From Dark Matter to Galaxies with Convolutional Networks

Xinyue Zhang*, Yanfang Wang*, Wei Zhang*, Yueqiu Sun**
Center for Data Science, New York University
xz2139,yw1007,wz1218,ys3202@nyu.edu

Gabriella Contardo, Francisco Villaescusa-Navarro
Center for Computational Astrophysics, Flatiron Institute
gcontardo,fvillaescusa@flatironinstitute.org

Siyu He
Department of Physics, Carnegie Mellon University
Center for Computational Astrophysics, Flatiron Institute
she@flatironinstitute.org

Shirley Ho
Center for Computational Astrophysics, Flatiron Institute
Department of Astrophysical Sciences, Princeton University
Department of Physics, Carnegie Mellon University
shirleyho@flatironinstitute.org

ABSTRACT

Cosmological surveys aim at answering fundamental questions about our Universe, including the nature of dark matter or the reason of unexpected accelerated expansion of the Universe. In order to answer these questions, two important ingredients are needed: 1) data from observations and 2) a theoretical model that allows fast comparison between observation and theory. Most of the cosmological surveys observe galaxies, which are very difficult to model theoretically due to the complicated physics involved in their formation and evolution; modeling realistic galaxies over cosmological volumes requires running computationally expensive hydrodynamic simulations that can cost millions of CPU hours. In this paper, we propose to use deep learning to establish a mapping between the 3D galaxy distribution in hydrodynamic simulations and its underlying dark matter distribution. One of the major challenges in this pursuit is the very high sparsity in the predicted galaxy distribution. To this end, we develop a two-phase convolutional neural network architecture to generate fast galaxy catalogues, and compare our results against a standard cosmological technique. We find that our proposed approach either outperforms or is competitive with traditional cosmological techniques. Compared to the common methods used in cosmology, our approach also provides a nice trade-off between time-consumption (comparable to fastest benchmark in the literature) and the quality and accuracy of the predicted simulation. In combination with current and upcoming data from cosmological observations, our method has the potential to answer fundamental questions about our Universe with the highest accuracy. [1]

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Applied computing** → **Astronomy**.

KEYWORDS

Convolutional neural networks, high sparsity, galaxy prediction, hydrodynamic simulation, dark matter

1 INTRODUCTION

Cosmology focuses on studying the origin and evolution of our Universe, from the Big Bang to today and its future. One of the holy grails of cosmology is to understand and define the physical rules and parameters that led to our actual Universe. Astronomers survey large volumes of the Universe [10, 12, 17, 32] and employ a large ensemble of computer simulations to compare with the observed data in order to extract the full information of our own Universe.

The constant improvement of computational power has allowed cosmologists to pursue elucidating the fundamental parameters and laws of the Universe by relying on simulations as their theory predictions. These simulations can help determine if a set of rules or specific parameters can lead to the observed Universe. An important type of simulations is gravo-hydrodynamical simulations, which aim at reproducing the formation and evolution of galaxies through time.

However, evolving trillions of galaxies over billions of light years including the forces of gravity, electromagnetism, and hydrodynamics, is a daunting task. The state-of-art fully gravo-hydrodynamical cosmological simulations that include most of the relevant physics can only simulate a small fraction of our Universe and still requires 19 million CPU hours (or about 2000 years on one single CPU) for the most recent one [31] to complete.

On the other hand, the standard cosmological model provides us with a solution to this challenge: most of the matter in the Universe is made up of dark matter, and the large scale cosmic structure of the Universe can be modeled quite accurately when we evolve dark matter through time with only physics of gravity. When we do add the gas into the mix, gas usually traces the matter density, and for large enough dark matter halos, gas falls to the center of dark matter halos, subsequently cool down and form stars and galaxies. In other words, dark matter halos form the skeleton inside which galaxies form, evolve, and merge. Hence, the behaviors, such as growth, internal properties, and spatial distribution of galaxies, are likely to be closely connected to the behaviors of dark matter halos.

A gravity-only N-body simulation is the most popular and effective numerical method to predict the full 6D (position and velocity) phase-space distribution of a large number of massive particles, whose position and velocity evolve over time in the Universe [11]. They are computationally significantly less expensive than when

*The authors contributed equally to this work.
[1]The source code of our implementation is available at: https://github.com/xz2139/From-Dark-Matter-to-Galaxies-with-Convolutional-Networks

we include other complex physics such as hydrodynamics and astrophysical processes. However, these simulations do not include 'baryonic' information (i.e. galaxies distributions). To overcome this problem, different approaches have been proposed to map from the dark-matter distribution (obtained with gravity-only N-body simulations) to the galaxy distribution (see Section 2 for more details), but they suffer from a trade-off between the accuracy of important physical properties of the Universe's expected structure and the scaling abilities and time consumption. Besides, they usually rely on assumptions like halo mass being the main quantity controlling galaxy properties such as clustering.

We propose in this paper a first machine-learning based approach for this problem. We explore the use of convolutional neural networks (CNN) to perform the mapping from the 3D matter field in an N-body simulation to galaxies in a full hydrodynamic simulation. The task can be formulated as a supervised learning problem. One main difficulty of this application is the very high sparsity of the 3D output, compared to the input. We design to this end a specific two-step architecture and learning scheme that alleviates this problem. We evaluate our resulting using different statistics evaluated on the hydrodynamic simulation (e.g. power-spectrum, bispectrum) to verify the accuracy of our predictions, and to compare with a benchmark method commonly used in cosmology. We show that our approach provides more accurate galaxy distribution than the benchmark on various criteria: positions, number of galaxies, power spectrum and bispectrum of galaxies. This illustrates a better fit of the different structures and physics properties one can evaluate on the galaxy distribution. Our method also benefits from great scaling ability: we could potentially generate large volumes of realistic galaxies in a very competitive time.

We provide background and review related works in cosmology and machine learning in Section 2. Section 3 presents the data used. Section 4 presents our model's architecture. We show quantitative results and visualizations of our predicted hydrodynamic simulation in Section 5. We conclude and discuss future works in Section 6.

2 BACKGROUND AND RELATED WORKS

2.1 Cosmology

The 21st century has brought us tools and methods to observe and analyze the Universe in far greater detail than before, allowing us to probe the fundamental properties of cosmology. We have a suite of cosmological observations that allow us to make serious inroads to the understanding of our own Universe, including the cosmic microwave background (CMB) [4, 9] supernova [21] and the large scale structure of galaxies [2, 8]. In particular, large scale structure involves measuring the positions and other properties of bright sources in great volumes of the sky. These observations established a best model of the Universe, which is currently described by less than 10 parameters in the standard ΛCDM cosmology model, where CDM stands for cold dark matter and Λ stands for the cosmological constant. The parameters that are important for this analysis include the matter density $\Omega_m \approx 0.3$ (normal matter and dark matter together constitute approximately 30% of the energy content of the Universe), the variance in the matter overdensities $\sigma_8 \approx 0.8$ (the variance of the matter field density on spheres of 8 Mpc/h), and the current Hubble parameter $H_0 = 100h \approx 70$km/s/Mpc (which describes the present rate of expansion of the Universe). The model also assumes a flat geometry for the Universe. Note that the unit of distance megaparsec/h (Mpc/h) used above is time-dependent, where 1 Mpc is equivalent to 3.26×10^6 light years and h is the dimensionless Hubble parameter that accounts for the expansion of the Universe.

The amount of information collected by modern astronomical surveys is overwhelming, and modern methods in machine learning and statistics can play an increasingly important role in modern cosmology. For example, the traditional method to compare large scale structure observation and theory relies on matching the compressed two-point correlation function of the observation with the theoretical prediction (which is very hard to model on small scales, where a significant cosmological information lies). There are several other methods in the cosmological community that allow cosmologists to utilize more information from the astronomical surveys and here we discuss some of them.

Mocking up the Universe with Halo Occupation Distribution Model Halo Occupation Distribution model[6, 27, 28] (hereafter HOD) is widely used to connect dark matter halos and galaxies. HOD is based on the assumption that the probability of the presence of a galaxy at certain position on the sky (or simulation) is based solely on the mass of the dark matter halo that the galaxy will sit in. The average number of galaxies in a halo of a certain mass is a function of only the halo mass. It describes how the distribution of the galaxies is related to the distribution of the dark matter halos, therefore providing us with a way to populate N-body simulation of dark matter particles with galaxies, allowing us a direct way to compare observed distribution of galaxies on the sky and our theoretical predictions represented by simulations. However, this method comes with its own limitation: multiple tuning parameters, all galaxies live in dark matter halos and the assumption that halo mass in the main property controlling the abundance and clustering of galaxies.

Mocking up Universe with Abundance Matching Abundance matching is a popular method to connect dark matter halos with galaxies, by ranking the dark matter halos by mass and the galaxies by luminosities. We then match brighter galaxies to the heavier halos, and we keep going down the ranked lists. Similar to the HOD, this is an easy way to populate N-body simulations of dark matter particles with galaxies, allowing direct comparisons between observed distribution of galaxies in the sky and our theoretical predictions represented by the simulations. This method also relies on assumptions, like monotonic relations between halo mass and galaxy abundances.

2.2 Machine Learning in Cosmology

Convolutional neural networks are traditionally used in computer vision tasks, such as image classification, detection, and segmentation. They are increasingly being adopted in cosmology researches nowadays, and work well in representing features of Universe. [23] estimates cosmology parameters from the volumetric representation of dark-matter simulations using 3D convolutional networks with high accuracy. They showed that machine learning techniques are comparable to, and can sometimes outperform cosmology models. The paper identifies ReLU, average pooling, batch normalization

Figure 1: Visualization of Illustris simulation at redshift $z = 0$ (left), and Zoom-in visualization of corresponding dark matters and galaxies. (Right), adapted from http://www.illustris-project.org/media/

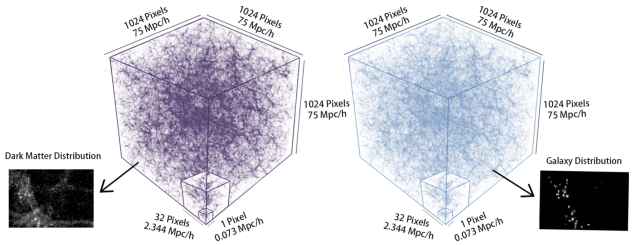

Figure 2: Spatial distribution of dark matter from the N-body simulation (left) and galaxies from the hydrodynamic simulation (right). The large boxes represent the entire simulations, while the small boxes correspond to the siezes of our voxels and training cubes.

and dropout as critical design choices in the neural network architecture to achieve highly competitive performance in estimating cosmology parameters. [18] used Extremely randomized Trees [14] to predict a hydrodynamical simulation of galaxies and found that ERT is very efficient in reproducing the statistical properties of galaxies in these hydrodynamical simulations. In [24], CNN has been demonstrated to give significantly better estimates of ω_m and σ_8 cosmological parameters from simulated convergence maps than the results from state-of-art methods, but is also free of systematic bias. Additionally, the CNN model could be interpreted by using the representations from internal layers. The similarity between a kernel and the Laplace operator inspired Ribli et al. to propose a new peak counting scheme that achieves better result than past peak counting schemes.

3 DATA

We utilize two types of simulations in this work: hydrodynamic (gravitational + hydrodynamic forces and astrophysical processes) and N-body (only gravitational forces), both from the Illustris project [13, 20, 30, 31]. The main purpose of this work is to train neural networks to predict the abundance and spatial distribution of galaxies from the very computationally expensive hydrodynamic simulation only using information from the much cheaper N-body simulation.

We use the level-1 simulations within Illustris, which is the simulation set with highest spatial and mass resolution of the suite. We focus our analysis at redshift $z = 0$, that corresponds with the current epoch of the Universe.

The cosmological model used for the Illustris simulation is in agreement with the constraints from WMAP9 (Nine-Year Wilkinson Microwave Anisotropy Probe Observations)[16].

At $z = 0$, the hydrodynamic simulation contains 5,280,615,062 gas cells, 595,243,070 stellar particles and 32,552 supermassive blackholes particles. The number of Dark Matter particles and galaxies within this snapshot is 6,028,568,000 and 4,366,546, respectively. The N-body simulation only contains 6,028,568,000 dark matter particles. Figure 1 shows the spatial distribution of dark matter in the N-body simulation as well as a close-up on a smaller region.

We compute the density fields of galaxies and dark matter by assigning each component to a regular grid with 1024^3 voxels using the nearest-grid point mass assignment scheme: if a galaxy or dark matter particle is inside a given voxel, the value of that cell is increased by 1 for the corresponding field.

Within the grid, the number of particles in each voxel ranges from 0 to 747,865 for dark matter and from 0 to 10 for galaxies. The percentage of non-zero cells is 44.99% and 0.37% for dark matter and galaxies, correspondingly. The low occupancy of galaxies in the grid poses an interesting challenge for our work. Fig. 2 shows the distribution of dark matter and galaxies from the N-body and hydrodynamic simulations, respectively. The voxels are shown to demonstrate the gridding we perform.

The simulation density fields are then separated into sub-cubes of size 32^3 voxels, corresponding to regions of size around 2.3 Mpc/h, and are used as independent samples. There are 32,768 unique sub-boxes and they are split spatially into three chunks. 62.6% of all boxes are used for training, 19.63% of the boxes are used for validation and then the other 17.76% are used for testing. Testing data are retained as a concatenated cube of size 42.4 Mpc/h for the ease of computation of relevant statistics.

4 METHODS

Here we present our approach for linking the 3D dark matter field from N-body simulations to the 3D galaxy distribution from hydrodynamic simulations.

The two key challenges in predicting galaxy positions from the dark matter field are (i) the inherently spatial nature of the data (dark matter and galaxies are structured spatially, on various correlated scales), (ii) the high sparsity of the galaxy (target) distribution.

To address the first aspect, we propose to rely on convolutional networks. They naturally provide interesting properties for our problem such as translational invariance [19]. Convolutional networks are also commonly employed for extraction of spatial patterns [23]. To address the second aspect, we developed a *two-phase* architecture and learning process. We present in the following section the details of this architecture, and we discuss the different convolutional networks we tested in our experiments.

4.1 Two-phase architecture

The high sparsity in our simulation dataset (99.6% of output voxels do not contain any galaxies) makes our training challenging. Because of imbalanced distribution between input and output, the model could easily achieve a high accuracy even if it fails in predicting all the galaxies. This slows down the training process to a great extent. In order to overcome this problem, we propose the following two-phase architecture.

The main idea is to break down the training into separate processes. The whole model is composed of two parts. The first part is a classifier, which predicts the presence or absence of galaxies as a probability for each voxel representing one part of Universe. Using a binary classifier as a first "layer" allows us to use special loss functions designed for such high sparsity prediction. Specifically, we use weighted cross-entropy loss, which penalizes wrong predictions with high probability, and introduces weights to correct imbalances in classes. For a single output voxel, it can be written as follows:

$$\mathbb{L}_{\text{CrossEnt}}(\hat{p}, \mathbf{y}) = -(w \cdot \mathbf{y} \cdot \log(\hat{p}) + (1 - \mathbf{y}) \cdot \log(1 - \hat{p})) \tag{1}$$

where $\hat{p}$ is the vector of predicted probability of the presence of at least one galaxy in the considered voxel. $\mathbf{y}$ is the actual target value (1 if there exists at least 1 galaxy in the voxel, 0 otherwise). w characterizes the weight applied for counter-balancing the large number of voxels without galaxies[2].

This first prediction is then used as a mask for the final prediction of the number of galaxies in each voxel. The second step of the network is optimized only on the voxels that are expected to contain at least a galaxy, according to the binary prediction result from the first phase. We propose to use a $L2$-loss since we decide to predict a probabilistic number of galaxies in each voxel, and expect to have a real value as output. The complete loss of the model for a single output voxel is illustrated below:

$$\mathbb{L}(n_g, \hat{p}, n_t) = M(\hat{p})(n_g - n_t)^2 \tag{2}$$

$$M(\hat{p}) = \begin{cases} 1 & \hat{p} > 0.5 \\ 0 & Otherwise \end{cases} \tag{3}$$

Where M is a function of the first-phase output ($\hat{p}$) for the given voxel, which returns 1(0) if we expect to see at least 1 galaxy in the voxel (or not). n_g is the prediction of the second-phase model and n_t is the actual target: number of galaxies in the output voxel.

From an experimental point of view, the training of both parts is done separately. We select the first classifier (e.g. with the highest recall) and then train the second part of the model. More details are given in Sec 5. A schema of this generic architecture is shown in Figure 3 for a better visualization of the process.

This two-phase set-up is quite generic, and allows us to build different types of architecture depending on the choice of networks for each phase.

[2]Details on the selection for w are given in Section 5 and in Supplementary Materials

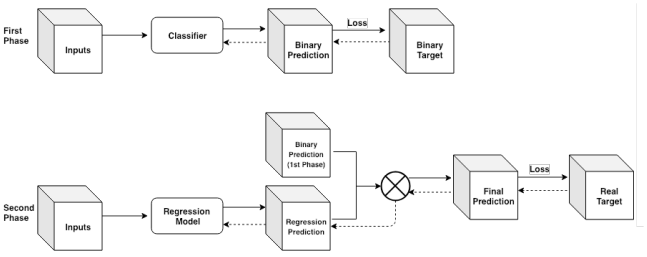

Figure 3: Two Phase Model Structure

4.2 Network architectures

We now present different types of convolutional networks we tested, with their physical motivations and specific modifications to better fit our problem. The models mentioned below can be used for both classification(first-phase) and regression(second-phase). We will compare different choices of networks in Section 5, as well as a more classical "one-phase" training.

U-Net. U-Net is a fully convolutional neural network, first proposed in [25] for bio-medical image segmentation. The networks is composed of a *contracting path* and a symmetric *expanding path*. The first path is a typical convolutional architecture, with convolutions followed by a rectified linear unit (ReLU) and max-pooling operation. The number of channels is increased at each step. This part aims at capturing spatial relations and context. The expanding path relies on up-sampling functions on the feature map, followed by *up-convolutions* that reduce the number of channels. Additionally, a *skip connection* is added at each level, which concatenates the up-sampled features and the corresponding map from the contracting path. This part provides the network with various levels of granularity for the final prediction, usually segmentation, which is in a similar shape as the input.

This type of network can be easily adapted to 3D data and has been successfully applied in different applications, for instance in cosmology [15], or for volumetric segmentation on medical data [7]. As its architecture is constructed to map between input and output with similar shapes and to extract spatial information on multiple scales, it appears as a good candidate to learn the relationship between dark-matter halos and the distribution of galaxies.

Our early experiments on this model structure showed that the prediction had a strong similarity in distribution with inputs instead of targets, which constrains the model from generalizing to larger scale. Thus, we proposed one modification to the original architecture. The last (topmost) skip-connection was removed to prevent the model from "feeding" too much information from the dark-matter halos on high resolution features. The final U-Net architecture used in the experiments is illustrated in Figure 4.

Recurrent Residual U-Net (R2Unet). Recurrent Residual U-Net (R2U-Net) were proposed in [1] as an upgrade of U-Net. The authors propose different variations around the U-Net architecture, but we focus here on the Recurrent-Residual one. The main idea is to change the convolution functions used in the U-Net architecture. Instead of using classical convolution functions, R2U-Net relies on a composition of two stacked Recurrent Convolutions (RCNN), as

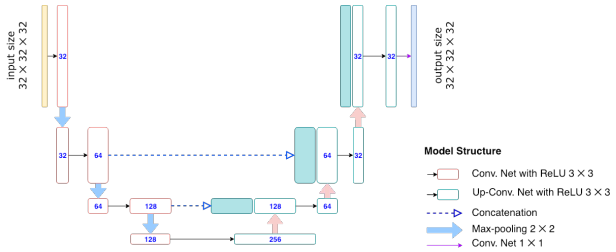

Figure 4: U-Net architecture with removed *skip*-connection on the upper layer.

