 2). We do not show error-bars arising from cosmic variance because all the cases are evolved from the same initial conditions. We have only shown the effect

of environmental variables in Fig. 5 and show the effect of internal properties of the halo on $P_{\mathrm{HI}}(k)$ in Fig. S4.

Instead of training the RF with a single environmental parameter, we now use all the environmental overdensity and anisotropy parameters at the scales $\{0.5,1,2,5,10,20\}$ $h^{-1}\,\mathrm{Mpc}$, and show the corresponding results in the bottom panel of Fig. 5. We also show results obtained by using Eq. 6 and we see that the relative improvement is smaller compared to RF. It is important to note that this is because the RF was trained separately for each of the four mass-bins in the figure, while the symbolic regressor was trained on a combined sample that included all halo masses—if we train the symbolic regressor for each individual mass bin, we expect to see better results. Finally, we make a combined sample comprising of halos from all mass bins and show the corresponding results in Fig. 6. Indeed, as the symbolic regressor was trained for the combined sample, it shows better results as compared to Fig. 5, and the improvement due to Eq. 6 and RF is comparable.

6. Discussion and Conclusions

Upcoming galaxy and line intensity mapping surveys will map large volumes of the Universe. We need accurate large-scale mock baryonic catalogues to provide the theory predictions needed to maximize the scientific return of these missions. Halo model tools like HOD are widely used for making large-scale baryonic maps and typically make the assumption that the baryonic content of a halo is a function of only the halo mass.

Using the IllustrisTNG simulation, we show that the neutral hydrogen (HI) content of a halo is dependent on secondary properties other than halo mass, like the environment of the halo, the mass fraction of substructure inside the halo and its concentration (see Figs. 3 and 4). We show that these secondary dependences also affect the overall clustering of HI and lead to HI assembly bias. We also provide physical explanations for the dependences, and model the effect of the halo environment on its HI mass using machine learning tools like random forests and symbolic regression. Our modeling can be easily used to augment a mass-based HOD model of HI, and leads to a significant improvement in the clustering of the modeled HI field (the real-space 21-cm power spectrum prediction is improved by $\gtrsim 10\%$ on scales $k \gtrsim 0.05\,h\,\mathrm{Mpc}^{-1}$, see Fig. 6). Modeling the assembly bias effects using parameters related to the halo environment has the additional advantage that these parameters can be easily computed in DM-only simulations, without there being a need to construct halo merger trees or resolve sub-halo structure.

In order to appropriately marginalize over assembly bias in a Bayesian analysis of survey data, it is crucial to encode its effects in a compact analytic expression. Symbolic regression enables such an encoding (see Eq. 6) and is therefore more advantageous to use over machine learning techniques like random forests or neural networks. Furthermore, the results from symbolic regression can provide an understanding of the underlying physical behavior and are readily generalizable. We expect symbolic regression to be an ideal tool for parameterizing assembly bias for any general case of baryonic tracers, directly by using data from hydrodynamic simulations or semi-analytic models.

Comparison with other works on modeling the halo environment. Although our study focuses on H I, it is worth comparing our approach of modeling the environment to that of other studies which modify the HOD formalism for galaxies based on the halo environment. There has been an ample interest on including the effect of the environment of the halo into the standard HOD model for populating galaxies (16, 20, 46–49, 96) in order to create more accurate galaxy mock catalogs. The previous studies however have only used a single parameter for modeling the halo environment, and they rely on the trends being monotonic with respect to that chosen parameter. As seen in some cases in Fig. 3, the trends can have turnovers and be non-monotonic for parameters like the environmental overdensity. In such cases studying the effect of a single parameter could give misleading results (for e.g., the increasing and the decreasing part of the trends can cancel out giving a null result overall). Furthermore, it is optimal at times to use a combination of two parameters to model the assembly bias effect (as seen in Eq. 6a where $\delta_{0.5}$ and $\alpha_{0.5}$ are both used). Symbolic regression is therefore an alternative approach to infer an optimal and physically motivated parameterization of assembly bias directly from simulations.

Future work. There are multiple ways in which our work can be extended. We have used the results from the IllustrisTNG simulation which uses a particular prescription for baryonic feedback. We plan to see how our results—in particular, the structure of Eq. 6—depend on astrophysical feedback parameters using the CAMELS suite of simulations (72) (which contain 2184 hydrodynamic simulations run for different astrophysical feedback and cosmology parameters). Such a test will also be useful in studying how the measurements of clustering of H I from upcoming surveys can be used to constrain feedback prescriptions in hydrodynamic simulations. We have performed our analysis for two particular redshifts ($z = \{1,5\}$) separately, and we plan to analyze intermediate redshifts and derive a single equation similar to the ones in Eq. 6, but which also includes a redshift dependence. We have studied the assembly bias effects in real space and it will be interesting to extend our analysis to redshift space and probe potential anisotropic assembly bias effects because of the correlation between the local tidal field and the H I mass of the halo. We will also extend our techniques from H I to galaxies (model the number of galaxies in a halo instead of the halo H I mass) in an upcoming paper.

