HIGAN: Cosmic Neutral Hydrogen with GANs

Juan Zamudio-Fernandez [*]
New York University

Atakan Okan [*†]
New York University

Francisco Villaescusa-Navarro
Flatiron Institute

Seda Bilaloglu
New York University

Asena Derin Cengiz
New York University

Siyu He
Flatiron Institute

Laurence Perreault Levasseur
Flatiron Institute

Shirley Ho
Flatiron Institute

Abstract

In this work we use a Wasserstein Generative Adversarial Network (WGAN) to generate new high-resolution 3D realizations of cosmic neutral Hydrogen (HI). The generator produces samples that match closely the fully non-linear abundance and clustering properties of cosmic HI from the state-of-the-art simulation IllustrisTNG. We show that different statistical properties of the generated samples – 1D PDF, power spectrum, bispectrum, and void size function – match very well to those of IllustrisTNG, and outperform state-of-the-art models such as Halo Occupation Distributions (HODs). Our WGAN samples reproduce the abundance of HI across 9 orders of magnitude, from the Lyα forest to Damped Lyman Absorbers.

1 Introduction

In the coming years, powerful cosmological surveys will be collecting data with the aim of constraining the value of cosmological parameters with the highest accuracy possible, in order to improve our understanding of the fundamental physical laws of the Universe. The abundance and spatial distribution of HI, directly observable with these surveys, contains a large amount of information on fundamental physical quantities. In order to extract it, one must compare the data from these missions against accurate and precise theory predictions. Numerical simulations are one of the most powerful ways to obtain the theory predictions.

The current state-of-the-art magneto-hydrodynamic simulations from ILLUSTRISTNG required more than 150 million CPU hours [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. In this work, we investigate the use of Generative Adversarial Networks (GANs) to quickly generate high-resolution 3D distributions of cosmic HI with the same statistical properties as the computationally expensive IllustrisTNG simulation. This constitutes a significant first step towards an efficient generation of accurate theoretical predictions of cosmic HI from large to small scales, needed to maximize the scientific return of aforementioned upcoming 21cm surveys.

2 Data

The simulation used as training data is TNG100-1, produced by the IllustrisTNG collaboration [7]. We focus our attention on cosmic HI at redshift 5 (1.2 Gyr after the Big Bang). The sparsity of the

[*]Equal Contribution

[†]Corresponding author: ao1512@nyu.edu

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

Table 1: Statistical properties of our training data

Min	Max	Median	Mean	SD
0	4.4×10^9	13.2	1.5×10^4	9.2×10^5

field is smaller at this redshift than in lower redshifts, facilitating the training of deep learning models. The data consists of a $2048 \times 2048 \times 2048$ cube where the value in each cell of the grid is the total HI mass. We have kept a cubic section with a volume equal to 1/8 times that of the entire cube as test set. For the training data, we randomly sample sub-cubes of $64 \times 64 \times 64$ cells, corresponding to a comoving volume of $\sim (2.34\, h^{-1}\mathrm{Mpc})^3$, from the region outside of the test set.

The probability distribution function of our data exhibits a very long tail for cells with large HI masses (see table 1). In order to facilitate the training of our model, we perform the following transformation to scale the data to the $[-0.23, 1]$ interval [3]:

$$\tilde{m}_{\mathrm{HI}} = \frac{\log_{10}(m_{\mathrm{HI}} + \epsilon)}{\log_{10}(m_{\mathrm{HI}}^{\mathrm{max}} + \epsilon)} \tag{1}$$

where $m_{\mathrm{HI}}^{\mathrm{max}}$ is the maximum HI mass from all cells and we have set $\epsilon = 0.01$.

3 Model & Training

In the Wasserstein GAN setup [13], the Critic, instead of trying to differentiate between real and generated samples, provides an approximation of how far the generated samples are from the real ones by using an approximation of the Earth Mover distance. [14] showed that using gradient penalty in the critic's loss function instead of weight clamping is a better approach to ensure that the Lipschitz constraint is met. Hence, the critic's loss function is given by:

$$L_D\left(\theta_d\right) = \mathbb{E}\left[D\left(G\left(\mathbf{z};\theta_g\right);\theta_d\right)\right] - \mathbb{E}\left[D\left(\mathbf{x};\theta_d\right)\right] + \lambda \mathbb{E}_{\hat{\mathbf{x}}\sim\mathbf{P}_{\hat{x}}}\left[\left(\left\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\right\|_2 - 1\right)^2\right] \tag{2}$$

where D is the critic parametrized by θ_d, G is the generator parametrized by θ_g, z is a vector of size 100 sampled from a standard Gaussian distribution, and λ is the gradient penalty coefficient which we set equal to 10. $\hat{x}$ is an interpolation between the real and generated samples: $\hat{x} = tx + (1-t)y$, where $t \sim U[0,1]$ and $x \sim P_r, y \sim P_g$.