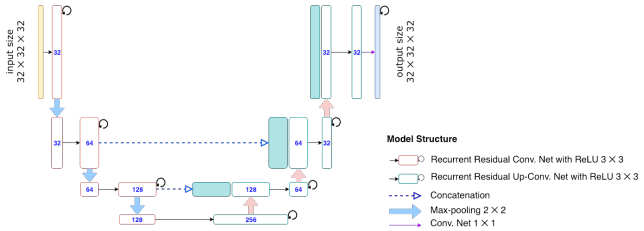

Figure 5: Recurrent Residual U-Net (R2U-Net) with removed upper skip-connection.

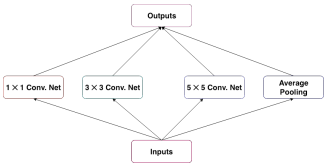

Figure 6: Inception Module v1 used in our network.

Figure 7: Architecture of the neural network used with Inception module (called in the remaining of the paper *Inception*).

presented in [22], with a residual connection from the input to the output. The recurrent aspect of RCNN allows to produce arbitrarily deeper architecture without increasing the number of parameters, while allowing the model to refine and aggregate the extraction of features through "time" (steps applied during the recurrent convolution). This helps to accumulate the feature representation. The residual connection on the other hand allows for the propagation of higher level information at each layer.

We propose to use this model, here again with the modified U-Net global architecture where we remove the upper skip connection.

Inception Net. Inception networks [26] were developed to handle the variation at the scale of the salient parts of images. The salient information can come in multiple scales . With "vanilla" convolutional networks, this leads to the difficulty of choosing the kernel size for the convolution function: if the information is structured on larger scales, one should choose a bigger kernel size, and inversely if the information is more locally distributed. Inception module proposes to horizontally stack convolution functions with different kernel sizes. Different variations around this key idea have been proposed in order to optimize the computation cost of widening the architecture.

In our application, it is very likely that the information extracted from the dark-matter halos is distributed on different scales, globally and locally. This motivates the use of the Inception module. More specifically, we propose to use the first original version of the Inception architecture module with 3 different kernel sizes of the 3 dimensional filters (1x1x1, 3x3x3, 5x5x5) and pooling layer. We use average pooling as we found it yields better results. The outputs of all the filters are concatenated and passed on as input to the subsequent layers. To limit the amount of parameters, we use convolutional functions as subsequent layers instead of fully connected ones, more specifically two convolutional layers to match the size of the target output. We use Sigmoid function as the final activation function of CNN network to output the probability that there is at least a galaxy in the voxel. Figure 7 shows the whole model structure.

5 RESULTS

In this section we describe the results we obtain with different architectures of convolutional neural networks and compare them with a benchmark method commonly used in cosmology, HOD (Halo Occupation Distribution). Our main goal is to accurately predict the abundance and spatial distribution of galaxies by using information only from the 3D dark-matter field. By using different cosmological observables, we show how our method outperforms, or is competitive with HOD on all the considered observables.

5.1 Experimental protocol

As described in Sec 3, our models take as input the density field of dark matter from the N-body simulation in 3D sub-boxes containing 2^{15} voxels, and predict the 3D galaxy field from the hydrodynamic simulation. We retain 62.6% of the 2^{15} total sub-boxes for training, 19.63% for validation and 17.76% for test. The split between training, validation, and test is made following a "global" cut, more specifically to enforce that the test sub-boxes form a larger cube with 42.2 Mpc/h on the side. This is motivated by the desire to compare our observables in a larger range of scales.

We compare our results to those from *halo occupation distribution* (HOD) algorithm (see e.g. [5, 29]), a method commonly used to link dark matter halos to galaxies in cosmology. The underlying idea behind the HOD is that all galaxies reside within halos, and galaxies can be split into centrals and satellites. Our HOD has three free-parameters: $M_{\rm min}$, M_1 and α and the algorithm is as follows. Only halos with masses greater than $M_{\rm min}$ will host a central galaxy, that will be placed on the halo center. The number of satellites galaxies follow a Poisson distribution with mean $(M/M_1)^\alpha$, that are placed randomly within the dark matter halo.

Table 1: Performance on binary prediction of galaxies

Model	Configuration	Accuracy	Recall	Precision
Inception	Weight: 80	96.32	**95.72**	10.15
U-Net	Weight: 5	99.6	59.8	39.2
R2Unet	Weight: 5	99.52	63.17	41.91
R2Unet	Weight: 10	99.29	74.8	32.42
R2Unet	Weight: 25	98.8	84.31	21.05
R2Unet	Weight: 80	97.41	92.49	13.52
HOD	—	**99.93**	86.80	**94.76**

Given a set of (M_1, α), we fit for M_{min} to match the predicted galaxy density with the true galaxy density from the target simulation with a threshold of 0.001. The only free-parameters are thus M_1 and α. They are optimized by minimizing the squared difference between the power spectrum computed on predicted galaxies and the power spectrum observed on the actual galaxies. In the following experiments, these parameters are optimized on the test sub-box observations.

5.2 Galaxy distribution - binary prediction and quantity prediction

Trying to optimize directly the number of galaxies per voxel proved ineffective due to the high sparsity of the output data. This motivates us to first perform a binary prediction, which predicts whether there is at least one galaxy in a voxel. We use accuracy, recall and prediction as metrics to evaluate the performance of different models. However, because of the sparsity of the data, high accuracy doesn't necessarily represent a good model as a model that predicts every voxel to have zero galaxies will achieve an accuracy of 99.57%. Table 1 shows accuracy, recall and prediction for different models. Our experiments show that the Inception-based network provides the best recall at 95.72%. We observe that in this binary-setup, HOD also performs well in terms of trade-off between recall and precision, with a high accuracy.

Following the results of binary prediction, we select the model with the highest recall as our first-phase model. By doing so, we alleviate the sparsity problem in the data, which allows for a small unrecoverable error in accuracy but manages to obtain a higher precision for the prediction in the second phase model. The second-phase model therefore focuses on reducing number of false positives (aka. improving precision) and predicting the a probabilistic number of galaxies in each voxel. Table 2 shows mean square error for different machine learning models and our HOD benchmark. The model that yields best performance is the two phase model with Inception network as first phase, and R2U-Net as second phase. Our approach significantly outperforms HOD in this set-up, which seems to indicate that while HOD predicts correctly the region of absence/presence of galaxies, it is much more imprecise when predicting the number of galaxies in each voxel. We will see in the next subsections that this will impact the different statistical measures one can make on the Universe.

A visualization of the predictions of different models is provided in Figure 9. Each row represents a "slice" of the simulations, with 8.89 Mpc/h on the side. As a reminder, the sub-boxes taken as

Table 2: Mean-Square Error evaluation for number of galaxies prediction

Model	Configuration	MSE
R2Unet+R2Unet	Weight: 80/0.6	0.00320
Inception+R2Unet	Weight: 80/0.8	**0.00308**
HOD	—	0.01007

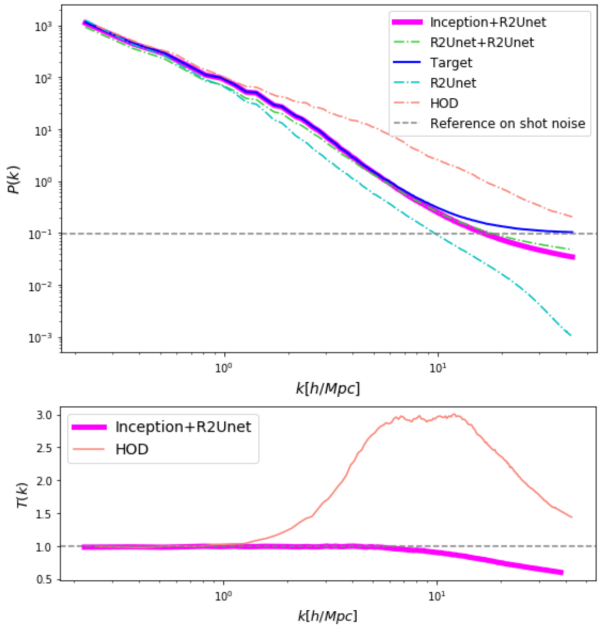

Figure 8: Power spectrum (Top) and Transfer Function (Bottom) comparison among different machine learning models, HOD model and target. The reference shot noise is the Poisson noise in the power spectrum resulting from Poisson sampling. Two-phase models performs as good as HOD on large scales (left of the upper plot, k<1 *h*/Mpc) and outperforms HOD on smaller scales.

inputs and outputs of our models are of size $2.34\,(\mathrm{Mpc/h})^3$, and are depicted as a red square on the upper-left-most of the Figures. The left-most column is the dark-matter halos' masses taken as input. The second column depicts the corresponding target number of galaxies, where brighter pixels represent a higher number of galaxies. We can directly observe the high sparsity of the problem. The two-phase models predictions are depicted in the third and fourth column. In fifth column are predictions from a single-phase model. This visualization also illustrates the behavior of a single-phase training, which has trouble refining the galaxies predictions. Last column are HOD predictions.

5.3 Two-Point Correlation and Power Spectrum

It is a standard practice in cosmology to extract information from observations via summary statistics. The most commonly used

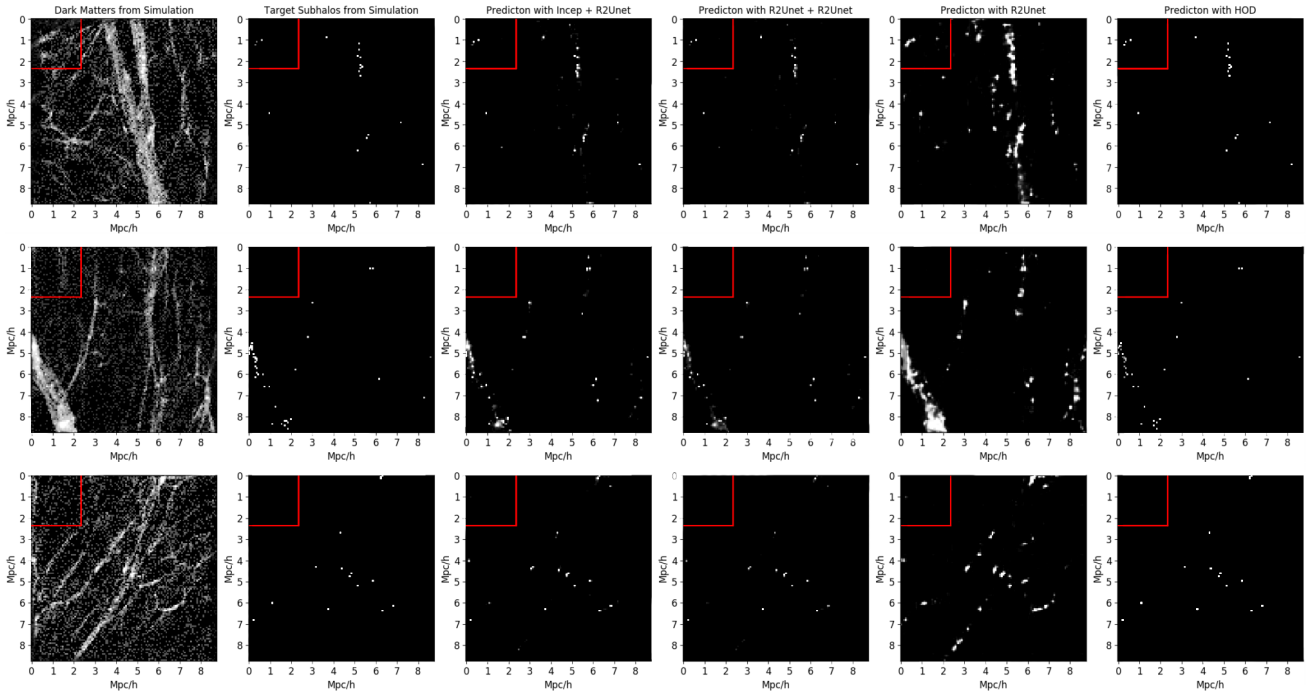

Figure 9: Visualization of slices of the simulations: first column are dark-matter halos, second column are the corresponding target galaxies. 3d and 4th columns are predictions from our two-phase models, 5th from a single-phase classifier, and last column are HOD predictions. Red square represents the size of the boxes taken as input by our models.

statistics is the two-point correlation function $\xi(r)$, defined as the excess probability, compared with that expected for a random distribution, of finding a pair of galaxies at a separation. It measures how the actual distribution of galaxies deviates from a simple random distribution. The power spectrum, $P(k)$, is the Fourier transform of the two-point correlation function:

$$\begin{aligned} \xi(|\boldsymbol{r}|) &= \langle \delta_A(\boldsymbol{r'}) \delta_B(\boldsymbol{r'} + \boldsymbol{r}) \rangle \\ P(|\boldsymbol{k}|) &= \int \mathrm{d}^3\boldsymbol{r}\ \xi(r) e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \end{aligned} \tag{4}$$

These two statistics are very important in cosmology, because they allow to extract all the information embedded into Gaussian density fields, as our Universe resembles on large-scales or at earlier times. In this paper we focus our attention on the power spectrum. We define the transfer function, $T(k)$, as

$$T(k) = \sqrt{\frac{P_{\text{pred}}(k)}{P_{\text{target}}(k)}} \tag{5}$$

and use it to quantify the performance of the models against the ground truth.

Figure. 8 shows the power spectrum and transfer function for the different models. Our two-phase model with Inception+R2Unet manages to reproduce the clustering of galaxies of the original data. Interestingly, it manages to obtain a good fit on a large range of scales, even though it is trained on relatively "small" sub-boxes. Comparing to the HOD results, our model achieves nearly the same performance when $k < 1$ h/Mpc, and outperforms when $k > 1$ h/Mpc. This is consistent with the fact that HOD is not being designed to work well on small scales. While the $P(k)$ of the benchmark method has significantly differ from the target's at $k =$ 1h/Mpc, $P(k)$ from our two-phase approach with Inception+R2Unet only begins to differ from the target at k $\approx$ 8 h/Mpc. This is likely due to the fact that the field is highly non-linear at those scales, which makes the model harder to learn. Furthermore, the fact that the galaxies are approximately poisson distributed produce a 'shot-noise' floor. This affects the power spectra prominently on small scales: it adds a white component, $1/\bar{n}_{gal}$, to the power spectra, where $\bar{n}_{gal}$ is the average number density of the galaxies in the simulation box. Its effect can be clearly seen in Fig. 8.

5.4 Three-Point Correlation and Bispectrum

The Universe, on large-scales, resembles a Gaussian field, and therefore, can be fully characterized by its 2pt correlation function or power spectrum. However, on small scales, non-linear gravitational evolution changes the density field into a non-Gaussian field. In order to characterize the spatial distribution of the Universe on small scales, where most of the cosmological information lies, higher-order statistics are needed. Here we concentrate on the bispectrum, the Fourier equivalent of 3 point correlation function, defined as

$$B(k_1, k_2, k_3)\delta(\boldsymbol{k}_{123}) = \langle \delta_{\boldsymbol{k}_1} \delta_{\boldsymbol{k}_2} \delta_{\boldsymbol{k}_3} \rangle \tag{6}$$

Figure 10 shows the bispectrum for our Inception+R2Unet model, the benchmark HOD model and the target. At large scales, small wavenumber (k), the bispectra of our model and benchmark models are consistent with that of the target. The mean relative bispectrum residual of our model and HOD model compared to the target at $k_1 = 0.5$ h/Mpc and $k_2 = 0.6$ h/Mpc is 2.7% and 5.0% respectively. On smaller scales, at $k_1 = 1.2$ h/Mpc and $k_2 = 1.3$ h/Mpc, the

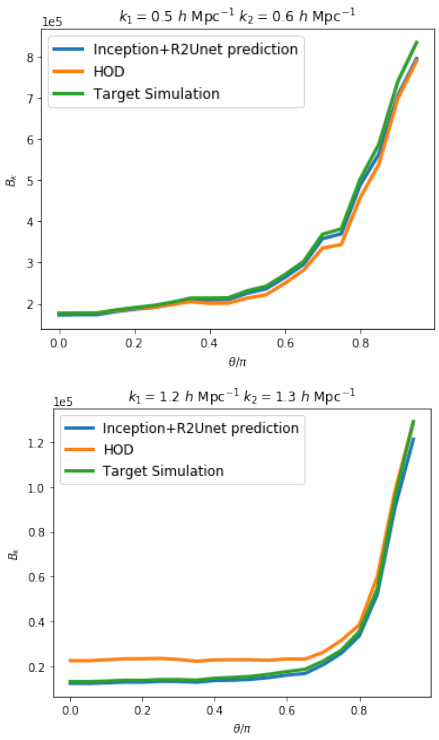

Figure 10: Bispectrum comparison between our Inception+R2Unet model, benchmark HOD model and Target Simulation. Upper panel: Bispectrum at $k_1 = 0.5\ h$/Mpc and $k_2 = 0.6\ h$/Mpc ("large" scale). Bottom Panel: Bispectrum at $k_1 = 1.2\ h$/Mpc and $k_2 = 1.3\ h$/Mpc ("small" scale). Inception+R2UNet outperfoms benchmark HOD model on both scales, and significantly so on smaller scales.

corresponding mean relative residual is 0.68% and 1193%. Our model reproduces the highly non-linear galaxy field on small scales far better than the benchmark model. This suggests that our machine learning model has large enough flexibility to reproduce the galaxy distribution from large to small scales even when we consider the higher order function, while the state-of-art benchmark produces bispectrum that is 5-6 times larger than the target at all scales.

5.5 Voids

The so-called cosmic web, i.e. the spatial distribution of matter on the Universe, is made up of high-density regions, clusters, where hundreds or thousands of galaxies resides. Galaxy clusters are connected by medium-density regions that contain highly ionized gas; the filaments. Finally, filaments are surrounded by enormous empty regions named voids. These voids are a very important element of the cosmic web, since most of the volume of the Universe resides on them. Given their unique nature, they embed a large amount of cosmological information. In this work we study the abundance of voids, as a function of their radii: the void size function. We identify voids in the galaxy distribution of the different models and the target using the algorithm described in [3].

We compare the void size function from our Inception+R2Unet model, the benchmark HOD model and the target simulation. The results are presented in Figure 11. The size function of voids from

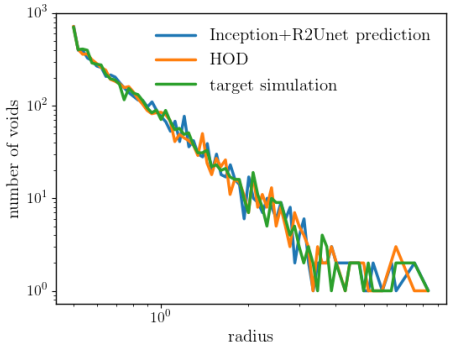

Figure 11: Number of voids as a function of their radii for the HOD (orange), Inception+R2UNet (blue) and the target (green). Both models are able to reproduce the abundance of voids from the target with high accuracy.

Table 3: Running time of different models

Model	Device	CPU/GPU Hours
Illustris	CPU	19 million
HOD	1CPU	8
Inception + R2Unet	1GPU (GTX1080)	3

Inception+R2Unet model and benchmark HOD are both consistent with that of the target simulation, indicating our R2Unet+Inception model is competitive against the benchmark in this large scale observable.