[‡‡]https://github.com/franciscovillaescusa/Pylians3

Supplemental material for 'Modeling the neutral hydrogen assembly bias with machine learning and symbolic regression'
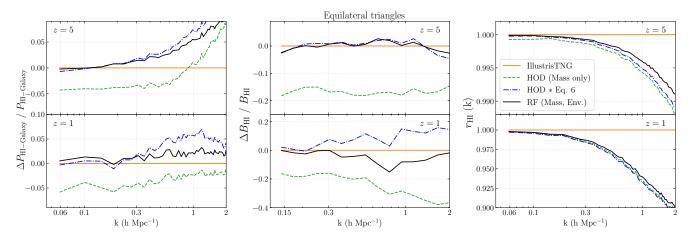

Fig. S1. Same as Fig. 6, but for different summary statistics of the modeled HI field. **Left:** Cross power between HI and galaxies, where we use galaxies in the TNG300-1 box with $M_* > 10^{10} M_\odot$. Note that the deviation in the cross power at high-k at $z = 5$ is due to low number density of galaxies (we have checked that increasing the number density improves our results at high-k for both redshifts). **Center:** Bispectrum as a function of side length for the equilateral triangle configuration. **Right:** The cross-correlation coefficient of the modeled HI field (defined as $r_{(A)} = P_{A-\mathrm{Illustris}}/\sqrt{P_A P_{\mathrm{Illustris}}}$). Overall, including the halo environment provides significant improvement for $P_{\mathrm{HI-Galaxy}}$ and B_{HI} but a modest change in r_{HI}. This hints at the fact that the halo environment primarily affects the bias of the modeled HI field.

A. Other summary statistics of the modeled HI field

In Sec. 5, we showed the comparison between the auto-power spectrum of the HI field modeled via HOD, random forests (RF) and symbolic regression. In this section, we discuss other useful summary statistics of the modeled HI field and show the corresponding results in Fig. S1.

1) Cross-power between HI and galaxy fields. $P_{\mathrm{HI-Galaxy}}$ is an important statistic to study because large regions of future HI surveys will overlap with those of galaxy surveys like DESI or the Roman space telescope. Furthermore, unlike the auto-power spectrum of 21cm which is yet to be detected, there have been multiple detections of the $P_{\mathrm{HI-Galaxy}}$ signal at $z \sim 1$ (97, 98). To calculate $P_{\mathrm{HI-Galaxy}}$, we use galaxies in the TNG300-1 sample with stellar masses $M_* > 1 \times 10^{10}$ $M_\odot$ (which corresponds to a number density of $n = 2 \times 10^{-4}$ h^3/Mpc^3 at $z = 5$ and $n = 9 \times 10^{-3}$ h^3/Mpc^3 at $z = 1$). We see that, similarly to $P_{\mathrm{HI}}(k)$ in Fig. 6, modeling the halo environment improves the $P_{\mathrm{HI-Galaxy}}$ prediction in the left panel of Fig. S1. It is worth noting that there is a deviation for all modeling techniques at high-k at $z = 5$. We have checked that increasing the number density of the galaxy sample removes this high-k deviation. This suggests that the deviation is a result of the large shot noise present in the $z = 5$ galaxy sample. We emphasize that we have not included the one halo term in our analysis and including it will likely improve the predictions for $k \gtrsim 1 \, h \, \mathrm{Mpc}^{-1}$ for all summary statistics.

2) Bispectrum. Late-time gravitational clustering causes a significant leakage of cosmological information, that initial was in the power spectrum, into higher order statistics (99–103). To recover this information, the lowest order statistic that one needs to compute in Fourier space is the bispectrum. Unlike the power spectrum, the bispectrum is sensitive to the shape of structures generated by late-time gravitational instability, and therefore provides complementary information. We show results for the bispectrum in the center panel of Fig. S1. The HOD model again shows a deviation, similar to the power spectrum case, at low-k and including the environmental effects improve the prediction. We only show results for the equilateral triangle configuration but have checked that the improvement is similar for other triangle configurations.

3) HI cross-correlation coefficient. The auto-power spectrum P_{HI} measures the amplitude of HI fluctuations (averaging over the squares of the mode amplitudes), being thus insensitive to fluctuations phases. We therefore calculate the cross-correlation coefficient of the HI field (which is sensitive to the phases of HI fluctuations), and show the results in the right panel of Fig. S1. We see that including the environmental effects provides marginal improvement at $z = 5$ and has a negligible effect at $z = 1$. Unlike the other statistics we discussed, the cross-correlation coefficient is not sensitive to the bias of HI. We therefore conclude that the modeling the environment of the halo improves the prediction of the HI bias, but has a small effect on the phases of the HI fluctuations.

B. Effect of internal halo properties on its HI mass

In Sec. 3, we discussed the effect of the environment properties of the halo on its HI mass. Here we discuss the effect of two internal properties: halo concentration and subhalo mass fraction. Apart from these two, there could also be correlations between M_{HI} and other internal halo properties like its formation epoch, spin, velocity dispersion and others, but we leave their study to a future work.

Mass fraction in subhalos. As shown earlier in Fig. 4, the fractional mass in subhalos is related to the total HI inside a halo. This is because the gas in subhalos is more dense and can more efficiently self-shield itself as compared to gas in the CGM (circumgalactic medium) of the halo. The subhalo mass fraction is correlated with the environment of the halo (increasing the environmental overdensity typically results in more mergers and therefore more substructure in halos). We show the magnitude of this correlation in Fig. S6; we have used an intermediate length scale (2.5 h^{-1} Mpc) to showcase the trends (we also checked that we obtain similar results using some other environmental parameters, like $\alpha_{0.5}$). Note that we use the subhalo data provided by IllustrisTNG in our analysis (104).