The loss function of the generator is given by:

$$L_G\left(\theta_g\right) = -\mathbb{E}\left[D\left(G\left(\mathbf{z};\theta_g\right);\theta_d\right)\right] \tag{3}$$

We use an all-convolutional architecture, similar to the deep convolutional GAN (DCGAN) implementation of [15], adapted to three-dimensional data. The generator consists of five 3D fractionally-strided convolutions (transpose convolution) layers with rectified linear unit (ReLU) activation and batch norm (3D) layers followed by two convolutional layers, with filter sizes 3 and 2, respectively. The final activation function is tanh. The number of channels in the first layer is 1024, halving in size until the last layer, which receives 128 channels and outputs 1. The critic has seven convolutional layers with leaky ReLU activations. The number of channels is symmetric to the generator. Adding the extra convolutional layers without changing the feature map sizes improved significantly the quality of the generated samples.

We used the Adam optimization algorithm with learning rate of 0.0005 and betas of 0.5 and 0.9. We trained for approximately 400 hours and 150,000 generator iterations.

4 Benchmark, Results & Validation Metrics

In Fig. 1, we show examples of 3D HI distributions from IllustrisTNG and generated from our trained WGAN. As it can be seen, the WGAN is able to generate realizations that visually look very similar the IllustrisTNG samples. Likewise, in Fig. 2, we can see that the generator has learned a

[3] Where -0.23 corresponds to the empty cells (zero mass) and 1 to the maximum value in the original scale.

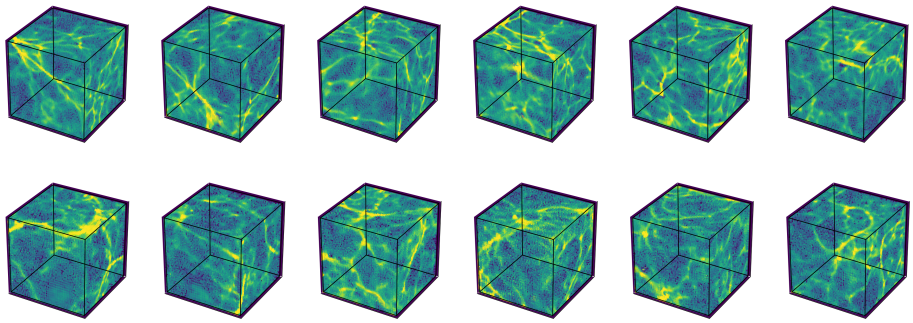


Figure 1: 3D HI distributions from IllustrisTNG (top row) and WGAN (bottom row). The Generator produces spatial distributions with all the elements of the HI web: filaments, voids and dense regions.

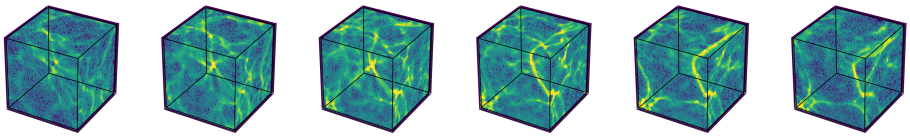
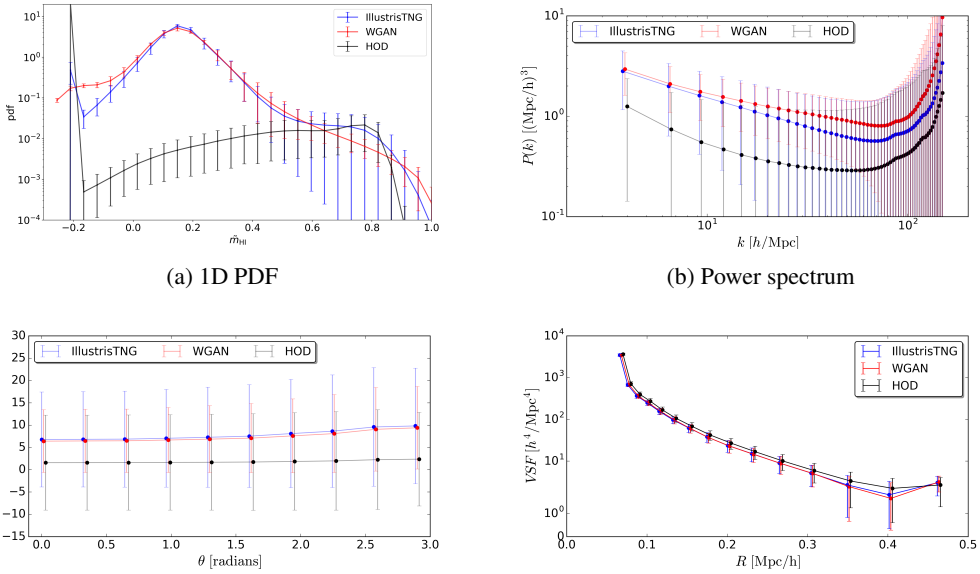


Figure 2: WGAN Samples generated by linearly interpolating between two latent space vectors. The generator learns a smooth mapping from the latent space to the output space.

smooth mapping from the latent space to the output space. In our case, the volume of each cell is $\simeq (35\ h^{-1}\mathrm{kpc})^3$, i.e. we are deep into the non-linear regime. In addition, in Fig. 3 we quantify the agreement between the 3D HI distribution of the different models using four different summary statistics that are sensitive to different regions of the underlying field.