5.6 Training time

Another key aspect of our approach is its scaling abilities. Table 3 shows the time needed to train and/or generate one simulation of galaxies using these various methods. HOD takes comparable amount of time to optimize. However, once trained, our method can generate large volumes of galaxy distribution with negligible time, and is flexible enough to generate excellent match to the target from small to large scales.

6 CONCLUSIONS

In this paper, we present a deep-learning based approach to model the link between the underlying dark matter from N-body simulations and the galaxies distribution from full hydrodynamic simulations. We design a specific learning scheme and model architecture to overcome the very high sparsity of the task. We show that our approach, by optimizing directly the number of galaxies prediction per voxel, manages to reproduce a large variety of important physical properties of the original data. It outperforms, or is as efficient as the standard benchmark method of the field, on various important cosmological statistics, while having much more scaling and generalization abilities. This is a first encouraging step to overcome the need for computationally expensive hydrodynamic simulations in the long run.

This work opens several trails for future research. First, it will be very interesting to extend our model to be able to predict not only the number and positions of the galaxies but also their internal properties, e.g. stellar mass, star-formation rate, metallicity, etc.

Training the model at different epochs of the Universe will allow us to better understand the complicated physics involved in galaxy formation/evolution. Training our model on simulations with different strengths of the most relevant astrophysical processes, such as active galactic nuclei and supernova feedback, will enable us to marginalize over these astrophysical complications and extract robust cosmological information.

Our approach can be used to populate the dark matter halos of very big gravity-only simulations with galaxies, without relying on the standard assumptions involved in the classical HODs. This will open new doors in cosmology, allowing us to investigate some of the most important theoretical systematics on cosmology such as baryonic effects. Since our framework allow us to model the spatial distribution of realistic galaxies down to very small scales, it can be used to extract the maximum information from cosmological observations. Our results can thus have a major impact on cosmology and will establish a new link between astrophysics and cosmology.

ACKNOWLEDGEMENT

We thank David Spergel, Siamak Ravanbakhsh and Barnabas Poczos for insightful discussions. This project is supported by Center for Computational Astrophysics of the Flatiron Institute in New York City. The Flatiron Institute is supported by the Simons Foundation.

REFERENCES

[1] Md. Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. 2018. Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation. *CoRR* abs/1802.06955 (2018). arXiv:1802.06955 http://arxiv.org/abs/1802.06955

[2] Lauren Anderson, Éric Aubourg, Stephen Bailey, Florian Beutler, Vaishali Bhardwaj, Michael Blanton, et al. 2014. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Releases 10 and 11 Galaxy samples. *Monthly Notices of the Royal Astronomical Society* 441, 1 (2014), 24–62.

[3] A. Banerjee and N. Dalal. 2016. Simulating nonlinear cosmological structure formation with massive neutrinos. *Journal of Cosmology and Astroparticle Physics* 11, Article 015 (Nov. 2016), 015 pages. https://doi.org/10.1088/1475-7516/2016/11/015 arXiv:1606.06167

[4] CL Bennett, D Larson, JL Weiland, N Jarosik, G Hinshaw, et al. 2013. Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: final maps and results. *The Astrophysical Journal Supplement Series* 208, 2 (2013), 20.

[5] Andreas A Berlind and David H Weinberg. 2002. The halo occupation distribution: toward an empirical determination of the relation between galaxies and mass. *The Astrophysical Journal* 575, 2 (2002), 587.

[6] Andreas A. Berlind and David H. Weinberg. 2002. The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass. *The Astrophysical Journal* 575 (Aug. 2002), 587–616. https://doi.org/10.1086/341469 arXiv:astro-ph/astro-ph/0109001

[7] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 424–432.

[8] Alison L. Coil, Brian F. Gerke, Jeffrey A. Newman, Chung-Pei Ma, Renbin Yan, Michael C. Cooper, Marc Davis, S. M. Faber, Puragra Guhathakurta, and David C. Koo. 2006. The DEEP2 Galaxy Redshift Survey: Clustering of Groups and Group Galaxies at z~1. *The Astrophysical Journal* 638 (Feb. 2006), 668–685. https://doi.org/10.1086/498885 arXiv:astro-ph/astro-ph/0507647

[9] Planck Collaboration, Y Akrami, F Arroja, M Ashdown, J Aumont, C Baccigalupi, M Ballardini, AJ Banday, RB Barreiro, N Bartolo, S Basak, et al. 2018. Planck 2018 results. I. Overview and the cosmological legacy of Planck. *arXiv preprint arXiv:1807.06205* (2018).

[10] Colless, M. et al. 2001. The 2dF Galaxy Redshift Survey: Spectra and redshifts. *Mon. Not. Roy. Astron. Soc.* 328 (2001), 1039. https://doi.org/10.1046/j.1365-8711.2001.04902.x arXiv:astro-ph/astro-ph/0106498

[11] M. Davis, G. Efstathiou, C. S. Frenk, and S. D. M. White. 1985. The evolution of large-scale structure in a universe dominated by cold dark matter. (May 1985), 371-394 pages. https://doi.org/10.1086/163168

[12] Olivier Doré, Christopher Hirata, Yun Wang, David Weinberg, Ivano Baronchelli, Andrew Benson, Peter Capak, Ami Choi, Tim Eifler, Shoubaneh Hemmati, et al. 2018. WFIRST Science Investigation Team" Cosmology with the High Latitude Survey" Annual Report 2017. *arXiv preprint arXiv:1804.03628* (2018).

[13] Shy Genel, Mark Vogelsberger, Volker Springel, Debora Sijacki, Dylan Nelson, Greg Snyder, Vicente Rodriguez-Gomez, Paul Torrey, and Lars Hernquist. 2014. Introducing the Illustris Project: the evolution of galaxy populations across cosmic time. MNRAS 2014 445 (2): 175-200. (2014). https://doi.org/10.1093/mnras/stu1654 arXiv:arXiv:1405.3749

[14] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63, 1 (01 Apr 2006), 3–42. https://doi.org/10.1007/s10994-006-6226-1

[15] Siyu He, Yin Li, Yu Feng, Shirley Ho, Siamak Ravanbakhsh, Wei Chen, and Barnabás Póczos. 2018. Learning to Predict the Cosmological Structure Formation. *arXiv preprint arXiv:1811.06533* (2018).

[16] Gary Hinshaw, D Larson, Eiichiro Komatsu, DN Spergel, et al. 2013. Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: cosmological parameter results. *The Astrophysical Journal Supplement Series* 208, 2 (2013), 19.

[17] Jones, H.D. et al. 2009. The 6dF Galaxy Survey: Final Redshift Release (DR3) and Southern Large-Scale Structures. *Mon. Not. Roy. Astron. Soc.* 399 (2009), 683. https://doi.org/10.1111/j.1365-2966.2009.15338.x arXiv:astro-ph.CO/0903.5451

[18] Harshil M Kamdar, Matthew J Turk, and Robert J Brunner. 2016. Machine learning and cosmological simulations–II. Hydrodynamical simulations. *Monthly Notices of the Royal Astronomical Society* 457, 2 (2016), 1162–1179.

[19] Stéphane Mallat. 2016. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society of London Series A* 374 (April 2016), 20150203. https://doi.org/10.1098/rsta.2015.0203 arXiv:stat.ML/1601.04920

[20] Dylan Nelson, Annalisa Pillepich, Shy Genel, Mark Vogelsberger, Volker Springel, Paul Torrey, Vicente Rodriguez-Gomez, Debora Sijacki, Gregory F. Snyder, Brendan Griffen, Federico Marinacci, Laura Blecha, Laura Sales, Dandan Xu, and Lars Hernquist. 2015. The Illustris Simulation: Public Data Release. Astronomy and Computing (2015), pp. 12-37. (2015). https://doi.org/10.1016/j.ascom.2015.09.003 arXiv:arXiv:1504.00362

[21] S Perlmutter, Supernova Cosmology Project Collaboration, et al. [n. d.]. Measurements of omega and lambda from 42 high-redshift supernovae 1999 Astrophys. *J* 517 ([n. d.]), 565.

[22] Pedro HO Pinheiro and Ronan Collobert. 2014. Recurrent convolutional neural networks for scene labeling. In *31st International Conference on Machine Learning (ICML)*.

[23] Siamak Ravanbakhsh, Junier B Oliva, Sebastian Fromenteau, Layne Price, Shirley Ho, Jeff G Schneider, and Barnabás Póczos. 2016. Estimating Cosmological Parameters from the Dark Matter Distribution.. In *ICML*. 2407–2416.

[24] Dezső Ribli, Bálint Ármin Pataki, and István Csabai. 2018. Learning from deep learning: better cosmological parameter inference from weak lensing maps. *arXiv preprint arXiv:1806.05995* (2018).

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[26] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[27] Jeremy L. Tinker. 2007. Redshift-space distortions with the halo occupation distribution - II. Analytic model. *Monthly Notices of the Royal Astronomical Society* 374 (Jan. 2007), 477–492. https://doi.org/10.1111/j.1365-2966.2006.11157.x arXiv:astro-ph/astro-ph/0604217

[28] Mohammadjavad Vakili and ChangHoon Hahn. 2016. How are galaxies assigned to halos? Searching for assembly bias in the SDSS galaxy clustering. *arXiv e-prints*, Article arXiv:1610.01991 (Oct. 2016), arXiv:1610.01991 pages. arXiv:astro-ph.CO/1610.01991

[29] Francisco Villaescusa-Navarro, Federico Marulli, Matteo Viel, Enzo Branchini, Emanuele Castorina, Emiliano Sefusatti, and Shun Saito. 2014. Cosmology with massive neutrinos I: towards a realistic modeling of the relation between matter, haloes and galaxies. *Journal of Cosmology and Astroparticle Physics* 2014, 03 (2014), 011.

[30] Mark Vogelsberger, Shy Genel, Volker Springel, Paul Torrey, Debora Sijacki, Dandan Xu, Gregory F. Snyder, Simeon Bird, Dylan Nelson, and Lars Hernquist. 2014. Properties of galaxies reproduced by a hydrodynamic simulation. (2014). https://doi.org/10.1038/nature13316 arXiv:arXiv:1405.1418

[31] Mark Vogelsberger, Shy Genel, Volker Springel, Paul Torrey, Debora Sijacki, Dandan Xu, Gregory F. Snyder, Dylan Nelson, and Lars Hernquist. 2014. Introducing the Illustris Project: Simulating the coevolution of dark and visible matter in the Universe. (2014). https://doi.org/10.1093/mnras/stu1536 arXiv:1405.2921

[32] Hu Zhan and J. Anthony Tyson. 2018. Cosmology with the Large Synoptic Survey Telescope: an overview. *Reports on Progress in Physics* 81, Article 066901 (June 2018), 066901 pages. https://doi.org/10.1088/1361-6633/aab1bd arXiv:astro-ph.CO/1707.06948

SUPPLEMENT

Data Access and Preparation

The data used for this work is publicly available at Illustris website[3].

We use Illustris-1 (hydrodynamical simulation) and Illustris-1-Dark (N-body Dark matter particle only simulation), which have the largest resolutions. The simulation consists of 6,028,568,000 dark matter particles and of an equal number of hydrodynamic voronoi cells in the hydrodynamic simulation.

We use the snapshot at redshift $z = 0$, which is the current universe. It has a volume of $75^3(\text{Mpc/h})^3$ (*Megaparsec*, 1 Mpc = 3.09×10^{22} meters). The governing cosmological parameters are Ω_m = 0.2726, Ω_Λ=0.7274, Ω_b=0.0456, σ_8 = 0.809, n_s = 0.963, H_0 = 100 $h\ kms^{-1}Mpc^{-1}$ and h = 0.704.

We take the positions of dark matter particles (from Illustris-1-Dark) and subhalos (from Illustris-1, where galaxies reside in), and grid the box into $1024 \times 1024 \times 1024$ pixels, obtaining 32786 non-overlapping sub-cubes of size $32 \times 32 \times 32$. We count the number of particles/subhalos in each pixel using Nearest Gird Point method. We use the top-left $13 \times 13 \times 31$ subboxes for validation, the bottom right $18 \times 18 \times 18$ subboxes for test, and the rest of subboxes are used for training.

Hyper-parameters search and Model Configuration

Table. 4 shows the various hyper-parameters and the search space for each parameter.

Table 4: Hyper-parameters space search

Configurations	Description	Search Space
lr	learning rate	$10^{-5} - 10^{-3}$
epochs	number of epochs	20-40
batch_size	batch size	16, 32
loss_weight	weight w of the loss function	0.6-80
conv1_out	number of hidden units for the 1x1x1 kernel	3-40
conv3_out	number of hidden units for the 3x3x3 kernel	4-60
conv5_out	number of hidden units for the 5x5x5 kernel	5-80
optimizer	optimizer for training	SGD, Adam

The best configuration for the classifier (the first phase) in the two-phase model (Inception + R2Unet) is: batch_size=16, conv1_out=6, conv3_out=8, conv5_out=10, epochs=20, loss_weight=80, lr=0.001, optimizer = Adam.

And the best configuration for the regression (the second phase) in the two-phase model (Inception + R2Unet) is: batch_size=16, epochs=20, loss_weight=0.8, lr=0.001, optimizer = Adam.

Training Setup

The experiments are carried out on NYU HPC cluster with one Intel Xeon E5-2690v4 2.6GHz CPU, one NVIDIA GTX 1080 GPU and 60GB of RAM. We use Python version 3.5.3 and Pytorch version 0.4.1. The code is available in Github Repository: https://github.com/xz2139/From-Dark-Matter-to-Galaxies-with-Convolutional-Networks. The codes for evaluation on HOD (power spectrum, bispectrum and void finder) are also available publicly at the following Github Repository https://github.com/franciscovillaescusa/Pylians.

As we described in Section 4, our model has two phases: classifier (first phase) and regression (second phase). In the first classifier phase, we train a classifier to get the binary prediction for the location of the galaxies (Algorithm. 2). Since the output of the model is the binary prediction, we convert target density fields into binary values before training the model (Algorithm 1).

Data: All Training Sub-cubes (Targets Only)
Result: Binary Targets
for *i in all training range(sub-cubes)* **do**
  if *Targets[i] > 0* **then**
    Binary_Targets[i] = 1;
  else
    Binary_Targets[i] = 0;
  end
end

Algorithm 1: Turn targets into binary value

Data: All Training Sub-cubes (Inputs and Binary_Targets)
Result: Binary_Prediction
for *i in all training range(sub-cubes)* **do**
  Binary_Prediction = Inception_Net(sub-cubes[i]);
  Loss = Cross_Entropy(Prediction, Binary_Targets[i]);
  Back-propagation;
end

Algorithm 2: Classifier Running Process

The binary prediction from the first phase will then serve as a mask for the second phase, where only the masked region is trained on. In the second phase regression model, the input of the model is the same as the classifier model, but the targets are the real number of galaxies.

To achieve the masking, we multiply the outputs from the regression model with our binary prediction from the classifier model. The algorithm for the second phase is shown in Algorithm 3.

Data: All Training Sub-cubes (Inputs and Real_Targets),
  Binary_Prediction from the first phase
Result: Final_Prediction
for *i in all training range(sub-cubes)* **do**
  Prediction = R2UNet(sub-cubes[i]);
  Final_Prediction = Binary_Prediction · Prediction;
  Loss = L2_Loss(Final_Prediction, Real_Targets[i]);
  Back-propagation;
end

Algorithm 3: Regression Model Running Process

[3] http://www.illustris-project.org/data/

Learning to Predict the Cosmological Structure Formation

Siyu He[a,b,c,1], **Yin Li**[d,e,f], **Yu Feng**[d,e], **Shirley Ho**[c,e,d,a,b,1], **Siamak Ravanbakhsh**[g], **Wei Chen**[c], **and Barnabás Póczos**[h]

[a]Physics Department, Carnegie Mellon University, Pittsburgh PA 15213; [b]McWilliams Center for Cosmology, Carnegie Mellon University, Pittsburgh, PA 15213, USA; [c]Center for Computational Astrophysics, Flatiron Institute, New York, NY 10010; [d]Berkeley Center for Cosmological Physics, University of California, Berkeley, CA 94720, USA; [e]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; [f]Kavli Institute for the Physics and Mathematics of the Universe, University of Tokyo Institutes for Advanced Study, The University of Tokyo, Chiba 277-8583, Japan; [g]Computer Science Department, University of British Columbia, Vancouver, BC V6T1Z4, Canada; [h]Machine Learning Department, Carnegie Mellon University, Pittsburgh PA 15213

Matter evolved under influence of gravity from minuscule density fluctuations. Non-perturbative structure formed hierarchically over all scales, and developed non-Gaussian features in the Universe, known as the Cosmic Web. To fully understand the structure formation of the Universe is one of the holy grails of modern astrophysics. Astrophysicists survey large volumes of the Universe and employ a large ensemble of computer simulations to compare with the observed data in order to extract the full information of our own Universe. However, to evolve billions of particles over billions of years even with the simplest physics is a daunting task. We build a deep neural network, the Deep Density Displacement Model (hereafter D^{3}M), which learns from a set of pre-run numerical simulations, to predict the non-linear large scale structure of the Universe with Zel'dovich Approximation (hereafter ZA), an analytical approximation based on perturbation theory, as the input. Our extensive analysis, demonstrates that D^{3}M outperforms the second order perturbation theory (hereafter 2LPT), the commonly used fast approximate simulation method, in predicting cosmic structure in the non-linear regime. We also show that D^{3}M is able to accurately extrapolate far beyond its training data, and predict structure formation for significantly different cosmological parameters. Our study proves that deep learning is a practical and accurate alternative to approximate 3D simulations of the gravitational structure formation of the Universe.

cosmology | deep learning | simulation

Astrophysicists require a large amount of simulations to extract the information from observations (1–8). At its core, modeling structure formation of the Universe is a computationally challenging task; it involves evolving billions of particles with the correct physical model over a large volume over billions of years (9–11). To simplify this task, we either simulate a large volume with simpler physics or a smaller volume with more complex physics. In order to produce the cosmic web (12) in large volume, we select gravity, the most important component of the theory, to simulate at large scales. A gravity-only N-body simulation is the most popular; and effective numerical method to predict the full 6D phase space distribution of a large number of massive particles whose position and velocity evolve over time in the Universe (13). Nonetheless, N-body simulations are relatively computationally expensive, thus making the comparison of the N-body simulated large-scale structure (of different underlying cosmological parameters) with the observed Universe a challenging task. We propose to use a deep model that predicts the structure formation as an alternative to N-body simulations.

Deep learning (14) is a fast growing branch of machine learning where recent advances have lead to models that reach and sometimes exceed human performance across diverse areas, from analysis and synthesis of images (15–17), sound (18, 19), text (20, 21) and videos (22, 23) to complex control and planning tasks as they appear in robotics and game-play (24–26). This new paradigm is also significantly impacting a variety of domains in the sciences, from biology (27, 28) to chemistry (29, 30) and physics (31, 32). In particular, in astronomy and cosmology, a growing number of recent studies are using deep learning for a variety of tasks, ranging from analysis of cosmic microwave background (33–35), large-scale structure (36, 37), and gravitational lensing effects (38, 39) to classification of different light sources (40–42).

The ability of these models to learn complex functions has motivated many to use them to understand the physics of interacting objects leveraging image, video and relational data (43–53). However, modeling the dynamics of billions of particles in N-body simulations poses a distinct challenge.

In this paper we show that a variation on the architecture of a well-known deep learning model (54), can efficiently transform the first order approximations of the displacement field and approximate the exact solutions, thereby producing accurate estimates of the large-scale structure. Our key

Significance Statement

To understand the evolution of the Universe requires a concerted effort of accurate observation of the sky and fast prediction of structures in the Universe. N-body simulation is an effective approach to predicting structure formation of the Universe, though computationally expensive. Here we build a deep neural network to predict structure formation of the Universe. It outperforms the traditional fast analytical approximation, and accurately extrapolates far beyond its training data. Our study proves that deep learning is an accurate alternative to traditional way of generating approximate cosmological simulations. Our study also used deep learning to generate complex 3D simulations in cosmology. This suggests deep learning can provide a powerful alternative to traditional numerical simulations in cosmology.