In order to show that the subhalo mass fraction also affects the clustering of HI, we train a random forest regressor using it and show the results in Fig. S4. We find improvements in the prediction of P_{HI} in most of the cases. There is however one particular case which is an exception: the low mass halos in $z = 1$, where ram pressure stripping has the largest effect. Ram pressure stripping exclusively affects the gas distribution and not the DM distribution in the halo. Information of the subhalo mass fraction does not therefore capture this effect, leading to slightly biased predictions.
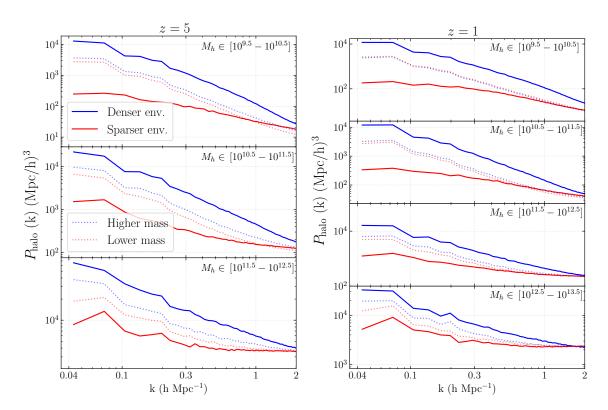
Fig. S2. Difference in the clustering of halos in denser and sparser environments for different halo masses (M_h). We split the halos in each mass bin into two sets based on their environmental overdensity: those with $\delta_{2.5}$ values above (Denser env.) or below (Sparser env.) the median; we show the halo power spectrum for the two samples in solid lines. The halos in denser environments cluster a lot more strongly. Even if there is slightly more Hɪ than average in halos in denser environments (which amounts to upweighting their contribution to the power spectrum of Hɪ), the overall Hɪ clustering will increase. To confirm that the differences observed in halo clustering are not just due to variation in halo mass within our mass bins, we show in dotted lines the results from splits based purely on the halo mass. All masses are in units of $h^{-1} M_\odot$.

Halo concentration. We use the following formula as a proxy to estimate halo concentration, c: $c = R_{200c}/R_{\max}$, where $R_{\max}$ is the comoving radius at the point where the maximum circular velocity is attained for the largest subhalo inside the halo. Note that this definition is less accurate for large mass halos; a more accurate way of finding the halo concentration is to fit a NFW profile to the halo mass distribution and measure the corresponding scale radius r_s (the concentration can then be calculated as $c = R_{200c}/r_s$). We have used the former manner to estimate concentration as the public IllustrisTNG data only provides information about $R_{\max}$.

One might expect the concentration of a halo to affect its Hɪ mass, as a deeper gravitational potential well can lead to larger amounts of gas being accumulated at the center (which helps in shielding from the UV radiation). This is however only true for the low-mass halos, as seen in the right panel of Fig. 4. The situation becomes complicated for larger halos because higher concentrations lead to increased star formation and larger ionizing feedback due to supernovae and AGN. We indeed see this effect in the right panel of Fig. 4 where, beyond a certain threshold, larger concentration leads to lower M_{HI}. Upon training the RF with the halo concentration, we find very little improvement in the Hɪ clustering prediction.

C. Calibrating the HOD model for halos from N-body simulations

In Sec. 2, we discussed a mass based HOD model for Hɪ. In order to calibrate the model parameters $\{M_0, \alpha, M_{\min}\}$ in Eq 2., we had used the data for the total mass M_h and the Hɪ mass M_{HI} of halos from the hydrodynamic IllustrisTNG simulation. However, if this calibrated model is directly used on halos in an N-body simulation, one would get spurious results. This is because there is a significant change in the halo mass when baryons and their feedback are included in the simulation, as is seen in multiple hydrodynamic simulations (31, 86). To quantify this change, we take advantage of

the fact that TNG provides both the hydrodynamical (or full-physics, FP) simulation output as well as the N-body (or dark matter only, DMO) one, evolved from the same set of initial conditions. TNG also provides the matches for subhalos in the FP and DMO simulation (based upon their origin from the same Lagrangian patch of initial conditions). This helps us to match the corresponding halos in the FP and DMO simulations (following (36), we match two halos if their central subhaloes are matched).

We show the ratio of masses of the matched halos in the left panel of Fig. S3. We see that the halo masses are different by $\gtrsim 10\%$ in the two simulations at both high and low redshifts. This difference also affects the HOD modeling of the Hɪ field. Our model was calibrated for the halo masses from the FP simulation (M_h^{FP}). Using this model on halos from the DMO version of the TNG300-1 box, we find that P_{HI} of the modeled field is discrepant, as seen in the purple curves in the center and right panels of Fig. S3. We also find that Ω_{HI} is overpredicted by $\sim 15\%$. This discrepancy was also encountered by VN18 who did a similar analysis for the TNG100-1 box (see their Fig. 24).

In order to appropriately recalibrate the HOD model such that it can be used on a N-body simulation, we refit the formula in Eq. 2 using M_{HI} from the FP simulation and M_h^{DMO} from the DMO simulation. We obtain the model parameters $\{\alpha, M_0/(10^{10}\,h^{-1}\,M_\odot), M_{\min}/(10^{10}\,h^{-1}\,M_\odot)\}$ to be $\{0.531, 0.589, 26.285\}$ for $z = 1$ and $\{0.727, 0.395, 5.216\}$ for $z = 5$. Using the recalibrated HOD model on the halos from the DMO simulation, we find that the results for P_{HI} are now improved, as seen from the red curves in Fig. S3.

D. Machine learning techniques used

In the introduction section of the main text, we outlined the motivation of using machine learning models like random forests or symbolic regression instead of deep neural networks. In this appendix we provide more details on these models.
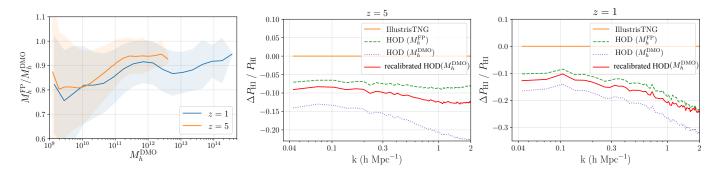
Fig. S3. This figure illustrates that an HOD relation, which is calibrated using halo masses from a hydrodynamic simulation, will give discrepant results if directly used on halos in a N-body simulation. We first match the halos in the hydrodynamic (full physics, FP) simulation to their counterparts in the N-body (DM-only, DMO) simulation. **Left:** The ratio of the halo masses in the two simulations (the mean is in solid and the shaded area encloses 68 % of the data). There is a significant difference in halo masses in the two simulations due to the effects of baryonic feedback. **Center:** The relative difference in P_{HI} for the HI field modeled using the HOD relation on halos from the FP (DMO) simulation is shown in green (purple). Note that the HOD relation was calibrated on data from the FP simulation (see Fig. 2), and therefore underperforms when used on the halos from the DMO simulation. We recalibrate the HOD model using halo masses from the DMO simulation and show the results in red, where the performance of the HOD model is now improved. **Right:** Same as the center panel but for $z = 1$.

Symbolic regression. Symbolic regression (SR) is a tool that searches the space of mathematical expressions to find the best equation that fits the data. The difference between it and ordinary "least squares" regression is that knowledge of the underlying functional form of the fitting function is not required a priori. Let us briefly describe the procedure to fit a function (e.g. Eq. 1) with the PYSR package. First we specify the relevant input parameters (e.g., $\delta_{0.5}, \alpha_{0.5}$ or δ_5) and the operations (e.g., sum $(+)$, multiplication $(\cdot)$, exponential or sinus). Using genetic programming (105), the SR then generates multiple iterations of formulae, like $2.7 \cdot \delta_{0.5} + \alpha_{0.5} \cdot \delta_5$ for example, and outputs a final list of equations which have the lowest mean squared error when compared to the data. The equations in the final list are ranked on the basis of their complexity (more complex operations like exponentials are penalized over standard ones like $+$).

Because it is easy to overfit given the large scatter in the simulation (see Fig. 3) data, we restrict ourselves to the simplest case of sum and multiplication operators. Furthermore, in the final list of formulae obtained from PYSR, we choose the most simplest ones to present in Eqs. 6. We also show the performance of the expressions in Fig. S5. Note also that dimensionality of the input dataset needs to be relatively small in order for the use of SR to be feasible as the cost scales exponentially with the number of input parameters and operators. We therefore needed to reduce assembly bias dependence to a few most important parameters before using SR.

Random forests. We use the random forest algorithm from the publicly available package `Scikit-Learn` (106). A random forest regressor (RF) is a collection of decision trees; each tree is in itself a regression model and is trained on a different random subset of the training data (107). The output from a RF is the mean of the predictions from the individual trees. Note that a single decision tree is prone to overfitting and using the ensemble mean of the different trees helps to reduce overfitting. RFs have been used for applications to cosmological problems (108–110), and allow for an easy measurement of the relative importance of each feature. This makes them better suited for interpretation as compared to deep neural networks.

We train the RFs to model the ratio of the HI mass of the halo to the prediction from the mass-only HOD model from Eq. 2 ($M_{\mathrm{HI}}/M_{\mathrm{HOD\text{-}HI}}$), as a function of different secondary properties of the halo. We use 100 decision trees in our model. For the hyperparameters which control the complexity of the model (and avoid overfitting), we use: `max_depth`=100 and `min_samples_leaf`=10.
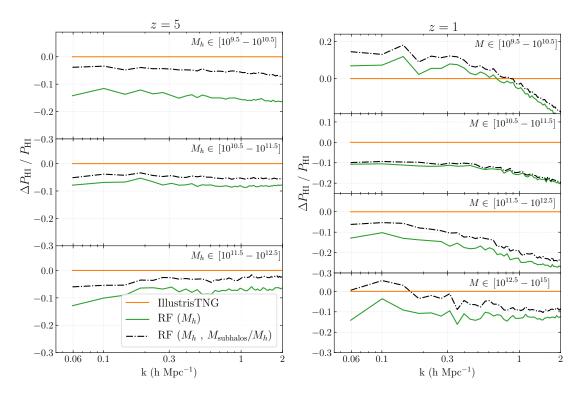
Fig. S4. Same as the top panel of Fig. 5 but training the random forest (RF) regressors using the subhalo mass fraction of the halos ($M_{\mathrm{subhalos}}/M_h$). In most cases, we find improvements similar to those seen for environmental parameters in Fig. 5. Note however that the predictions become worse for the case of low-mass halos in $z = 1$; see the text for an explanation. We found that training the RF using the halo concentration had a small effect on the P_{HI} predictions, and hence we do not show results for that case.
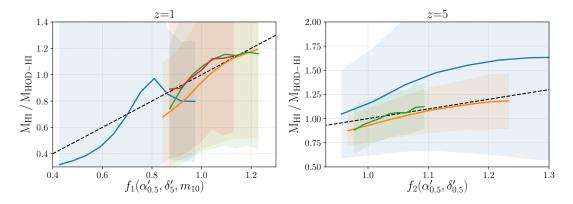