S. He, Y.L., S. Ho, and B.P. designed research; S. He, Y.L., Y.F., and S. Ho performed research; S. He, Y.L., S.R., and B.P. contributed new reagents/analytic tools; S. He, Y.L., Y.F., and W.C. analyzed data; and S. He, Y.L., Y.F., S. Ho, S.R., and W.C. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The source code of our implementation is available at https://github.com/siyucosmo/ML-Recon. The code to generate the training data is available at https://github.com/rainwoodman/fastpm.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1821458116/-/DCSupplemental.

[1]To whom correspondence may be addressed. Email: shirleyho@flatironinstitute.org or siyuh@andrew.cmu.edu.

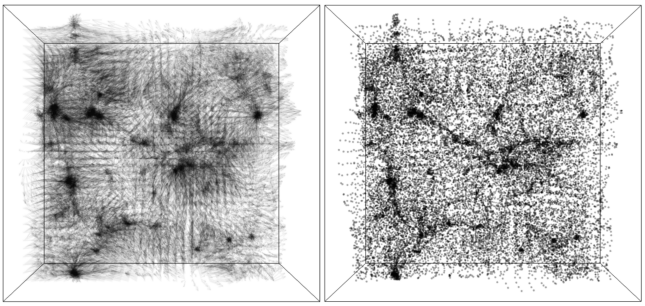

Fig. 1. (left) The displacement vector-field and (right) the resulting density field produced by D^3M. The vectors in the left figure are uniformly scaled down for better visualization.

objective is to prove that this approach is an accurate and computationally efficient alternative to expensive cosmological simulations, and to this end we provide an extensive analysis of the results in the following section.

The outcome of a typical N-body simulation depends on both the initial conditions and on cosmological parameters which affect the evolution equations. A striking discovery is that D^3M, trained using a single set of cosmological parameters generalizes to new sets of significantly different parameters, minimizing the need for training data on diverse range of cosmological parameters.

Setup

We build a deep neural network, D^3M, with similar input and output of an N-body simulation. The input to our D^3M is the displacement field from ZA (55). A displacement vector is the difference of a particle position at target redshift $z = 0$, i.e., the present time, and its Lagrangian position on a uniform grid. ZA evolves the particles on linear trajectories along their initial displacements. It is accurate when the displacement is small, therefore ZA is frequently used to construct the initial conditions of N-body simulations (56). As for the ground truth, the target displacement field is produced using FastPM (57), a recent approximate N-body simulation scheme that is based on a particle-mesh (PM) solver. FastPM quickly approaches a full N-body simulation with high accuracy and provides a viable alternative to direct N-body simulations for the purpose of our study.

A significantly faster approximation of N-body simulations is produced by second-order Lagrangian perturbation theory (hereafter 2LPT), which bends each particle's trajectory with a quadratic correction (58). 2LPT is used in many cosmological analyses to generate a large number of cosmological simulations for comparison of astronomical dataset against the physical model (59, 60) or to compute the covariance of the dataset (61–63). We regard 2LPT as an effective way to efficiently generate a relatively accurate description of the large-scale structure and therefore we select 2LPT as the reference model for comparison with D^3M.

We generate 10,000 pairs of ZA approximations as input and accurate FastPM approximations as target. We use simulations of 32^3 N-body particles in a volume of $128\,h^{-1}$Mpc (600 million light years, where $h = 0.7$ is the Hubble parameter). The particles have a mean separation of $4\,h^{-1}$Mpc per dimension.

An important choice in our approach is training with displacement field rather than density field. Displacement field $\mathbf{\Psi}$ and density field ρ are two ways of describing the same distribution of particles. And an equivalent way to describe density field is the over-density field, defined as $\delta = \rho/\bar{\rho} - 1$, with $\bar{\rho}$ denoting the mean density. The displacement field and over-density field are related by eq. 1.

$$\begin{aligned} \mathbf{x} &= \mathbf{\Psi}(\mathbf{q}) + \mathbf{q} \\ \delta(\mathbf{x}) &= \int d^3q \delta_D(\mathbf{x} - \mathbf{q} - \mathbf{\Psi}(\mathbf{q})) - 1 \end{aligned} \quad [1]$$

When the displacement field is small and has zero curl, the choice of over-density vs displacement field for the output of the model is irrelevant, as there is a bijective map between these two representations, described by the equation:

$$\mathbf{\Psi} = \int \frac{d^3k}{(2\pi)^3} e^{i\mathbf{k}\cdot\mathbf{q}} \frac{i\mathbf{k}}{k^2} \delta(\mathbf{k}) \quad [2]$$

However as the displacements grow into the non-linear regime of structure formation, different displacement fields can produce identical density fields (e.g. 64). Therefore, providing the model with the target displacement field during the training eliminates the ambiguity associated with the density field. Our inability to produce comparable results when using the density field as our input and target attests that relevant information resides in the displacement field (See SI Appendix, Fig. S1) .

Results and Analysis

Figure 1 shows the displacement vector field as predicted by D^3M (left) and the associated point-cloud representation of the structure formation (right). It is possible to identify structures such as clusters, filaments and voids in this point-cloud representation. We proceed to compare the accuracy of D^3M and 2LPT compared with ground truth.

Point-Wise Comparison. Let $\mathbf{\Psi} \in \mathbb{R}^{d\times d\times d\times 3}$ denote the displacement field, where d is the number of spatial resolution elements in each dimension ($d = 32$). A natural measure of error is the relative error $|\hat{\mathbf{\Psi}} - \mathbf{\Psi}_t|/|\mathbf{\Psi}_t|$, where $\mathbf{\Psi}_t$ is the true displacement field (FastPM) and $\hat{\mathbf{\Psi}}$ is the prediction from 2LPT or D^3M. Figure 2 compares this error for different approximations in a 2-D slice of a single simulation. We observe that D^3M predictions are very close to the ground truth, with a maximum relative error of 1.10 over all 1000 simulations. For 2LPT this number is significantly higher at 4.23. In average, the result of D^3M comes with a 2.8% relative error while for 2LPT it equals 9.3%.

2-Point Correlation Comparison. As suggested by Figure 2 the denser regions seem to have a higher error for all methods – that is, more non-linearity in structure formation creates larger errors for both D^3M and 2LPT. The dependence of error on scale is computed with 2-point and 3-point correlation analysis.

Cosmologists often employ compressed summary statistics of the density field in their studies. The most widely used of these statistics are the 2-point correlation function (2PCF) $\xi(r)$ and its Fourier transform, the power spectrum $P_{\delta\delta}(k)$:

$$\begin{aligned} \xi(|\boldsymbol{r}|) &= \langle \delta_A(\boldsymbol{r}')\delta_B(\boldsymbol{r}' + \boldsymbol{r})\rangle, \\ P_{\delta\delta}(|\boldsymbol{k}|) &= \int \mathrm{d}^3\boldsymbol{r}\; \xi(r) e^{i\boldsymbol{k}\cdot\boldsymbol{r}}, \end{aligned} \quad [3]$$

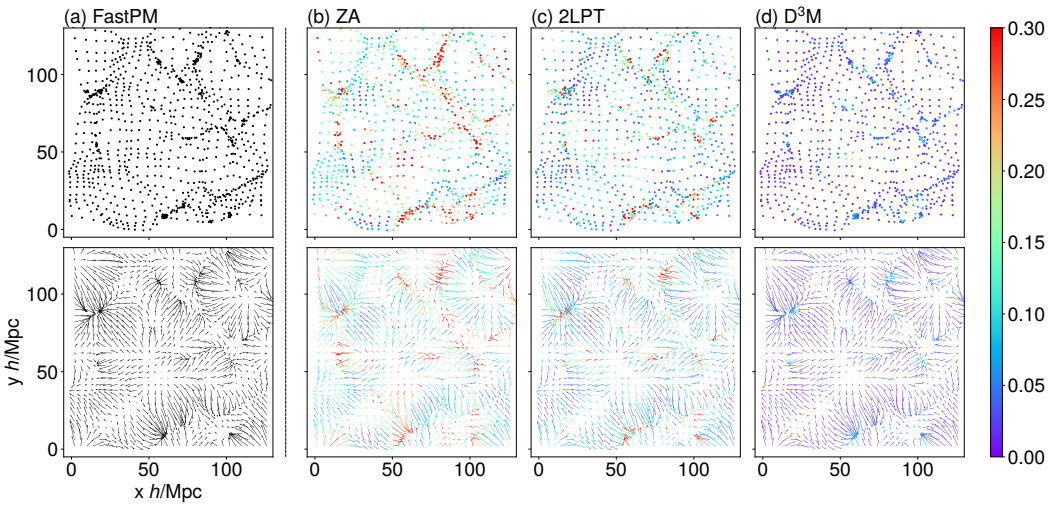

Fig. 2. The columns show 2-D slices of full particle distribution (top) and displacement vector (bottom) by various models, from left to right:
(a) FastPM: the target ground truth, a recent approximate N-body simulation scheme that is based on a particle-mesh (PM) solver ;
(b) Zel'dovich approximation (ZA): a simple linear model that evolves particle along the initial velocity vector;
(c) Second order Lagrangian perturbation theory (2LPT): a commonly used analytical approximatation;
(d) Deep learning model (D^3M) as presented in this work.
While FastPM (A) served as our ground truth, B–D include color for the points or vectors. The color indicates the relative difference $(q_{model} - q_{target})/q_{target}$ between the target (a) location or displacement vector and predicted distributions by various methods (b-d). The error-bar shows denser regions have a higher error for all methods. which suggests that it is harder to predict highly non-linear region correctly for all models: D^3M, 2LPT and ZA. Our model D^3M has smallest differences between predictions and ground truth among the above models (b)-(d).

where the ensemble average $\langle\,\rangle$ is taken over all possible realizations of the Universe. Our Universe is observed to be both homogeneous and isotropic on large scales, i.e. without any special location or direction. This allows one to drop the dependencies on $\boldsymbol{r}'$ and on the direction of $\boldsymbol{r}$, leaving only the amplitude $|\boldsymbol{r}|$ in the final definition of $\xi(r)$. In the second equation, $P_{\delta\delta}(k)$ is simply the Fourier transform of $\xi(r)$, and captures the dispersion of the plane wave amplitudes at different scales in the Fourier space. $\boldsymbol{k}$ is the 3D wavevector of the plane wave, and its amplitude k (the wavenumber) is related to the wavelength λ by $k = 2\pi/\lambda$. Due to isotropy of the Universe, we drop the vector form of $\boldsymbol{r}$ and $\boldsymbol{k}$.

Because FastPM, 2LPT and D^3M take the displacement field as input and output, we also study the two-point statistics for the displacement field. The displacement power spectrum is defined as:

$$P_{\boldsymbol{\Psi}\boldsymbol{\Psi}}(k) = \langle\Psi_x(k)\Psi_x^*(k)\rangle + \langle\Psi_y(k)\Psi_y^*(k)\rangle + \langle\Psi_z(k)\Psi_z^*(k)\rangle \quad [4]$$

We focus on the Fourier-space representation of the 2-point correlation. Because the matter and the displacement power spectrum take the same form, in what follows we drop the subscript for matter and displacement field and use $P(k)$ to stand for both matter and displacement power spectrum. We employ the transfer function $T(k)$ and the correlation coefficient $r(k)$ as metrics to quantify the model performance against the ground truth (FastPM) in the 2-point correlation. We define the transfer function $T(k)$ as the square root of the ratio of two power spectra,

$$T(k) = \sqrt{\frac{P_{\rm pred}(k)}{P_{\rm true}(k)}}, \quad [5]$$

where $P_{\rm pred}(k)$ is the density or displacement power spectrum as predicted by 2LPT or D^3M, and analogously $P_{\rm true}(k)$ is the ground truth predicted by FastPM. The correlation coefficient r(k) is a form of normalized cross power spectrum,

$$r(k) = \frac{P_{\rm pred\times true}(k)}{\sqrt{P_{\rm pred}(k)P_{\rm true}(k)}}, \quad [6]$$

where $P_{\rm pred\times true}(k)$ is the cross power spectrum between 2LPT or D^3M predictions and the ground truth (FastPM) simulation result. The transfer function captures the discrepancy between amplitudes, while the correlation coefficient can indicate the discrepancy between phases as functions of scales. For a perfectly accurate prediction, $T(k)$ and $r(k)$ are both 1. In particular, $1 - r^2$ describes stochasticity, the fraction of the variance in the prediction that cannot be explained by the true model.

Figures 3(a) shows the average power spectrum, transfer function $T(k)$ and stochasticity $1 - r^2(k)$ of the displacement field and the density field over 1000 simulations. The transfer function of density from 2LPT predictions is 2% smaller than that of FastPM on large scales ($k \approx 0.05\,h\mathrm{Mpc}^{-1}$). This is expected since 2LPT performs accurately on very large scales ($k < 0.01\,h\mathrm{Mpc}^{-1}$). The displacement transfer function of 2LPT increases above 1 at $k \approx 0.35\,h\mathrm{Mpc}^{-1}$ and then drops sharply. The increase of 2LPT displacement transfer function is because 2LPT over-estimates the displacement power at small scales (see, e.g. 65). There is a sharp drop of power near the voxel scale because smoothing over voxel scales in our predictions automatically erases power at scales smaller than the voxel size.

Now we turn to the D^3M predictions: both the density and displacement transfer functions of the D^3M differ from 1 by a mere 0.4% at scale $k \lesssim 0.4\,h\mathrm{Mpc}^{-1}$, and this discrepancy only increases to 2% and 4% for density field and displacement field respectively, as k increases to the Nyquist frequency around $0.7\,h\mathrm{Mpc}^{-1}$. The stochasticity hovers at approximately 10^{-3}

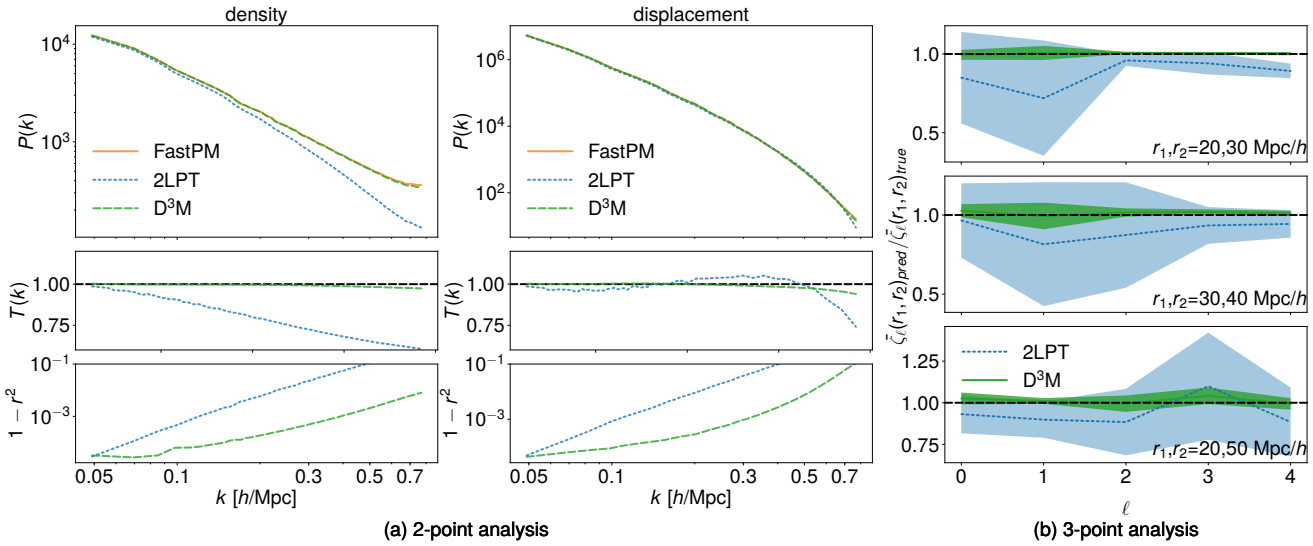

Fig. 3. (**a**) From top to bottom: (top) displacement and density power-spectrum of FastPM (orange), 2LPT (blue), and D^{3}M (green); (middle) transfer function – i.e., the square root of the ratio of the predicted power-spectrum to the ground truth; (bottom) 1-r^2 where r is the correlation coefficient between the predicted fields and the true fields. Results are the averaged values of 1000 test simulations. The transfer function and correlation coefficient of the D^{3}M predictions is nearly perfect from large to intermediate scales and outperforms our benchmark 2LPT significantly.
(**b**) The ratios of the multipole coefficients ($\zeta_l(r_1, r_2)$) (to the target) of the two 3-point correlation functions for several triangle configurations. The results are averaged over 10 test simulations. The error-bars (padded regions) are the standard deviations derived from 10 test simulations. The ratio shows the 3-point correlation function of D^{3}M is closer than 2LPT to our target FastPM with lower variance.

and 10^{-2} for most scales. In other words, for both the density and displacement fields the correlation coefficient between the D^{3}M predictions and FastPM simulations, all the way down to small scales of $k = 0.7\,h\mathrm{Mpc}^{-1}$ is greater than 90%. The transfer function and correlation coefficient of the D^{3}M predictions shows that it can reproduce the structure formation of the Universe from large to semi-non-linear scales. D^{3}M significantly outperforms our benchmark model 2LPT in the 2 point function analysis. D^{3}M only starts to deviate from the ground truth at fairly small scales. This is not surprising as the deeply nonlinear evolution at these scales is more difficult to simulate accurately and appears to be intractable by current analytical theories(66).

3-Point Correlation Comparison. The 3-point correlation function (3PCF) expresses the correlation of the field of interest among 3 locations in the configuration space, which is equivalently defined as bispectrum in Fourier space. Here we concentrate on the 3PCF for computational convenience:

$$\zeta(r_1, r_2, \theta) = \langle \delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{r_1})\delta(\boldsymbol{x}+\boldsymbol{r_2})\rangle. \quad [7]$$

where r_1=$|\boldsymbol{r_1}|$ and r_2=$|\boldsymbol{r_2}|$. Translation invariance guarantees that ζ is independent of $\boldsymbol{x}$. Rotational symmetry further eliminates all direction dependence except dependence on θ, the angle between $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$. The multipole moments of $\zeta(r_1, r_2, \theta)$, $\zeta_\ell(r_1, r_2) = (2\ell+1)\int d\theta P_\ell(\cos\theta)\zeta(r_1, r_2, \theta)$ where $P_\ell(\cos\theta)$ is the Legendre polynomial of degree ℓ, can be efficiently estimated with pair counting (67). While the input (computed by ZA) do not contain significant correlations beyond the second order (power spectrum level), we expect D^{3}M to generate densities with a 3PCF that mimics that of ground truth.