Fig. S5. We train a symbolic regressor to predict the ratio of the H I mass of a halo to the output $M_{\mathrm{HOD\text{-}HI}}$ from the mass-based HOD model in Eq. 2. We compare the values predicted from Eq. 6 against the true values from IllustrisTNG (the black dashed line represents predicted=true). f_1 and f_2 are the expressions in the RHS of Eq. 6b and Eq. 6a respectively. The color coding of the curves is the same as the figure below. Note that the offset seen in the blue line in the right panel is because the HOD model in Eq. 2 does not fit the data well in the low M_h regime for $z = 5$. Note that the ratio $M_{\mathrm{HI}}/M_{\mathrm{HOD\text{-}HI}}$ corresponds to the level of assembly bias seen in the hydrodynamic simulation and the symbolic regressor is thus efficient at finding parameter combinations which can be used for model assembly bias.

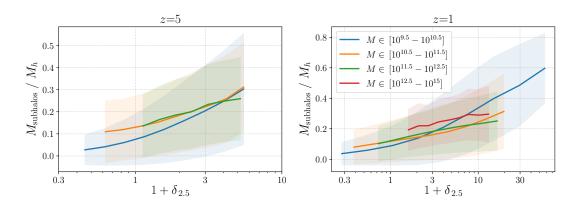**Fig. S6.** Effect of the environmental overdensity of the halo, parameterized by $\delta_{2.5}$, on its subhalo mass fraction. Increasing the environmental overdensity typically results in more mergers and therefore more substructure in halos. These trends are shown to complement the physical explanations outlined in Sec. 3A for the effect of the environment seen on the halo H I mass in Fig. 3.

References

1. Pillepich A, et al. (2018) Simulating galaxy formation with the IllustrisTNG model. MNRAS 473(3):4077–4106.
2. Scoccimarro R, Sheth RK, Hui L, Jain B (2001) How Many Galaxies Fit in a Halo? Constraints on Galaxy Formation Efficiency from Spatial Clustering. ApJ 546(1):20–34.
3. Seljak U (2000) Analytic model for galaxy and dark matter clustering. MNRAS 318(1):203–213.
4. Peacock JA, Smith RE (2000) Halo occupation numbers and galaxy bias. MNRAS 318(4):1144–1156.
5. Berlind AA, Weinberg DH (2002) The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass. ApJ 575(2):587–616.
6. Zheng Z, et al. (2005) Theoretical Models of the Halo Occupation Distribution: Separating Central and Satellite Galaxies. ApJ 633(2):791–809.
7. Reid BA, Seo HJ, Leauthaud A, Tinker JL, White M (2014) A 2.5 per cent measurement of the growth rate from small-scale redshift space clustering of SDSS-III CMASS galaxies. MNRAS 444(1):476–502.
8. Rodríguez-Torres SA, , et al. (2016) The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: modelling the clustering and halo occupation distribution of BOSS CMASS galaxies in the Final Data Release. Mon. Not. Roy. Astron. Soc. 460(2):1173–1187.
9. Saito S, et al. (2016) Connecting massive galaxies to dark matter haloes in BOSS - I. Is galaxy colour a stochastic process in high-mass haloes? MNRAS 460(2):1457–1475.
10. Alam S, Miyatake H, More S, Ho S, Mandelbaum R (2017) Testing gravity on large scales by combining weak lensing with galaxy clustering using CFHTLenS and BOSS CMASS. MNRAS 465(4):4853–4865.
11. Hearin AP, Zentner AR, van den Bosch FC, Campbell D, Tollerud E (2016) Introducing decorated HODs: modelling assembly bias in the galaxy-halo connection. MNRAS 460(3):2552–2570.
12. Avila S, , et al. (2020) The Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: exploring the Halo Occupation Distribution model of Emission Line Galaxies.
13. DeRose J, et al. (2019) The AEMULUS Project. I. Numerical Simulations for Precision Cosmology. ApJ 875(1):69.
14. Zhai Z, et al. (2019) The Aemulus Project. III. Emulation of the Galaxy Correlation Function. ApJ 874(1):95.
15. Lange JU, et al. (2019) Cosmological Evidence Modelling: a new simulation-based approach to constrain cosmology on non-linear scales. MNRAS 490(2):1870–1878.
16. Yuan S, Hadzhiyska B, Bose S, Eisenstein DJ, Guo H (2020) Evidence for galaxy assembly bias in BOSS CMASS redshift-space galaxy correlation function. arXiv e-prints p. arXiv:2010.04182.
17. Yuan S, Eisenstein DJ, Leauthaud A (2020) Can assembly bias explain the lensing amplitude of the BOSS CMASS sample in a Planck cosmology? MNRAS 493(4):5551–5564.
18. Wibking BD, et al. (2020) Cosmology with galaxy-galaxy lensing on non-perturbative scales: emulation method and application to BOSS LOWZ. MNRAS 492(2):2872–2896.
19. Alam S, Zu Y, Peacock JA, Mand elbaum R (2019) Cosmic web dependence of galaxy clustering and quenching in SDSS. MNRAS 483(4):4501–4517.
20. Salcedo AN, et al. (2020) Elucidating Galaxy Assembly Bias in SDSS. arXiv e-prints p. arXiv:2010.04176.
21. Villaescusa-Navarro F, Viel M, Datta KK, Choudhury TR (2014) Modeling the neutral hydrogen distribution in the post-reionization Universe: intensity mapping. J. Cosmology Astropart. Phys. 2014(9):050.
22. Villaescusa-Navarro F, et al. (2015) Cross-correlating 21cm intensity maps with Lyman Break Galaxies in the post-reionization era. J. Cosmology Astropart. Phys. 2015(3):034.
23. Castorina E, Villaescusa-Navarro F (2017) On the spatial distribution of neutral hydrogen in the Universe: bias and shot-noise of the H I power spectrum. MNRAS 471(2):1788–1796.
24. Villaescusa-Navarro F, et al. (2018) Ingredients for 21 cm Intensity Mapping. ApJ 866(2):135.
25. Modi C, Castorina E, Feng Y, White M (2019) Intensity mapping with neutral hydrogen and the Hidden Valley simulations. J. Cosmology Astropart. Phys. 2019(9):024.
26. Spinelli M, Zoldan A, De Lucia G, Xie L, Viel M (2020) The atomic hydrogen content of the post-reionization Universe. MNRAS 493(4):5434–5455.
27. Vale A, Ostriker JP (2004) Linking halo mass to galaxy luminosity. MNRAS 353(1):189–200.
28. Conroy C, Wechsler RH, Kravtsov AV (2006) Modeling Luminosity-dependent Galaxy Clustering through Cosmic Time. ApJ 647(1):201–214.
29. Zhu G, et al. (2006) The Dependence of the Occupation of Galaxies on the Halo Formation Time. ApJ 639(1):L5–L8.
30. Pujol A, Gaztañaga E (2014) Are the halo occupation predictions consistent with large-scale galaxy clustering? MNRAS 442(3):1930–1941.
31. Schaller M, et al. (2015) Baryon effects on the internal structure of ΛCDM haloes in the EAGLE simulations. Mon. Not. Roy. Astron. Soc. 451(2):1247–1267.
32. Croton DJ, Gao L, White SDM (2007) Halo assembly bias and its effects on galaxy clustering. MNRAS 374(4):1303–1309.
33. Vakili M, Hahn C (2019) How Are Galaxies Assigned to Halos? Searching for Assembly Bias in the SDSS Galaxy Clustering. ApJ 872(1):115.
34. Kobayashi Y, Nishimichi T, Takada M, Takahashi R (2020) Cosmological information content in redshift-space power spectrum of SDSS-like galaxies in the quasinonlinear regime up to k =0.3 h Mpc^{-1}. Phys. Rev. D 101(2):023510.
35. Wechsler RH, Tinker JL (2018) The Connection Between Galaxies and Their Dark Matter Halos. ARA&A 56:435–487.
36. Hadzhiyska B, Bose S, Eisenstein D, Hernquist L, Spergel DN (2020) Limitations to the 'basic' HOD model and beyond. MNRAS 493(4):5506–5519.
37. Wechsler RH, Zentner AR, Bullock JS, Kravtsov AV, Allgood B (2006) The Dependence