We compare the 3PCF calculated from FastPM, 2LPT and D^{3}M by analyzing the 3PCF through its multipole moments $\zeta_\ell(r_1, r_2)$. Figure 3(b) shows the ratio of the binned multipole coefficients of the two 3PCF for several triangle configurations, $\bar{\xi}_\ell(r_1, r_2)_{pred}/\bar{\xi}_\ell(r_1, r_2)_{true}$, where $\bar{\xi}_\ell(r_1, r_2)_{pred}$ can be the 3PCF for D^{3}M or 2LPT and $\bar{\xi}_\ell(r_1, r_2)_{true}$ is the 3PCF for FastPM. We used 10 radial bins with $\Delta r = 5\,h^{-1}$Mpc. The results are averaged over 10 test simulations and the errorbars are the standard deviation. The ratio shows the 3PCF of D^{3}M is more close to FastPM than 2LPT with smaller errorbars. To further quantify our comparison, we calculate the relative 3PCF residual defined by

$$\text{3PCF relative residual} = \frac{1}{9\times N_r}\sum_{\ell=0}^{8}\sum_{r_1,r_2}\frac{|\zeta_\ell(r_1,r_2)_{\rm pred} - \zeta_\ell(r_1,r_2)_{\rm true}|}{|\zeta_\ell(r_1,r_2)_{\rm true}|} \quad [8]$$

where N_r is the number of (r_1,r_2) bins. The mean relative 3PCF residual of the D^{3}M and 2LPT predictions compared to FastPM are 0.79% and 7.82% respectively. The D^{3}M accuracy on 3PCF is also an order of magnitude better than 2LPT, which indicates that the D^{3}M is far better at capturing the non-Gaussian structure formation.

Generalizing to New Cosmological Parameters

So far, we train our model using a "single" choice of cosmological parameters $A_s = 2.142\times10^{-9}$ (hereafter $A_0 = 2.142\times10^{-9}$) and $\Omega_m = 0.3089$ (68). A_s is the primordial amplitude of the scalar perturbation from cosmic inflation, and Ω_m is the fraction of the total energy density that is matter at the present time, and we will call it matter density parameter for short. The true exact value of these parameters are unknown and different choices of these parameters change the large-scale structure of the Universe; see Figure 4.

Here, we report an interesting observation: the D^{3}M trained on a single set of parameters in conjunction with ZA (which depends on A_s and Ω_m) as input, can predict the structure

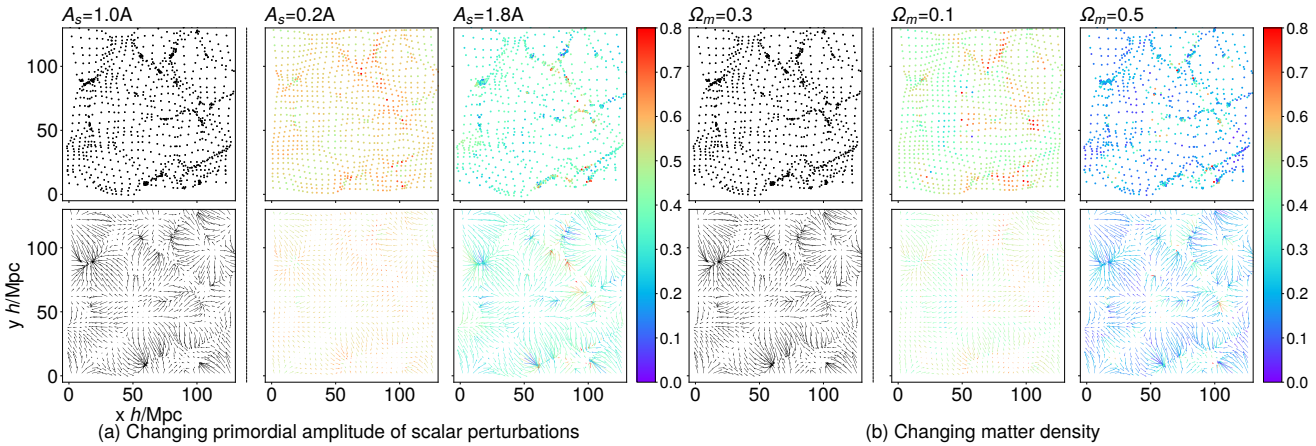

Fig. 4. We show the differences of particle distributions and displacement fields when we change the cosmological parameters A_s and Ω_m.
(**a**) The errorbar shows the difference of particle distribution (upper panel) and displacement fields (lower panel) between $A_s = A_0$ and the two extremes for $A_s = .2A_0$(Center) and $A_s = 1.8A_0$ (Right).
(**b**) A similar comparison showing the difference of the particle distributions (upper panel) and displacement fields (lower panel) for smaller and larger values of $\Omega_m \in \{.1, .5\}$ with regard to $\Omega_m = 0.3089$, which was used for training.
While the difference for smaller value of A_s (Ω_m) is larger, the displacement for larger A_s (Ω_m) is more non-linear. This non-linearity is due to concentration of mass and makes the prediction more difficult.

formation for widely different choices of A_s and Ω_m. From a computational point of view this suggests a possibility of producing simulations for a diverse range of parameters, with minimal training data.

Varying Primordial Amplitude of Scalar Perturbations A_s. After training the D^{3}M using $A_s = A_0$, we change A_s in the input of our test set by nearly one order of magnitude: $A_s = 1.8A_0$ and $A_s = 0.2A_0$. Again, we use 1000 simulations for analysis of each test case. The average relative displacement error of D^{3}M remains less than 4% per voxel (compared to $< 3\%$ when train and test data have the same parameters). This is still well below the error for 2LPT, which has relative errors of 15.5% and 6.3% for larger and smaller values of A_s respectively.

Figure 5(a) shows the transfer function and correlation coefficient for both D^{3}M and 2LPT. The D^{3}M performs much better than 2LPT for $A_s = 1.8A_0$. For small $A_s = 0.2A_0$, 2LPT does a better job than D^{3}M predicting the density transfer function and correlation coefficient at the largest scales, otherwise D^{3}M predictions are more accurate than 2LPT at scales larger than $k = 0.08\,h\text{Mpc}^{-1}$. We observe a similar trend with 3PCF analysis: the 3PCF of D^{3}M predictions are notably better than 2LPT ones for larger A_s, compared to smaller A_s where it is only slightly better. These results confirm our expectation that increasing A_s increases the non-linearity of the structure formation process. While 2LPT can predict fairly well in linear regimes, compared to D^{3}M its performance deteriorates with increased non-linearity. It is interesting to note that D^{3}M prediction maintains its advantage despite being trained on data from more linear regimes.

Varying matter density parameter Ω_m. We repeat the same experiments, this time changing Ω_m to 0.5 and 0.1, while the model is trained on $\Omega_m = 0.3089$, which is quite far from both of the test sets. For $\Omega_m = 0.5$ the relative residual displacement errors of the D^{3}M and 2LPT averaged over 1000 simulations are 3.8% and 15.2% and for $\Omega_m = 0.1$ they are 2.5% and 4.3%. Figures 5(c)(d) show the two-point statistics for density field predicted using different values of Ω_m. For $\Omega_m = 0.5$, the results show that the D^{3}M outperforms 2LPT at all scales, while for smaller $\Omega_m = 0.1$, D^{3}M outperforms 2LPT on smaller scales ($k > 0.1\,h\text{Mpc}^{-1}$). As for the 3PCF of simulations with different values of Ω_m, the mean relative 3PCF residual of the D^{3}M for $\Omega_m = 0.5$ and $\Omega_m = 0.1$ are 1.7% and 1.2% respectively and for 2LPT they are 7.6% and 1.7% respectively. The D^{3}M prediction performs better at $\Omega_m = 0.5$ than $\Omega_m = 0.1$. This is again because the Universe is much more non-linear at $\Omega_m = 0.5$ than $\Omega_m = 0.1$. The D^{3}M learns more non-linearity than is encoded in the formalism of 2LPT.

type	point-wise	$T(k)$ $k = 0.11$ †	$r(k)$ $k = 0.11$	$T(k)$ $k = 0.50$	$r(k)$ $k = 0.50$	3PCF
test phase						
2 LPT Density	N/A	0.96	1.00	0.74	0.94	0.0782
D^{3}M Density	N/A	1.00	1.00	0.99	1.00	0.0079
2 LPT Displacement	0.093	0.96	1.00	1.04	0.90	N/A
D^{3}M Displacement	0.028	1.00	1.00	0.99	1.00	N/A
$\boldsymbol{A_s = 1.8A_0}$						
2LPT Density	N/A	0.93	1.00	0.49	0.78	0.243
D^{3}M Density	N/A	1.00	1.00	0.98	1.00	0.039
2LPT Displacement	0.155	0.97	1.00	1.07	0.73	N/A
D^{3}M Displacement	0.039	1.00	1.00	0.97	0.99	N/A
$\boldsymbol{A_s = 0.2A_0}$						
2LPT Density	N/A	0.99	1.00	0.98	0.99	0.024
D^{3}M Density	N/A	1.00	1.00	1.03	1.00	0.022
2LPT Displacement	0.063	0.99	1.00	0.95	0.98	N/A
D^{3}M Displacement	0.036	1.00	1.00	1.01	1.00	N/A
$\boldsymbol{\Omega_m = 0.5}$						
2LPT Density	N/A	0.94	1.00	0.58	0.87	0.076
D^{3}M Density	N/A	1.00	1.00	1.00	1.00	0.017
2LPT Displacement	0.152	0.97	1.00	1.10	0.80	N/A
D^{3}M Displacement	0.038	1.00	1.00	0.98	0.99	N/A
$\boldsymbol{\Omega_m = 0.1}$						
2LPT Density	N/A	0.97	1.00	0.96	0.99	0.017
D^{3}M Density	N/A	0.99	1.00	1.04	1.00	0.012
2LPT Displacement	0.043	0.97	1.00	0.97	0.98	N/A
D^{3}M Displacement	0.025	0.99	1.00	1.02	1.00	N/A

†The unit of k is $h\text{Mpc}^{-1}$. N/A, not applicable

Table 1. A summary of our analysis.

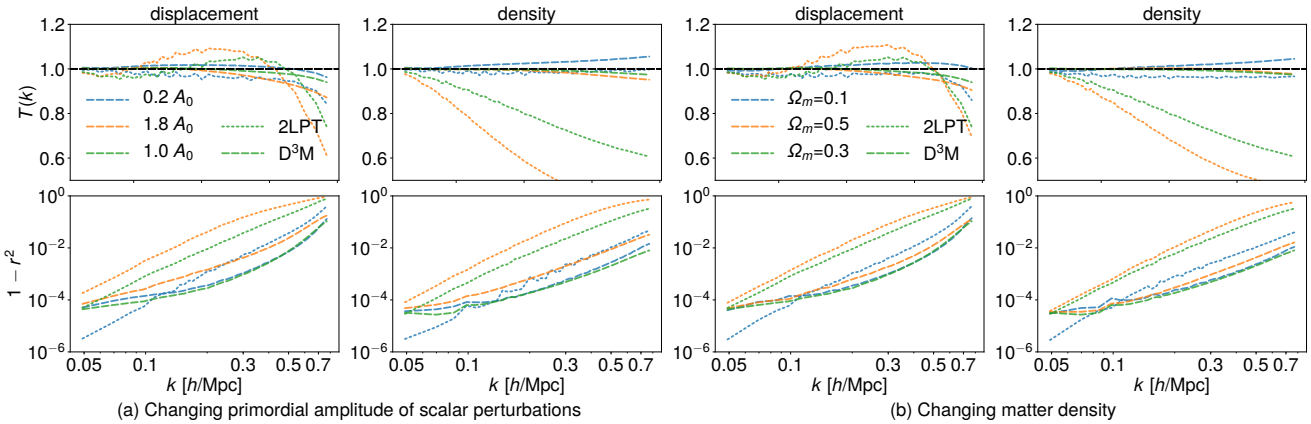

Fig. 5. Similar plots as in Figure. 3(a), except we test the 2 point statistics when we vary the cosmological parameters without changing the training set (which has differnet cosmological parameters) nor the trained model. We show predictions from D^{3}M and 2LPT when tested on different (**a**) A_s and, (**b**) Ω_m. We show (top) the transfer function – i.e., the square root of the ratio of the predicted power-spectrum to the ground truth and (bottom) 1-r^2 where r is the correlation coefficient between the predicted fields and the true fields. D^{3}M prediction outperforms 2LPT prediction at all scales except in the largest scales as the perturbation theory works well in linear regime (large scales).

Conclusions

To summarize, our deep model D^{3}M can accurately predict the large-scale structure of the Universe as represented by FastPM simulations, at all scales as seen in the summary table in Table. 1. Furthermore, D^{3}M learns to predict cosmic structure in the non-linear regime more accurately than our benchmark model 2LPT. Finally, our model generalizes well to test simulations with cosmological parameters (A_s and Ω_m) significantly different from the training set. This suggests that our deep learning model can potentially be deployed for a ranges of simulations beyond the parameter space covered by the training data (Table 1). Our results demonstrate that the D^{3}M successfully learns the nonlinear mapping from first order perturbation theory to FastPM simulation beyond what higher order perturbation theories currently achieve.

Looking forward, we expect replacing FastPM with exact N-body simulations would improve the performance of our method. As the complexity of our D^{3}M model is linear in the number of voxels, we expect to be able to further improve our results if we replace the FastPM simulations with higher resolution simulations. Our work suggests that deep learning is a practical and accurate alternative to the traditional way of generating approximate simulations of the structure formation of the Universe.

Materials and Methods

Dataset. The full simulation data consists of 10,000 simulations of boxes with ZA and FastPM as input-output pairs, with an effective volume of 20 $(\mathrm{Gpc/h})^3$ $(190 \times 10^9 \mathrm{ly}^3)$, comparable to the volume of a large spectroscopic sky survey like Dark Energy Spectroscopic Instrument or EUCLID. We split the full simulation data set into 80%, 10% and 10% for training, validation and test, respectively. We also generated 1000 simulations for 2LPT for each set of tested cosmological parameters.

Model and Training. The D^{3}M adopts the U-Net architecture (54) with 15 convolution or deconvolution layers and approximately 8.4×10^6 trainable parameters. Our D^{3}M generalizes the standard U-Net architecture to work with three-dimensional data (69–71). The details of the architecture are described in the following sections and a schematic figure of the architecture is shown in SI Appendix, Figure. S2. In the training phase, we employ the Adam Optimizer (72) with a learning rate of 0.0001, and first and second moment exponential decay rates equal to 0.9 and 0.999, respectively. We use the Mean-Squared Error as the loss function (See Loss Function) and $L2$ regularization with regularization coefficient 0.0001.

Details of the* D^{3}M *Architecture. The contracting path follows the typical architecture of a convolution network. It consists of two blocks, each of which consists of two successive convolutions of stride 1 and a down-sampling convolution with stride 2. The convolution layers use 3×3×3 filters with a periodic padding of size 1 (see Padding and Periodic Boundary) on both sides of each dimension. Notice that at each of the two down-sampling steps, we double the number of feature channels. At the bottom of the D^{3}M, another two successive convolutions with stride 1 and the same periodic padding as above are applied. The expansive path of our D^{3}M is an inverted version of the contracting path of the network. (It includes two repeated applications of the expansion block, each of which consists of one up-sampling transposed convolution with stride 1/2 and two successive convolution of stride 1. The transposed convolution and the convolution are constructed with 3×3×3 filters.)

We take special care in the padding and cropping procedure to preserve the shifting and rotation symmetry in the up-sampling layer in expansive path. Before the transposed convolution we apply a periodic padding of length 1 on the right, down and back sides of the box (`padding=(0,1,0,1,0,1)` in pytorch), and after the transposed convolution, we discard one column on the left, up and front sides of the box and two columns on the right, down and back sides (`crop=(1,2,1,2,1,2)`).

A special feature of the D^{3}M is the concatenation procedure, where the up-sampling layer halves the feature channels and then concatenates them with the corresponding feature channels on the contracting path, doubling the number of feature channels.

The expansive building block then follows a 1×1×1 convolution without padding, which converts the 64 features to

the the final 3-D displacement field. All convolutions in the network except the last one are followed by a rectified linear unit activation and batch normalization (BN).

Padding and Periodic Boundary. It is common to use constant or reflective padding in deep models for image processing. However, these approaches are not suitable for our setting. The physical model we are learning is constructed on a spatial volume with a periodic boundary condition. This is sometimes also referred to as a torus geometry, where the boundaries of the simulation box are topologically connected – that is $x_{i+L} = x_i$ where i is the index of the spatial location, and L is the periodicity (size of box). Constant or reflective padding strategies break the connection between the physically nearby points separated across the box, which not only loses information but also introduces noise during the convolution, further aggravated with an increased number of layers.

We find that the periodic padding strategy significantly improves the performance and expedites the convergence of our model, comparing to the same network using a constant padding strategy. This is not surprising, as one expects it is easier to train a model that can explain the data than to train a model that does not.

Loss Function. We train the $\mathrm{D^3M}$ to minimize the mean square error on particle displacements

$$\mathcal{L} = \frac{1}{N}\sum_i (\hat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_{t,i})^2, \qquad [9]$$

where i labels the particles and the N is the total number of particles. This loss function is proportional to the integrated squared error, and using a Fourier transform and Parseval's theorem it can be rewritten as

$$\int (\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_t)^2 \mathrm{d}^3\boldsymbol{q} = \int \left|\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_t\right|^2 \mathrm{d}^3\boldsymbol{k} =$$
$$\int \mathrm{d}^3\boldsymbol{k} \left(\left|\boldsymbol{\Psi}_t\right|^2 (1-T)^2 + 2\left|\hat{\boldsymbol{\Psi}}\right|\left|\boldsymbol{\Psi}_t\right|(1-r) \right) \qquad [10]$$

where $\boldsymbol{q}$ is the Lagrangian space position, and $\boldsymbol{k}$ its corresponding wavevector. T is the transfer function defined in Eq. 5, and r is the correlation coefficient defined in Eq. 6, which characterize the similarity between the predicted and the true fields, in amplitude and phase respectively. Eq. 10 shows that our simple loss function jointly captures both of these measures: as T and r approach 1, the loss function approaches 0.

Data availability. The source code of our implementation is available at https://github.com/siyucosmo/ML-Recon. The code to generate the training data is also available at https://github.com/rainwoodman/fastpm.

ACKNOWLEDGMENTS. We thank Angus Beane, Peter Braam, Gabriella Contardo, David Hogg, Laurence Levasseur, Pascal Ripoche, Zack Slepian and David Spergel for useful suggestions and comments, Angus Beane for comments on the paper, Nick Carriero for help on Center for Computational Astrophysics (CCA) computing clusters. The work is funded partially by Simons Foundation. The FastPM simulations are generated on the computer cluster Edison at the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. The training of neural network model is performed on the CCA computing facility and the Carnegie Mellon University AutonLab computing facility. The open source software toolkit `nbodykit` (73) is employed for the clustering analysis. YL acknowledges support from the Berkeley Center for Cosmological Physics and the Kavli Institute for the Physics and Mathematics of the Universe, established by World Premier International Research Center Initiative (WPI) of the MEXT, Japan. S.Ho thanks NASA for their support in grant number: NASA grant 15-WFIRST15-0008 and NASA Research Opportunities in Space and Earth Sciences grant 12-EUCLID12-0004, and Simons Foundation.

References

1. Colless, M., , et al. (2001) The 2dF Galaxy Redshift Survey: Spectra and redshifts. *Mon. Not. Roy. Astron. Soc.* 328:1039.
2. Eisenstein, D.J., , et al. (2011) SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems. *Astron. J.* 142:72.
3. Jones, H.D., , et al. (2009) The 6dF Galaxy Survey: Final Redshift Release (DR3) and Southern Large-Scale Structures. *Mon. Not. Roy. Astron. Soc.* 399:683.
4. Liske, J., , et al. (2015) Galaxy and mass assembly (gama): end of survey report and data release 2. *Monthly Notices of the Royal Astronomical Society* 452(2):2087.
5. Scodeggio, M., , et al. (2016) The VIMOS Public Extragalactic Redshift Survey (VIPERS). Full spectroscopic data and auxiliary information release (PDR-2). *ArXiv e-prints*.
6. Ivezić Ž, et al. (2008) LSST: from Science Drivers to Reference Design and Anticipated Data Products. *ArXiv e-prints*.
7. Amendola L, et al. (2018) Cosmology and fundamental physics with the Euclid satellite. *Living Reviews in Relativity* 21:2.
8. Spergel D, et al. (2015) Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report. *ArXiv e-prints*.
9. MacFarland T, Couchman HMP, Pearce FR, Pichlmeier J (1998) A new parallel $\mathrm{P^3M}$ code for very large-scale cosmological simulations. *New Astronomy* 3(8):687–705.
10. Springel V, Yoshida N, White SDM (2001) GADGET: a code for collisionless and gasdynamical cosmological simulations. *New Astronomy* 6(2):79–117.
11. Bagla JS (2002) TreePM: A Code for Cosmological N-Body Simulations. *Journal of Astrophysics and Astronomy* 23:185–196.
12. Bond JR, Kofman L, Pogosyan D (1996) How filaments of galaxies are woven into the cosmic web. Nature 380:603–606.
13. Davis M, Efstathiou G, Frenk CS, White SDM (1985) The evolution of large-scale structure in a universe dominated by cold dark matter. ApJ 292:371–394.
14. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *nature* 521(7553):436.
15. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. in *CVPR*. Vol. 1, p. 3.
16. Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.
17. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC (2017) Improved training of wasserstein gans in *Advances in Neural Information Processing Systems*. pp. 5767–5777.
18. Van Den Oord A, et al. (2016) Wavenet: A generative model for raw audio. *CoRR abs/1609.03499*.
19. Amodei D, et al. (2016) Deep speech 2: End-to-end speech recognition in english and mandarin in *International Conference on Machine Learning*. pp. 173–182.
20. Hu Z, Yang Z, Liang X, Salakhutdinov R, Xing EP (2017) Toward controlled generation of text. *arXiv preprint arXiv:1703.00955*.
21. Vaswani A, et al. (2017) Attention is all you need in *Advances in Neural Information Processing Systems*. pp. 5998–6008.
22. Denton E, Fergus R (2018) Stochastic video generation with a learned prior. *arXiv preprint arXiv:1802.07687*.
23. Donahue J, et al. (2015) Long-term recurrent convolutional networks for visual recognition and description in *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2625–2634.
24. Silver D, et al. (2016) Mastering the game of go with deep neural networks and tree search. *nature* 529(7587):484.
25. Mnih V, et al. (2015) Human-level control through deep reinforcement learning. *Nature* 518(7540):529.
26. Levine S, Finn C, Darrell T, Abbeel P (2016) End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research* 17(1):1334–1373.
27. Ching T, et al. (2018) Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface* 15(141):20170387.
28. Alipanahi B, Delong A, Weirauch MT, Frey BJ (2015) Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology* 33(8):831.
29. Segler MH, Preuss M, Waller MP (2018) Planning chemical syntheses with deep neural networks and symbolic ai. *Nature* 555(7698):604.
30. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE (2017) Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212*.
31. Carleo G, Troyer M (2017) Solving the quantum many-body problem with artificial neural networks. *Science* 355(6325):602–606.
32. Adam-Bourdarios C, et al. (2015) The higgs boson machine learning challenge in *NIPS 2014 Workshop on High-energy Physics and Machine Learning*. pp. 19–55.
33. He S, Ravanbakhsh S, Ho S (2018) Analysis of Cosmic Microwave Background with Deep Learning. *International Conference on Learning Representations Workshop*.
34. Perraudin N, Defferrard M, Kacprzak T, Sgier R (2018) Deepsphere: Efficient spherical convolutional neural network with healpix sampling for cosmological applications. *arXiv preprint arXiv:1810.12186*.

35. Caldeira J, et al. (2018) Deepcmb: Lensing reconstruction of the cosmic microwave background with deep neural networks. *arXiv preprint arXiv:1810.01483*.
36. Ravanbakhsh, S., et al. (2017) Estimating cosmological parameters from the dark matter distribution. *ArXiv e-prints*.
37. Mathuriya A, et al. (2018) Cosmoflow: using deep learning to learn the universe at scale. *arXiv preprint arXiv:1808.04728*.
38. Hezaveh, Y. D., Levasseur, L. P., Marshall, P. J. (2017) Fast automated analysis of strong gravitational lenses with convolutional neural networks. *Nature* 548:555-557.
39. Lanusse, F., et al. (2018) Cmu deeplens: deep learning for automatic image-based galaxy-galaxy strong lens finding. MNRAS pp. 3895–3906.
40. Kennamer N, Kirkby D, Ihler A, Sanchez-Lopez FJ (2018) ContextNet: Deep learning for star galaxy classification in *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research. Vol. 80, pp. 2582–2590.
41. Kim EJ, Brunner RJ (2016) Star-galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society* p. stw2672.
42. Lochner M, McEwen JD, Peiris HV, Lahav O, Winter MK (2016) Photometric supernova classification with machine learning. *The Astrophysical Journal Supplement Series* 225(2):31.
43. Battaglia PW, Hamrick JB, Tenenbaum JB (2013) Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences* p. 201306572.
44. Battaglia P, Pascanu R, Lai M, Rezende DJ, , et al. (2016) Interaction networks for learning about objects, relations and physics in *Advances in neural information processing systems*. pp. 4502–4510.
45. Mottaghi R, Bagherinezhad H, Rastegari M, Farhadi A (2016) Newtonian scene understanding: Unfolding the dynamics of objects in static images in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3521–3529.
46. Chang MB, Ullman T, Torralba A, Tenenbaum JB (2016) A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*.
47. Wu J, Yildirim I, Lim JJ, Freeman B, Tenenbaum J (2015) Galileo: Perceiving physical object properties by integrating a physics engine with deep learning in *Advances in neural information processing systems*. pp. 127–135.
48. Wu J, Lim JJ, Zhang H, Tenenbaum JB, Freeman WT (2016) Physics 101: Learning physical object properties from unlabeled videos. in *BMVC*. Vol. 2, p. 7.
49. Watters N, et al. (2017) Visual interaction networks: Learning a physics simulator from video in *Advances in Neural Information Processing Systems*. pp. 4539–4547.
50. Lerer A, Gross S, Fergus R (2016) Learning physical intuition of block towers by example. *arXiv preprint arXiv:1603.01312*.
51. Agrawal P, Nair AV, Abbeel P, Malik J, Levine S (2016) Learning to poke by poking: Experiential learning of intuitive physics in *Advances in Neural Information Processing Systems*. pp. 5074–5082.
52. Fragkiadaki K, Agrawal P, Levine S, Malik J (2015) Learning visual predictive models of physics for playing billiards. *arXiv preprint arXiv:1511.07404*.
53. Tompson J, Schlachter K, Sprechmann P, Perlin K (2016) Accelerating eulerian fluid simulation with convolutional networks. *arXiv preprint arXiv:1607.03597*.
54. Ronneberger, O.; Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* 9351:234–241.
55. Zel'dovich YB (1970) Gravitational instability: An approximate theory for large density perturbations. A&A 5:84–89.
56. White M (2014) The Zel'dovich approximation. MNRAS 439:3630–3640.
57. Feng Y, Chu MY, Seljak U, McDonald P (2016) FASTPM: a new scheme for fast simulations of dark matter and haloes. MNRAS 463:2273–2286.
58. Buchert T (1994) Lagrangian Theory of Gravitational Instability of Friedman-Lemaitre Cosmologies - a Generic Third-Order Model for Nonlinear Clustering. MNRAS 267:811.
59. Jasche J, Wandelt BD (2013) Bayesian physical reconstruction of initial conditions from large-scale structure surveys. MNRAS 432:894–913.
60. Kitaura FS (2013) The initial conditions of the Universe from constrained simulations. MNRAS 429:L84–L88.
61. Dawson KS, et al. (2013) The Baryon Oscillation Spectroscopic Survey of SDSS-III. AJ 145:10.
62. Dawson KS, et al. (2016) The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data. AJ 151:44.
63. DESI Collaboration, et al. (2016) The DESI Experiment Part I: Science,Targeting, and Survey Design. *ArXiv e-prints*.
64. Feng Y, Seljak U, Zaldarriaga M (2018) Exploring the posterior surface of the large scale structure reconstruction. J. Cosmology Astropart. Phys. 7:043.
65. Chan KC (2014) Helmholtz decomposition of the Lagrangian displacement.
66. Perko A, Senatore L, Jennings E, Wechsler RH (2016) Biased Tracers in Redshift Space in the EFT of Large-Scale Structure. *arXiv e-prints* p. arXiv:1610.09321.
67. Slepian Z, Eisenstein DJ (2015) Computing the three-point correlation function of galaxies in $\mathcal{O}(N^2)$ time.
68. Planck Collaboration, et al. (2016) Planck 2015 results. XIII. Cosmological parameters.
69. Milletari F, Navab N, Ahmadi SA (2016) V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *ArXiv e-prints*.
70. Berger P, Stein G (2019) A volumetric deep Convolutional Neural Network for simulation of mock dark matter halo catalogues. MNRAS 482:2861–2871.
71. Aragon-Calvo MA (2018) Classifying the Large Scale Structure of the Universe with Deep Neural Networks. *ArXiv e-prints*.
72. Kingma D, Ba J (2014) Adam: A method for stochastic optimization. *eprint arXiv: 1412.6980*.
73. Hand N, et al. (2018) nbodykit: an open-source, massively parallel toolkit for large-scale structure. *The Astronomical Journal* 156(4):160.

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 |

In 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.

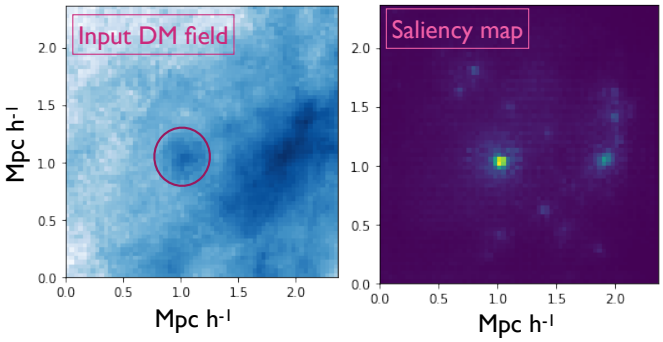

Fig. 1. Left: Dark matter (DM) density field in a particular sub-cube of the TNG100 simulation where we have identified a $10^{12}\,h^{-1}\,M_\odot$ halo (red circle). This field was provided as input to the U-Net trained in W20 to predict the HI field. **Right:** Saliency map where brightness roughly 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 $2.34\,h^{-1}\,\mathrm{Mpc}$. 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 non-trivial 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 ($M_{\rm HI}$). Our goal is in this paper is to model $M_{\rm HI}$ by approximating the function

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

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 $M_{\rm HI}$. 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 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.

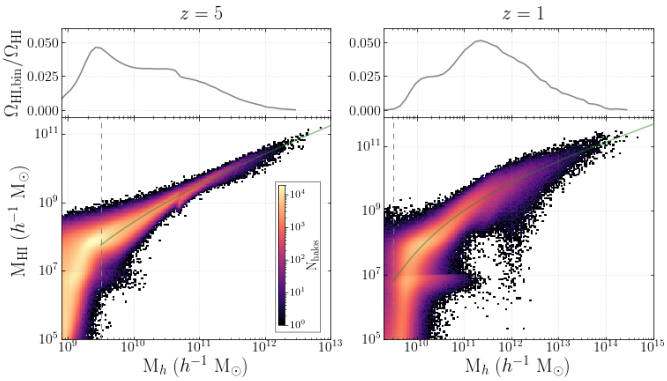

Fig. 2. Bottom: Distribution of HI mass ($M_{\rm HI}$) versus total mass (M_h) for halos in the TNG300-1 simulation at redshifts $z = 5$ (*left*) and $z = 1$ (*right*). Color coding indicates the number of halos in that region of parameter space. The best-fit for the HI–halo mass function (see Eq. 2) 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 $\Omega_{\rm HI}$ of each halo mass bin.

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 $M_{\rm HOD\text{-}HI}$) using the HI−halo mass relation given by:

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

where M_0 is a normalization factor, α is the power-law slope, and $M_{\rm min}$ 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: $M_0 = 0.64\,(0.2) \times 10^{10}\,h^{-1}\,M_\odot$,

[†] https://github.com/MilesCranmer/PySR
[‡] https://www.tng-project.org/data/
[§] HOD in the traditional literature is used for modeling the number of galaxies in a particular halo and we use the term here for modeling the mass of HI in a particular halo.
[¶] It gets harder to retain neutral gas in halos below this mass which can self-shield itself from the ionizing metagalactic radiation.

$M_{\rm min} = 25.41\,(2.36) \times 10^{10}\,h^{-1}\,M_\odot$ 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 $M_{\rm HOD\text{-}HI}$ relation. One can immediately note from Fig. 2 that there is a large scatter in the HI -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 HI field. As we are interested in analyzing large scales in the paper, we have ignored the one-halo term (which account for distribution of HI within the halo) in our model throughout the paper and assumed the entire HI is located at the center of halo; the one halo term only becomes important on scales $k \gtrsim 1\,h\,{\rm Mpc}^{-1}$ that are not relevant for this work.

We emphasize that the best-fit values quoted earlier for $\{M_0, \alpha, M_{\rm min}\}$ 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., $\Omega_{\rm HI}$ 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 HI mass

VN18 showed that halos with the same mass but different HI content cluster differently. This hints at the fact that the HI 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 $M_{\rm HI}$ 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 HI 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 $M_{\rm HI}$ 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 HI 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 $M_R < M_{\rm halo}$ which roughly happens when the $R < R_{\rm virial}$ 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).

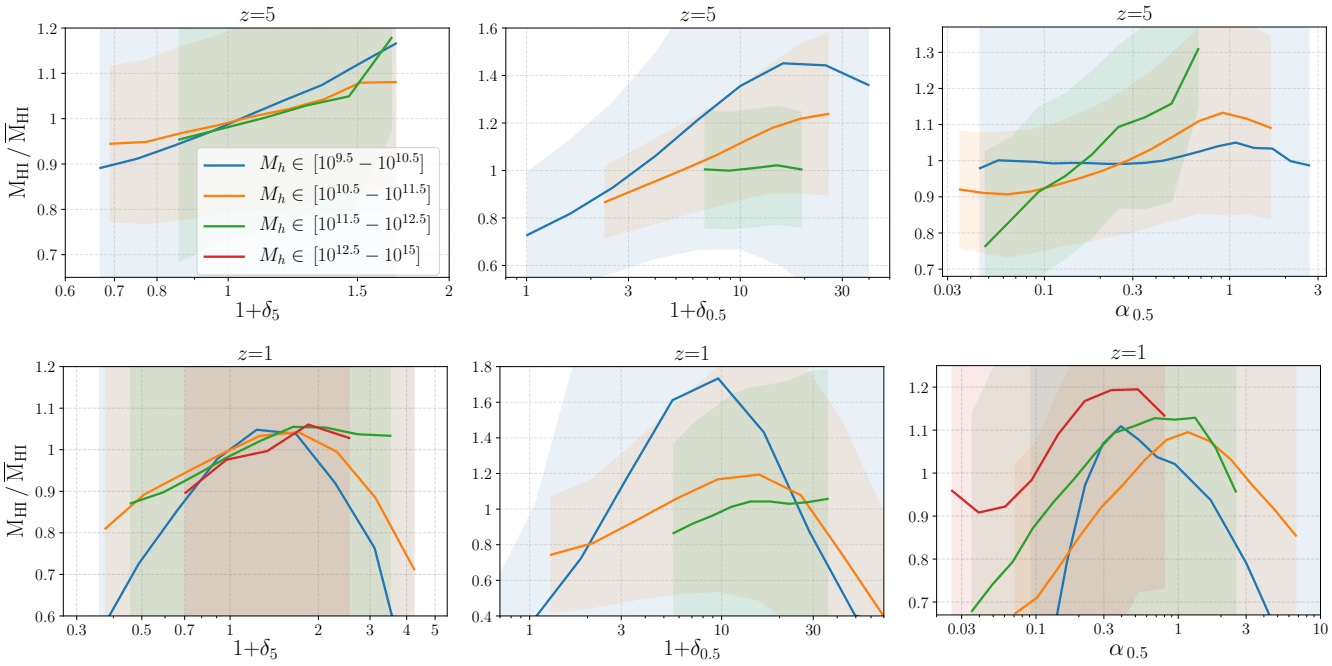

Fig. 3. This figure shows the effect of various halo environmental properties on their HI masses (M_{HI}). δ_R (α_R) is the environmental overdensity (anisotropy) within a radius $R/(\,h^{-1}\,\mathrm{Mpc})$ surrounding the halo as defined in Eq. 3 (Eq. 5). We split the halos in the TNG300-1 volume into 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 $h^{-1}\,M_\odot$.

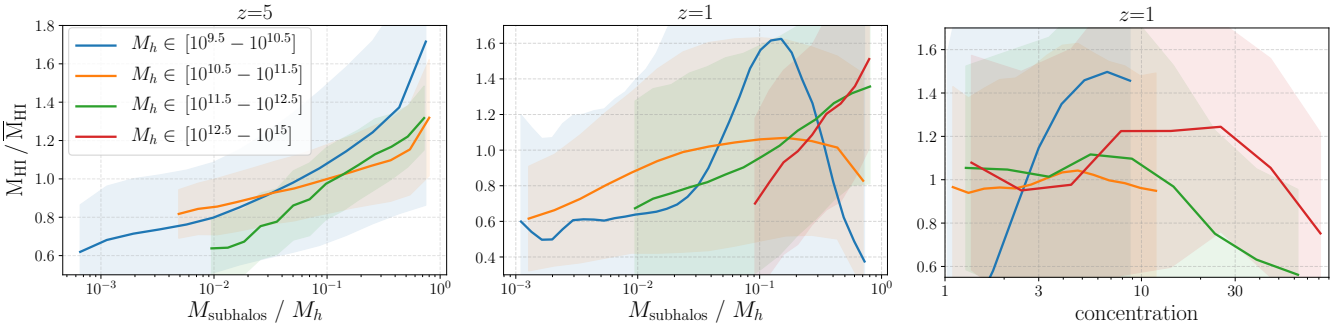

Fig. 4. Same as Fig. 3 but for the effect of internal halo properties like the fractional mass in subhalos (left and middle) and concentration on the halo HI mass (right). These trends are shown to complement the physical explanations outlined in Sec. 3A for the effect of the environment seen on the 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 overdensity 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 HI 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 HI 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 HI 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, $\delta'_{\rm R} \equiv \log(2+\delta_{\rm R})$ 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

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

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

We have presented the most concise expressions that include the effect of environment on the HI 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 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 HI 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 HI-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 HI by a significant amount. See Ref. (71) for a more comprehensive modeling of HI-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 HI mass. In this section, we investigate how this dependence propagates into the clustering of HI (we refer to this phenomenon as HI assembly bias in this paper). We will now focus on the real-space HI 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 HI. 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 HI than those in sparser regions, the overall clustering of HI would increase; this is because the calculation of $P_{\rm HI}$ involves weighting halos by their HI mass and the halos in denser regions would get upweighted. We thus conclude that halo environment not only can affect the HI content of a halo, but also the overall clustering signal of HI.

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

[††] 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 $\alpha \to 0$.