38. of Halo Clustering on Halo Formation History, Concentration, and Occupation. ApJ 652(1):71–84.
38. Dalal N, Doré O, Huterer D, Shirokov A (2008) Imprints of primordial non-gaussianities on large-scale structure: Scale-dependent bias and abundance of virialized objects. Phys. Rev. D 77(12):123514.
39. Paranjape A, Hahn O, Sheth RK (2018) Halo assembly bias and the tidal anisotropy of the local halo environment. MNRAS 476(3):3631–3647.
40. Han J, et al. (2019) The multidimensional dependence of halo bias in the eye of a machine: a tale of halo structure, assembly, and environment. MNRAS 482(2):1900–1919.
41. Sheth RK, Tormen G (2004) On the environmental dependence of halo formation. MNRAS 350(4):1385–1390.
42. Gao L, Springel V, White SDM (2005) The age dependence of halo clustering. MNRAS 363(1):L66–L70.
43. Leauthaud A, , et al. (2017) Lensing is Low: Cosmology, Galaxy Formation, or New Physics? Mon. Not. Roy. Astron. Soc. 467(3):3024–3047.
44. Amodeo S, , et al. (2020) The Atacama Cosmology Telescope: Modelling the Gas Thermodynamics in BOSS CMASS galaxies from Kinematic and Thermal Sunyaev-Zel'dovich Measurements.
45. Paranjape A, Kovač K, Hartley WG, Pahwa I (2015) Correlating galaxy colour and halo concentration: a tunable halo model of galactic conformity. MNRAS 454(3):3030–3048.
46. McEwen JE, Weinberg DH (2018) The effects of assembly bias on the inference of matter clustering from galaxy-galaxy lensing and galaxy clustering. MNRAS 477(4):4348–4361.
47. Xu X, Zehavi I, Contreras S (2020) Dissecting and Modelling Galaxy Assembly Bias. arXiv e-prints p. arXiv:2007.05545.
48. Salcedo AN, et al. (2020) Cosmology with stacked cluster weak lensing and cluster-galaxy cross-correlations. MNRAS 491(3):3061–3081.
49. Salcedo AN, et al. (2020) Cosmology with stacked cluster weak lensing and cluster-galaxy cross-correlations. MNRAS 491(3):3061–3081.
50. Hadzhiyska B, Bose S, Eisenstein D, Hernquist L (2020) Extensions to models of the galaxy-halo connection. arXiv e-prints p. arXiv:2008.04913.
51. Wadekar D, Villaescusa-Navarro F, Ho S, Perreault-Levasseur L (2020) HInet: Generating neutral hydrogen from dark matter with neural networks. arXiv e-prints p. arXiv:2007.10340.
52. Zhang X, et al. (2019) From Dark Matter to Galaxies with Convolutional Networks. arXiv e-prints p. arXiv:1902.05965.
53. Giusarma E, et al. (2019) Learning neutrino effects in Cosmology with Convolutional Neural Networks. arXiv e-prints p. arXiv:1910.04255.
54. Yip JHT, et al. (2019) From Dark Matter to Galaxies with Convolutional Neural Networks. arXiv e-prints p. arXiv:1910.07813.
55. Zamudio-Fernandez J, et al. (2019) HIGAN: Cosmic Neutral Hydrogen with Generative Adversarial Networks. arXiv e-prints p. arXiv:1904.12846.
56. He S, et al. (2019) Learning to predict the cosmological structure formation. Proceedings of the National Academy of Science 116(28):13825–13832.
57. Modi C, Feng Y, Seljak U (2018) Cosmological reconstruction from galaxy light: neural network based light-matter connection. J. Cosmology Astropart. Phys. 2018(10):028.
58. Kodi Ramanah D, Charnock T, Villaescusa-Navarro F, Wandelt BD (2020) Super-resolution emulator of cosmological simulations using deep physical models. MNRAS 495(4):4227–4236.
59. Tröster T, Ferguson C, Harnois-Déraps J, McCarthy IG (2019) Painting with baryons: augmenting N-body simulations with gas using deep generative models. MNRAS 487(1):L24–L29.
60. Thiele L, Villaescusa-Navarro F, Spergel DN, Nelson D, Pillepich A (2020) Teaching neural networks to generate Fast Sunyaev Zel'dovich Maps. arXiv e-prints p. arXiv:2007.07267.
61. Li Y, et al. (2020) AI-assisted super-resolution cosmological simulations. arXiv e-prints p. arXiv:2010.06608.
62. Berger P, Stein G (2019) A volumetric deep Convolutional Neural Network for simulation of mock dark matter halo catalogues. MNRAS 482(3):2861–2871.
63. Zorrilla Matilla JM, Sharma M, Hsu D, Haiman Z (2020) Interpreting deep learning models for weak lensing. arXiv e-prints p. arXiv:2007.06529.
64. Alber M, et al. (2018) iNNvestigate neural networks! arXiv e-prints p. arXiv:1808.04260.
65. Gunn JE, Gott, J. Richard I (1972) On the Infall of Matter Into Clusters of Galaxies and Some Effects on Their Evolution. ApJ 176:1.
66. Schmidt MD, Lipson H (2009) Distilling free-form natural laws from experimental data. Science 324:81 − 85.
67. Udrescu SM, Tegmark M (2020) AI Feynman: A physics-inspired method for symbolic regression. Science Advances 6(16):eaay2631.
68. Wu T, Tegmark M (2018) Toward an AI Physicist for Unsupervised Learning. arXiv e-prints p. arXiv:1810.10525.
69. Cranmer M, et al. (2020) Discovering Symbolic Models from Deep Learning with Inductive Biases. arXiv e-prints p. arXiv:2006.11287.
70. Cranmer MD, Xu R, Battaglia P, Ho S (2019) Learning Symbolic Physics with Graph Networks. arXiv e-prints p. arXiv:1909.05862.
71. La Torre V, Villaescusa-Navarro F (2020). in preparation.
72. Villaescusa-Navarro F, , et al. (2020) The CAMELS project: Cosmology and Astrophysics with MachinE Learning Simulations. arXiv e-prints p. arXiv:2010.00619.
73. Kim S, et al. (2019) Integration of Neural Network-Based Symbolic Regression in Deep Learning for Scientific Discovery. arXiv e-prints p. arXiv:1912.04825.
74. Liu Z, Tegmark M (2020) AI Poincaré: Machine Learning Conservation Laws from Trajectories. arXiv e-prints p. arXiv:2011.04698.
75. Cranmer M (2020) Pysr: Fast & parallelized symbolic regression in python/julia.
76. Montero-Dorta AD, et al. (2020) The manifestation of secondary bias on the galaxy population from IllustrisTNG300. MNRAS 496(2):1182–1196.
77. Contreras S, Angulo R, Zennaro M (2020) A flexible modelling of galaxy assembly bias. arXiv e-prints p. arXiv:2005.03672.
78. Shi J, et al. (2020) Power Spectrum of Intrinsic Alignments of Galaxies in IllustrisTNG. arXiv

e-prints p. arXiv:2009.00276.