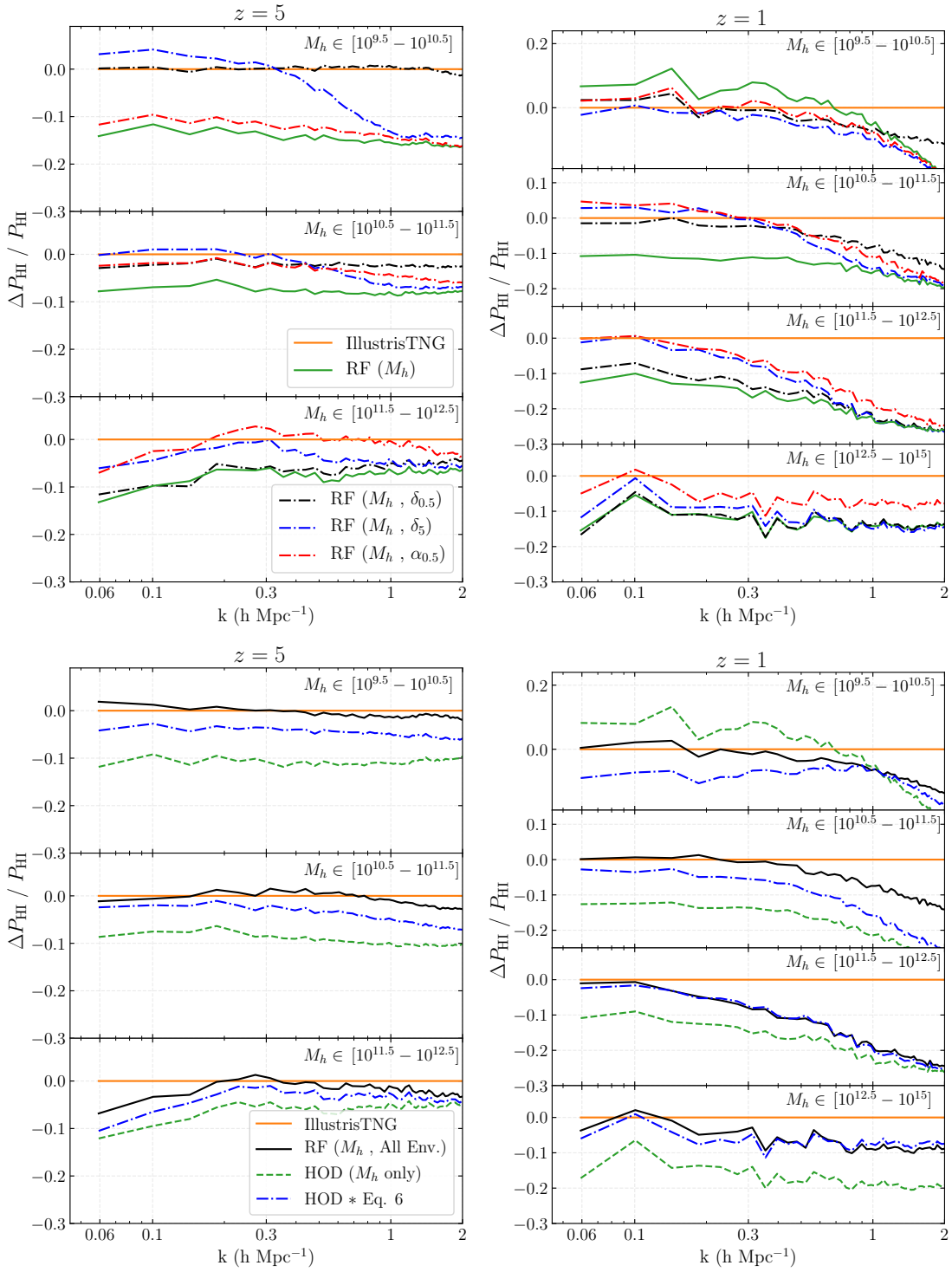

Fig. 5. This figure illustrates the effects of halo environment on the power spectrum of the modeled HI field. We train random forest (RF) regressors to predict the HI mass of halos as a function of different environmental properties (see legend). The line legends are the same for the left and right 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 HI mass.

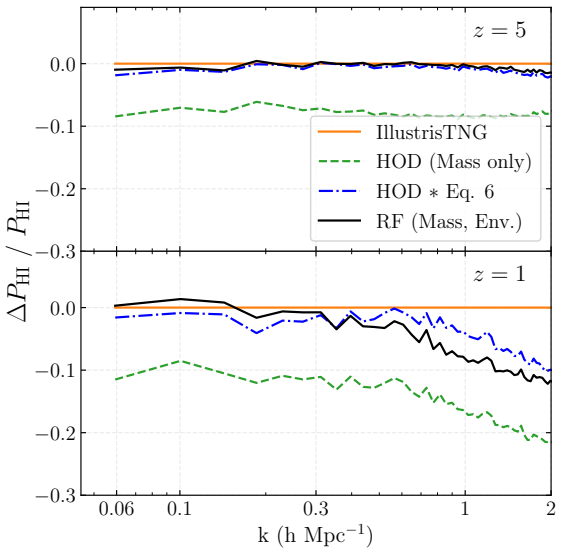

Fig. 6. Same as bottom panels of Fig. 5 but for a combined set of halos of all masses. Note that modeling the effect of the halo environment using either random forests (RF) or symbolic regression significantly improves the accuracy of the predicted HI field at both high and low redshifts; we observe a $\sim$ 10% improvement in both cases. See Fig. S1 for a comparison of summary statistics other than $P_{\rm HI}$.

dom forest regressor (RF). We first split the TNG300-1 box into training and test sets: the full box has a side-length $205\,h^{-1}$ Mpc, from which we use a box with a side-length $150\,h^{-1}$ 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 ($k \gtrsim 1\,h\,{\rm Mpc}^{-1}$), 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 $P_{\rm HI}(k)$ 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} 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 $k \gtrsim 0.05\,h\,{\rm Mpc}^{-1}$, 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 HI, 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 HI 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 HI mass of the halo. We will also extend our techniques from HI to galaxies (model the number of galaxies in a halo instead of the halo HI mass) in an upcoming paper.

ACKNOWLEDGMENTS. We thank Hamsa Padmanabhan, Boryana Hadzhiyska, Sownak Bose, Daniel Eisenstein, Neal Dalal, Sandy Huan, Roman Scoccimarro, Chang Hahn, Andrej Obuljen, Jeremy Tinker and Miles Cranmer for fruitful discussions. We also thank Hamsa Padmanabhan for detailed comments on the draft of this paper. FVN acknowledges funding from the WFIRST program through NNG26PJ30C and NNN12AA01C. The work of SH is supported by Center for Computational Astrophysics of the Flatiron Institute in New York City. The Flatiron Institute is supported by the Simons Foundation. This work was also supported in part through the NYU IT High Performance Computing resources. We thank the IllustrisTNG collaboration for making their simulation data publicly available. We have used the publicly available Pylians3 libraries[‡‡] to carry out the analysis of the simulations and PySR[†] package for symbolic regression.

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

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

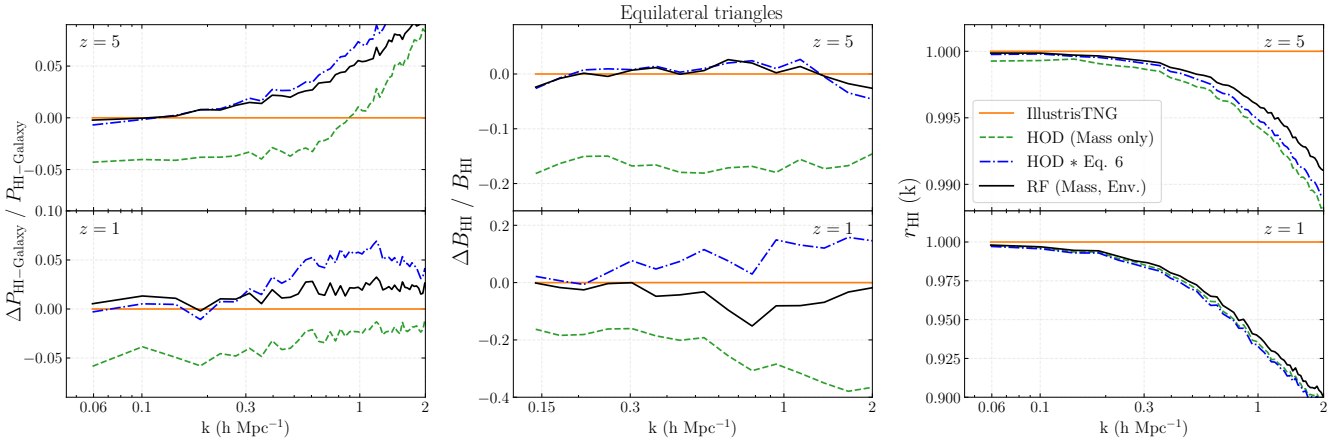

Fig. S1. Same as Fig. 6, but for different summary statistics of the modeled HI field. **Left:** Cross power between HI and galaxies, where we use galaxies in the TNG300-1 box with $M_* > 10^{10} M_\odot$. Note that the deviation in the cross power at high-k at $z = 5$ is due to low number density of 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 $r_{\rm (A)} = P_{\rm A-Illustris}/\sqrt{P_{\rm A} P_{\rm Illustris}}$). Overall, including the halo environment provides significant improvement for $P_{\rm HI-Galaxy}$ and $B_{\rm HI}$ but a modest change in $r_{\rm HI}$. 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. $P_{\rm HI-Galaxy}$ 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 $P_{\rm HI-Galaxy}$ signal at $z \sim 1$ (97, 98). To calculate $P_{\rm HI-Galaxy}$, we use galaxies in the TNG300-1 sample with stellar masses $\rm M_* > 1 \times 10^{10}$ $\rm M_\odot$ (which corresponds to a number density of $n = 2 \times 10^{-4}$ $\rm h^3/Mpc^3$ at $z = 5$ and $n = 9 \times 10^{-3}$ $\rm h^3/Mpc^3$ at $z = 1$). We see that, similarly to $P_{\rm HI}(k)$ in Fig. 6, modeling the halo environment improves the $P_{\rm HI-Galaxy}$ 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 $k \gtrsim 1\, h\,{\rm Mpc}^{-1}$ 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 $P_{\rm HI}$ 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 $M_{\rm HI}$ 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 $P_{\rm HI}$ 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.

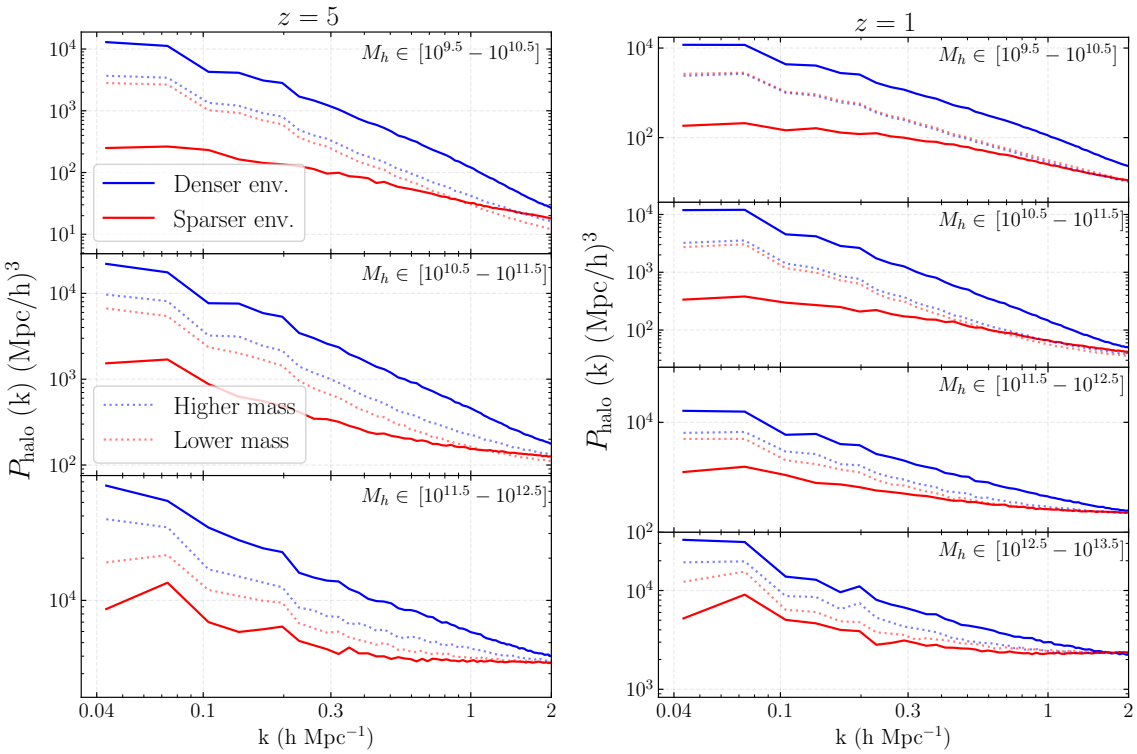

Fig. S2. Difference in the clustering of halos in denser and sparser environments for different halo masses (M_h). We split the halos in each mass bin into two sets based on their environmental overdensity: those with $\delta_{2.5}$ values above (Denser env.) or below (Sparser env.) the median; we show 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 HI than average in halos in denser environments (which amounts to upweighting their contribution to the power spectrum of HI), the overall HI 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: $c = R_{200c}/R_{\rm max}$, where $R_{\rm max}$ 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 $R_{\rm max}$.

One might expect the concentration of a halo to affect its HI 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 $M_{\rm HI}$. Upon training the RF with the halo concentration, we find very little improvement in the HI clustering prediction.

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

In Sec. 2, we discussed a mass based HOD model for HI. In order to calibrate the model parameters $\{M_0, \alpha, M_{\rm min}\}$ in Eq 2., we had used the data for the total mass M_h and the HI mass $M_{\rm HI}$ 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$10% in the two simulations at both high and low redshifts. This difference also affects the HOD modeling of the HI field. Our model was calibrated for the halo masses from the FP simulation ($M_h^{\rm FP}$). Using this model on halos from the DMO version of the TNG300-1 box, we find that $P_{\rm HI}$ 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 $\Omega_{\rm HI}$ is overpredicted by ~ 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 $M_{\rm HI}$ from the FP simulation and $M_h^{\rm DMO}$ from the DMO simulation. We obtain the model parameters $\{\alpha, M_0/(10^{10}\,h^{-1}\,M_\odot), M_{\rm min}/(10^{10}\,h^{-1}\,M_\odot)\}$ 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 $P_{\rm HI}$ 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.

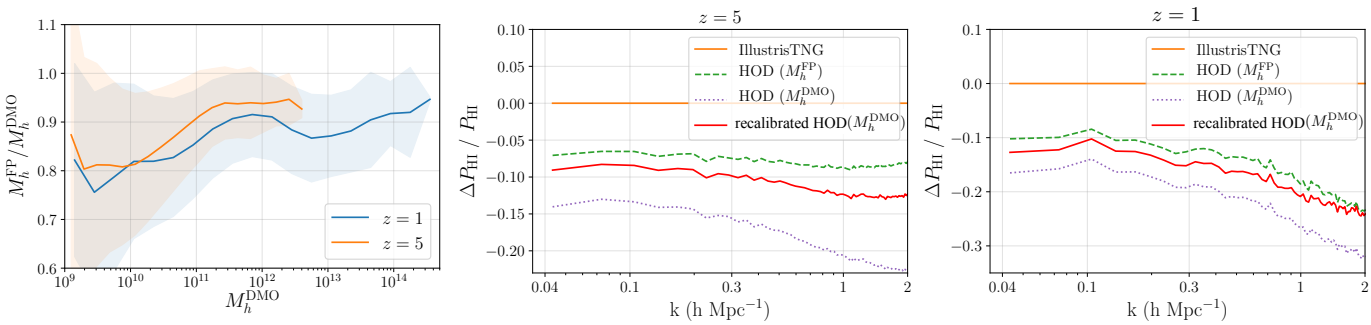

Fig. S3. This figure illustrates that an HOD relation, which is calibrated using halo masses from a hydrodynamic simulation, will give discrepant results if directly used on halos in a N-body simulation. We first match the halos in the hydrodynamic (full physics, FP) simulation to their counterparts 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 $P_{\rm HI}$ 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 ($M_{\rm HI}/M_{\rm HOD\text{-}HI}$), 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.

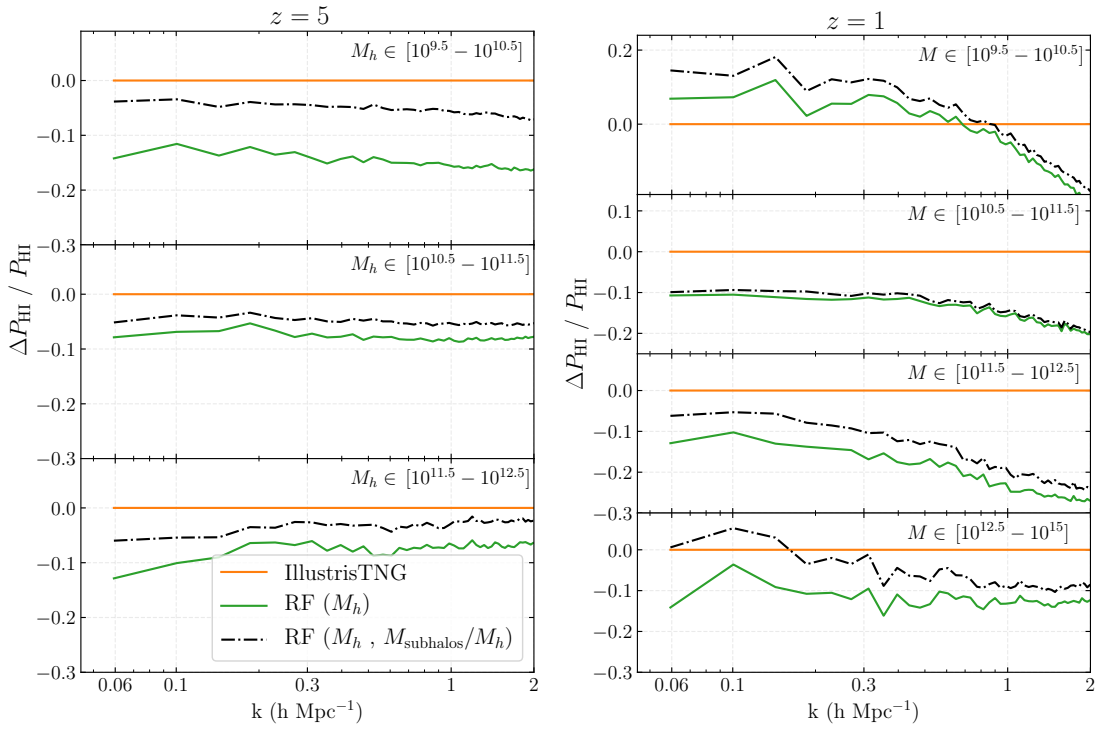

Fig. S4. Same as the top panel of Fig. 5 but training the random forest (RF) regressors using the subhalo mass fraction of the halos ($M_{\rm subhalos}/M_h$). In most cases, we find improvements similar to those seen for environmental parameters in Fig. 5. Note however that the predictions become worse 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 $P_{\rm HI}$ predictions, and hence we do not show results for that case.

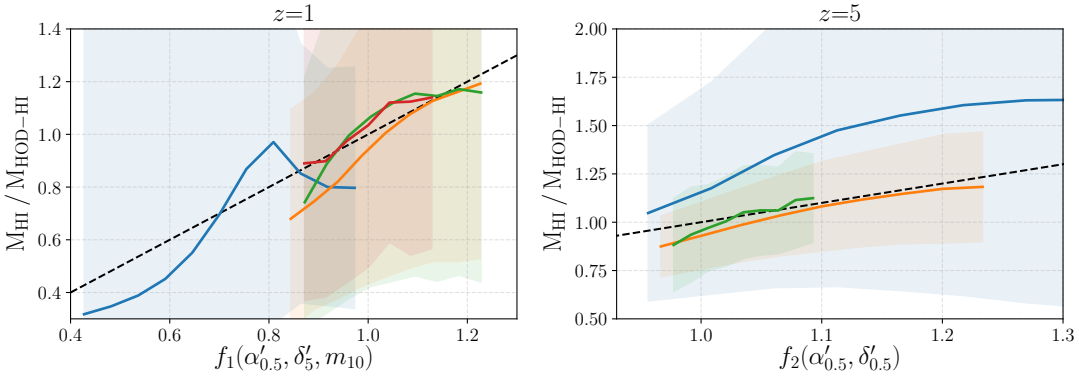

Fig. S5. We train a symbolic regressor to predict the ratio of the HI mass of a halo to the output $M_{\rm HOD-HI}$ from the mass-based HOD model in Eq. 2. We compare the values predicted from Eq. 6 against the true values from IllustrisTNG (the black dashed line represents predicted=true). f_1 and f_2 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 $M_{\rm HI}/M_{\rm HOD-HI}$ 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.

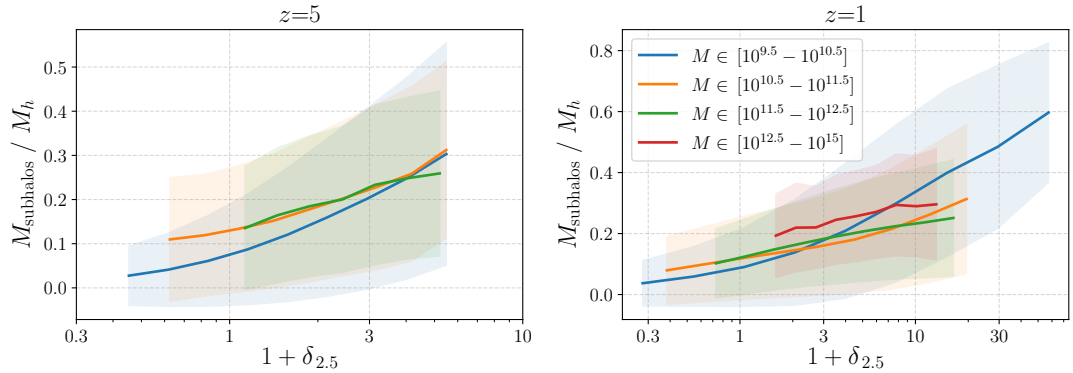

Fig. S6. Effect of the environmental overdensity of the halo, parameterized by $\delta_{2.5}$, on its subhalo mass fraction. Increasing the environmental overdensity typically results in more mergers and therefore more substructure in halos. These trends are shown to complement the physical explanations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

From Dark Matter to Galaxies with Convolutional Neural Networks

Jacky H. T. Yip*
Department of Physics, The Chinese University of Hong Kong
1155092406@link.cuhk.edu.hk

Xinyue Zhang,* Yanfang Wang,* Wei Zhang,* Yueqiu Sun*
Center for Data Science, New York University

Gabriella Contardo, Francisco Villaescusa-Navarro, Siyu He, Shy Genel, Shirley Ho
Center for Computational Astrophysics, Flatiron Institute

Abstract

Cosmological simulations play an important role in the interpretation of astronomical data, in particular in comparing observed data to our theoretical expectations. However, to compare data with these simulations, the simulations in principle need to include gravity, magneto-hydrodyanmics, radiative transfer, etc. These ideal large-volume simulations (*gravo-magneto-hydrodynamical*) are incredibly computationally expensive which can cost tens of millions of CPU hours to run. In this paper, we propose a deep learning approach to map from the dark-matter-only simulation (computationally cheaper) to the galaxy distribution (from the much costlier cosmological simulation). The main challenge of this task is the high sparsity in the target galaxy distribution: space is mainly empty. We propose a cascade architecture composed of a classification filter followed by a regression procedure. We show that our result outperforms a state-of-the-art model used in the astronomical community, and provides a good trade-off between computational cost and prediction accuracy.

1 Introduction

Most cosmological surveys observe galaxies, which are however very difficult to model directly due to the complicated physics involved in their formation and evolution. For even a small fraction of the Universe, evolving tens of billions of resolution particles interacting under coupled effects of gravity, magneto-hydrodynamics and radiative processes over cosmic time is incredibly computationally costly; a state-of-the-art simulation today, such as IllustrisTNG [1], requires up to 40 million CPU hours to complete ($\sim$4500 years on a single CPU).

On the other hand, in the ΛCDM cosmological model, baryonic matter ("normal" matter, which galaxies are made out of) constitutes only about one-sixth of all the matter in the Universe. The rest is composed of (notably) dark matter, that interacts only through gravity and drives the growth and morphology of the large-scale structure of the Universe, such as galaxy clusters, filaments, and voids. Baryonic matter collapses within dark matter halos, which results in the formation of stars and galaxies. In other words, dark matter halos are the environments in which galaxies form, evolve and merge. Hence, the behaviors and properties of galaxies, such as their spatial distribution, are expected to be closely connected to the properties of dark matter halos they live in.

*These authors contributed equally.

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

In contrast to full hydrodynamical simulations, dark-matter-only N-body simulations are computationally much cheaper as gravity is the only interacting force. Therefore, it would be extremely interesting to find the mapping between the dark matter and galaxy fields in N-body and full hydrodynamical simulations, respectively. State-of-the-art Halo Occupation Distribution (hereafter HOD) [3] model has been developed to achieve this, but it is unclear whether it is comprehensive enough as the model usually relies on certain assumptions such as the halo mass being the main quantity controlling galaxy properties, and not, for instance, its local structure.

We propose a first deep learning approach for the above purpose. We explore the use of a cascade of convolutional neural networks (CNNs) to map from the dark matter distribution, obtained from a gravity-only N-body simulation, to the galaxy distribution, obtained from a full gravo-magneto-hydrodynamical simulation. We show that our model predicts a significantly more reliable galaxy distribution compared to the HOD algorithm on a variety of criteria, including statistics commonly used in cosmology.

2 Method

2.1 Data

From the IllustrisTNG project [2], we use the TNG300 full hydrodynamical simulation (TNG300-1) since it simulates the largest range of spatial scales for this type of simulations. We also employ the associated N-body simulation (TNG300-1-Dark). We use the level-1 simulations (highest spatial and mass resolution) at present day Universe (redshift $z = 0$). Note that our approach could be applied in similar fashion on any simulation pairs of dark matter and their respective hydrodynamical simulation.

The input data is the mass density field of the dark matter distribution, which allows us to be independent of the specific configurations of the simulation such as the mass resolution. The target data is the galaxy number density field. We define galaxies as the subhalos with a positive stellar mass (total mass of stars particles)[1]. In the following experiments, we do not apply a threshold on the minimum stellar mass to our target set of galaxies, i.e. we will predict all galaxies with non-zero stellar mass.

Both the dark matter particles and the galaxies from the TNG300 simulations are enclosed in a 3D simulation volume of (205 Mpc/h)^3. We grid this space into 1024^3 voxels, and generate sub-cubes (taken as inputs and targets) of 32^3 voxels each. The resulting dataset is composed of 32^3 pairs of samples. We split this dataset into training (63.1%), validation (19.1%) and test (17.8%) sets. The test set forms a coherent volume of (115.3 Mpc/h)^3 for computing the relevant cosmological statistics.

2.2 Cascade architecture

We physically expect that the formation of a galaxy should primarily depend on the local properties of the halo and the environment where it lives, and that it is independent of its absolute position. This motivates our choice of convolutional neural network (CNN) based architecture [6]. However, the percentage of non-empty voxels is 75.93% for the input field and 0.15% for the target galaxy field. The high sparsity in the latter is the major difficulty of the task, because the model would still achieve a high accuracy (99.85%) even if it fails to predict all the galaxies. We thus propose a two-phase cascade architecture to overcome this problem.

Classification phase. The first phase is an *Inception Network* [4] as a classifier to predict the probability of having at least one galaxy in each of the target voxels. We use the weighted cross-entropy loss for this binary classification problem:

$$\mathbb{L}_{\text{CrossEnt}}(\mathbf{x}, \mathbf{y}) = -mean\{[w_1 \cdot \mathbf{y} \cdot \log(\mathbf{x}) + (1 - \mathbf{y}) \cdot \log(1 - \mathbf{x})]/(w_1 + 1)\} \quad (1)$$

where $\mathbf{x}$ contains the predicted probabilities in the voxels for the presence of galaxies, $\mathbf{y}$ contains the target values (1 for presence; 0 for absence) and w_1 is the weight applied to emphasize the penalty on bad predictions in the voxels in which galaxies are present. This weight effectively corrects the imbalance in the target classes due to the high sparsity.

[1] The `SubhaloFlag` field from the catalog is also used to filter out non-cosmological subhalos.

Regression phase. The second phase is a *Recurrent Residual U-Net* [5]. Similarly, it takes the dark matter density field as input, and regresses the number of galaxies $\mathbf{n_g}$ in each voxel. The output from the previous phase is used as a mask to compute the loss and backpropagate when considering only the voxels that were predicted in the previous phase to have at least one galaxy. We optimize this phase using a weighted mean square error (MSE) loss as follows:

$$\mathbb{L}_{\mathrm{MSE}}(\mathbf{n_g}, \mathbf{n_t}) = mean\{W(\mathbf{n_t})(\mathbf{n_g} - \mathbf{n_t})^2\} \tag{2}$$

where $\mathbf{n_t}$ contains the numbers of galaxies in the target, and the weight function W is defined by $W(n_{t,i}) = w_2$ for $n_{t,i} > 0$ and $W(n_{t,i}) = 1$ otherwise. Similarly, this weight function provides a means (by varying w_2) to emphasize the loss from the voxels in which galaxies are present in the target.

The output of this cascade model is continuous (not categorical integer prediction), which can be interpreted as a probabilistic number of galaxies in each voxel. As we focus here on statistics which do not require exact numbers of galaxies (e.g. the overdensity field used for computing the power spectrum and bispectrum in the next Section), we keep the output as is. Otherwise, one may round the output to obtain a generic number density field.

Hyperparameter search and training. While the MSE (Eqn. (2)) is effective as the loss, it is not an ideal indicator of the model's performance: if the model predicts a galaxy's presence in a voxel next to where it truly is, this MSE would yield a significant error while the slightly misplaced galaxy prediction does not practically make a big difference in most statistics of interest here. Therefore, for determining which model performs best, we use the MSE in the total number of galaxies in each 32^3 sub-cube (MSE_{sub}), which is the baseline statistic that should be matched well by the prediction.

The classifier and the regression model are trained separately. We prioritize training the classifier for a high recall in order to minimize the number of non-empty voxels incorrectly removed (false negative) as they would become unrecoverable in the regression phase. For instance, with $w_1 = 500$, we obtain a result with 99.07% recall and 8.28% precision, i.e. only $\sim$1% of non-empty voxels are incorrectly classified while the sparsity is effectively reduced by $\sim$55 times. For the regression model, the weight w_2 is tuned to minimize the corresponding MSE_{sub} on the validation set. The optimal weight is found to be $w_2 = 1.05$.

The model in each phase is trained for 50 epochs with the Adam optimizer and a batch size of 16. The learning rate starts from 0.001 and is halved every 4 epochs. We augment the training data by using random rotations of the cubes. All hyper-parameters are selected based on the performance on the validation set. The total training time with 1 GPU is $\sim$10 hours.

3 Results

All results shown here are computed on the test set. We compare our model's result with the state-of-the-art Halo Occupation Distribution model. The HOD model identifies mass-related-only parameters to determine the number of galaxies that a dark matter halo holds, then the galaxies are placed randomly within a critical radius inside the halo.

Visualization. Figure 1 shows snapshots of a slice from the input, target and results of the cascade model and HOD. Both the cascade model and HOD successfully predict approximate positions of the galaxies. However, from the zoomed-in images (second row), it is evident that our cascade model (third column) learns small scale details on the distribution of galaxies, while HOD only predicts a spherical distribution of galaxies within the halo.

Power spectra and bispectra. Summary statistics such as the power spectrum and the bispectrum are most commonly used in cosmology to extract information from fields of fluctuation, such as the overdensity field of the galaxy distribution. The power spectrum $P(k)$ is the Fourier equivalent of the two-point correlation function which measures how the actual galaxy distribution deviates from a simple Gaussian random field. On the other hand, the bispectrum $B(k)$ is the Fourier equivalent of the three-point correlation function, which is a higher-order statistic commonly used for characterizing the galaxy distribution on smaller scales as a non-Gaussian field due to non-linear interactions. Figure 2 shows the power spectra (left) and bispectra (right) of the target and the results of our model and HOD.

For the power spectrum, our cascade model obtains a good fit on a large range of scales even though it is trained on relatively small sub-cubes. Over the full range of scales, its mean relative residual[2] is 26.9% which is much smaller than HOD's 246.6%. Our model achieves a comparable performance as HOD for $k < 0.3$ h/Mpc (larger scales), and outperforms for larger k (smaller scales). This shows that our model can better capture the non-linearities in the smaller scales of the field. This is further demonstrated in the bispectra on small scales ($k_1 = 1.2$ h/Mpc and $k_2 = 1.3$ h/Mpc), in which our model outperforms HOD by a significant margin (0.79% vs 435% in mean relative residual).

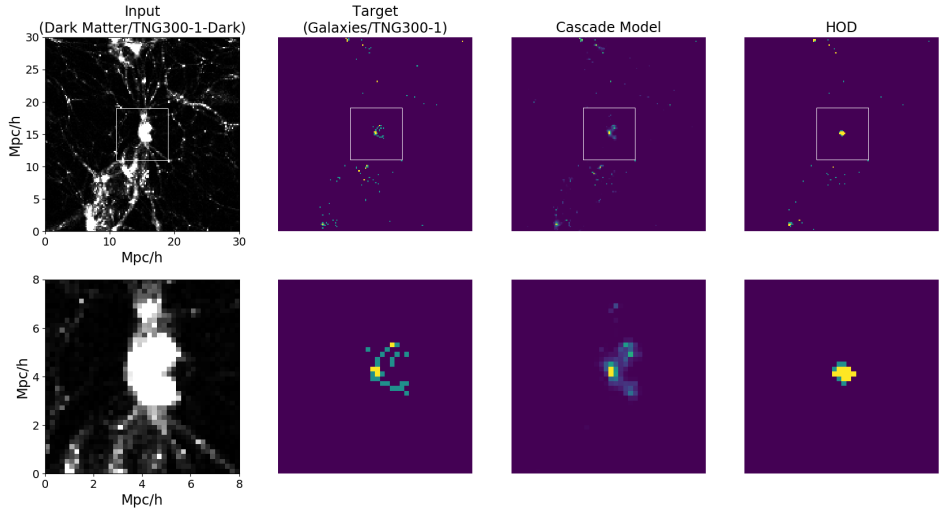

Figure 1: Snapshots of a slice from simulations and results: First column shows the dark matter input; Second shows the target galaxies; Third shows the prediction from our cascade model; Forth is from our benchmark model, the commonly deployed method in cosmology. Second row is a zoomed-in on the white squares in the first row. Brighter colors represent more dark matter particles/galaxies[3].

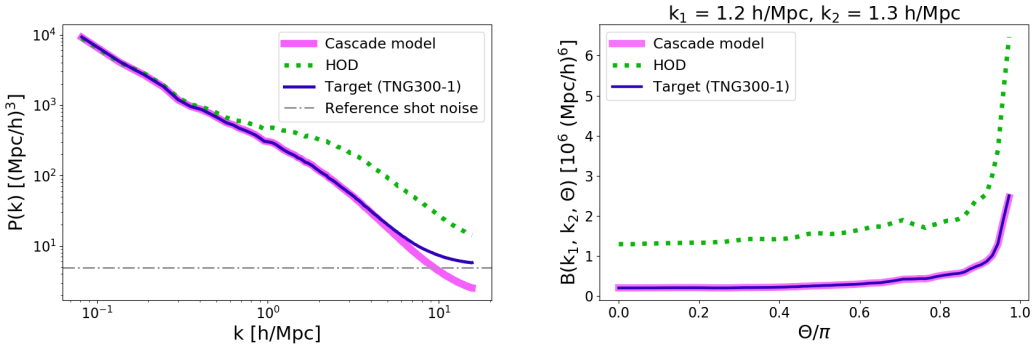

Figure 2: Power spectra $P(k)$ (left) and bispectra $B(k_1, k_2, \theta)$ (right)

4 Conclusion and future work

We show in this paper that our cascade model of convolutional neural networks can efficiently predict the number density field of galaxies given only the dark matter density field, and can outperform a benchmark method used in cosmology in a large range of scales.

We plan for several extensions of this work, notably focusing on predicting additional properties of the galaxies (e.g. stellar mass, star formation rate, etc). We are also looking into the ability of the model

[2]The mean relative residual is defined by $mean\{|\mathbf{y_m}/\mathbf{y_t} - 1|\}$, where $\mathbf{y_m}$ and $\mathbf{y_t}$ are spectrum values of the model's prediction and the target respectively.

[3]For illustration purposes, colors in the galaxies snapshots range from blue-violet (darkest) to blue-green and to yellow-orange (brightest). Yellow-orange represents voxels with 2 or more galaxies, blue-green represents voxels with 1 galaxy and blue-violet represents empty voxels in the backgrounds. For our cascade model's prediction, there are more intermediate colors as it predicts continuous numbers of galaxies.

to transfer between simulations of different volumes and resolutions. Not only this work paves the way for obtaining simulation results more efficiently, but also could help explore scientific questions. For example, we are interested in whether our model can learn about halo assembly bias [7], that HOD neglects. Our approach could be used to analyse any possible effects of the environment (e.g. structure of the dark matter halos) regarding galaxy formation.

Acknowledgments

We thank David Spergel, Siamak Ravanbakhsh and Barnabas Poczos for insightful discussions. This project is supported by the Center for Computational Astrophysics of the Flatiron Institute in New York City. The Flatiron Institute is supported by the Simons Foundation.

References

[1] Annalisa Pillepich, Volker Springel, Dylan Nelson, Shy Genel, Jill Naiman, Ruediger Pakmor, Lars Hernquist, Paul Torrey, Mark Vogelsberger, Rainer Weinberger, Federico Marinacci. Simulating Galaxy Formation with the IllustrisTNG Model. *Mon. Notices Royal Astron. Soc* 473 3 4077–4106, 2018

[2] Dylan Nelson, Volker Springel, Annalisa Pillepich, Vicente Rodriguez-Gomez, Paul Torrey, Shy Genel, Mark Vogelsberger, Ruediger Pakmor, Federico Marinacci, Rainer Weinberger, Luke Kelley, Mark Lovell, Benedikt Diemer, Lars Hernquist. The IllustrisTNG Simulations: Public Data Release. *ArXiv e-prints*, Apr 2019

[3] Andreas A. Berlind, David H. Weinberg. The Halo Occupation Distribution: Towards an Empirical Determination of the Relation Between Galaxies and Mass. *Astrophys. J.* 575 587-616, 2002

[4] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich. Going Deeper with Convolutions. *ArXiv e-prints*, Sep 2014

[5] Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, Vijayan K. Asari. Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation. *ArXiv e-prints*, May 2018

[6] Stéphane Mallat. Understanding deep convolutional networks. *Philos. Trans. Royal Soc. A Math. Phys. Sci.* 374 2065, Apr 2016

[7] Mohammadjavad Vakili, ChangHoon Hahn. How are galaxies assigned to halos? Searching for assembly bias in the SDSS galaxy clustering. *Astrophys. J.* 872 1, 2019