79. Bose S, et al. (2019) Revealing the galaxy-halo connection in IllustrisTNG. MNRAS 490(4):5693–5711.

80. Padmanabhan H, Choudhury TR, Refregier A (2016) Modelling the cosmic neutral hydrogen from DLAs and 21-cm observations. MNRAS 458(1):781–788.

81. Padmanabhan H, Refregier A (2017) Constraining a halo model for cosmological neutral hydrogen. MNRAS 464(4):4008–4017.

82. Padmanabhan H, Refregier A, Amara A (2017) A halo model for cosmological neutral hydrogen : abundances and clustering. MNRAS 469(2):2323–2334.

83. Padmanabhan H, Refregier A, Amara A (2019) Impact of astrophysics on cosmology forecasts for 21 cm surveys. MNRAS 485(3):4060–4070.

84. Padmanabhan H, Refregier A, Amara A (2020) Cross-correlating 21 cm and galaxy surveys: implications for cosmology and astrophysics. MNRAS 495(4):3935–3942.

85. Barnes LA, Haehnelt MG (2014) The bias of DLAs at z∼2.3: evidence for very strong stellar feedback in shallow potential wells. MNRAS 440(3):2313–2321.

86. Lovell MR, et al. (2018) The fraction of dark matter within galaxies from the IllustrisTNG simulations. MNRAS 481(2):1950–1975.

87. Ramakrishnan S, Paranjape A, Hahn O, Sheth RK (2019) Cosmic web anisotropy is the primary indicator of halo assembly bias. MNRAS 489(3):2977–2996.

88. Obuljen A, Dalal N, Percival WJ (2019) Anisotropic halo assembly bias and redshift-space distortions. J. Cosmology Astropart. Phys. 2019(10):020.

89. Hahn O, Porciani C, Dekel A, Carollo CM (2009) Tidal effects and the environment dependence of halo assembly. MNRAS 398(4):1742–1756.

90. Mansfield P, Kravtsov AV (2020) The three causes of low-mass assembly bias. MNRAS 493(4):4763–4782.

91. Chan KC, Scoccimarro R, Sheth RK (2012) Gravity and large-scale nonlocal bias. Phys. Rev. D 85(8):083509.

92. Baldauf T, Seljak U, Desjacques V, McDonald P (2012) Evidence for quadratic tidal tensor bias from the halo bispectrum. Phys. Rev. D 86(8):083540.

93. Catelan P, Theuns T (1996) Evolution of the angular momentum of protogalaxies from tidal torques: Zel'dovich approximation. MNRAS 282(2):436–454.

94. Heavens A, Peacock J (1988) Tidal torques and local density maxima. MNRAS 232:339–360.

95. Bond JR, Cole S, Efstathiou G, Kaiser N (1991) Excursion Set Mass Functions for Hierarchical Gaussian Fluctuations. ApJ 379:440.

96. Hadzhiyska B, Tacchella S, Bose S, Eisenstein DJ (2020) The galaxy-halo connection of emission-line galaxies in IllustrisTNG. *arXiv e-prints* p. arXiv:2011.05331.

97. Chang TC, Pen UL, Bandura K, Peterson JB (2010) Hydrogen 21-cm Intensity Mapping at redshift 0.8. *arXiv e-prints* p. arXiv:1007.3709.

98. Masui KW, et al. (2013) Measurement of 21 cm Brightness Fluctuations at z ∼0.8 in Cross-correlation. ApJ 763(1):L20.

99. Wadekar D, Scoccimarro R (2019) The Galaxy Power Spectrum Multipoles Covariance in Perturbation Theory. *arXiv e-prints* p. arXiv:1910.02914.

100. Scoccimarro R, Zaldarriaga M, Hui L (1999) Power Spectrum Correlations Induced by Nonlinear Clustering. ApJ 527:1–15.

101. Takada M, Jain B (2004) Cosmological parameters from lensing power spectrum and bispectrum tomography. MNRAS 348(3):897–915.

102. Villaescusa-Navarro F, et al. (2019) The Quijote simulations. *arXiv e-prints* p. arXiv:1909.05273.

103. Hahn C, Villaescusa-Navarro F, Castorina E, Scoccimarro R (2020) Constraining M_{ν} with the bispectrum. Part I. Breaking parameter degeneracies. J. Cosmology Astropart. Phys. 2020(3):040.

104. Springel V, White SDM, Tormen G, Kauffmann G (2001) Populating a cluster of galaxies - I. Results at [formmu2]z=0. MNRAS 328(3):726–750.

105. Koza J (1993) Genetic programming - on the programming of computers by means of natural selection in *Complex adaptive systems*.

106. Pedregosa F, et al. (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12(85):2825–2830.

107. Breiman L (2001) Random forests. *Mach. Learn.* 45(1):5–32.

108. Lucie-Smith L, Peiris HV, Pontzen A, Lochner M (2018) Machine learning cosmological structure formation. MNRAS 479(3):3405–3414.

109. Moster BP, Naab T, Lindström M, O'Leary JA (2020) GalaxyNet: Connecting galaxies and dark matter haloes with deep neural networks and reinforcement learning in large volumes. *arXiv e-prints* p. arXiv:2005.12276.

110. Nadler EO, Mao YY, Wechsler RH, Garrison-Kimmel S, Wetzel A (2018) Modeling the Impact of Baryons on Subhalo Populations with Machine Learning. ApJ 859(2):129.