Benchmark: Halo Occupation Distribution In the post-reionization era, most of the cosmic HI in the Universe resides within dark matter halos, this is the basis principle underlying the state-of-the-art



(a) 1D PDF

(b) Power spectrum

(c) Bispectrum

(d) Voids

Figure 3: Validation metrics for 1000 samples from IllustrisTNG, WGAN and HOD.

framework developed in [16] to model the abundance and spatial distribution of HI. This model is used as a benchmark to judge the performance of our generative model. The procedure to generate 3D HI mock distributions is as follows. We randomly select a sub-cube of TNG100-1-Dark, which is a dark-matter-only N-body simulation with the same mass, spatial resolution, and mode phases as the full hydrodynamical TNG100-1. We then identify all halos that lie inside that region by finding their positions and masses. For each halo, we then compute its total HI mass and then distribute the total HI mass inside the halo using 1000 particles. For each particle, its radius is drawn from the HI density exponential distribution, while its direction is taken randomly. We repeat this procedure for all halos that lie within the considered sub-cube. Finally, we construct a $64 \times 64 \times 64$ grid from the positions and HI masses of all the particle tracers, using the CIC interpolation scheme.

1-D Probability Mass Function We compute the mean and standard deviation in log scale as a function of HI mass for the three sample sets. The WGAN is able to reproduce the PDF from IllustrisTNG from $\tilde{m}_{\rm HI} \in [0, 1]$, i.e. for more than 9 orders of magnitude for HI masses. While the HOD fails to reproduce the central part of the PDF, which is dominated by low HI density cells, corresponding to the Lyα-forest. On the low mass tail, the WGAN model performs slightly worse. In particular, we can see how our model fails at reproducing a peak around -0.23, which corresponds to empty cells in IllustrisTNG. We notice however that those cells will have negligible contribution to the observed signal.

Power Spectrum In Fig. 3b we show the mean HI power spectrum and standard deviation as a function of wavenumber from the three sample sets. We find that the HI distribution from the WGANs samples have, on large scales, an average power spectrum which is very close in amplitude and in shape to that of the IllustrisTNG samples. The HOD model is not able to reproduce the power spectrum from IllustrisTNG on any of the above considered scale. We should note the scales considered here are highly non-linear; the HOD model is expected to not perform well on those scales. We also note that, for some HI masses, the variance of the WGAN samples are smaller than the ones from IllustrisTNG. We believe that the unconditional nature of the WGAN causes the standard deviations of the WGAN values to be smaller than the variance on the simulation in the intermediate scales. We believe that this could be addressed by conditioning the WGAN on the amplitude of the HI power spectrum in a cell.

Bispectrum The bispectrum carries information pertaining to departure of a field distribution from Gaussianity. For simplicity, we have selected a value of $k_1 = 6\ h/\mathrm{Mpc}$ and $k_2 = 6.5\ h/\mathrm{Mpc}$ and varied the angle θ, i.e. k_3, although we get similar results for other configurations. The figure shows the mean and standard deviation as a function of the angle between k_1 and k_2. We find that the bispectra of the WGAN samples are closer to those of IllustrisTNG than to those of the HOD.

Voids Abundance Voids are connected regions in the HI density field that are underdense with respect to the mean HI density. Void size function, defined as the number density of voids per unit of radius as a function of void radius. We show the results in Fig. 3d. The plot shows the mean and the standard deviation of the void size function (number density of voids per unit of radius) as a function of radius for the three sets. We find that both the WGAN and HOD models produce fields with the same abundance of underdense regions as IllustrisTNG.

5 Conclusions

In this work we demonstrated the usefulness of generative adversarial neural networks to make fast and accurate predictions of the distribution of cosmic HI, over scales deep into the non-linear regime. Our WGAN model learns a 100-dimension manifold that characterizes the abundance and clustering of HI in the fully non-linear regime. By sampling from that manifold, we produce new samples with very close statistical properties to the ones from the full hydrodynamic simulation IllustrisTNG. Moreover, we found that our generated samples exhibit better agreement with the IllustrisTNG simulated samples than the current benchmark HOD model when considering four summary statistics: the 1D PDF, the power spectrum, the bispectrum, and the voids abundance.

This works represents a first step towards a fast, accurate, and precise theory prediction pipeline needed to maximize the scientific return of upcoming cosmological 21cm survey missions. We are hoping that the variance of the produced samples can be made even closer to those of costly

hydrodynamical simulations by conditioning the GANs on the amplitude of the power spectrum in the predicted volume, an avenue which we plan to explore in follow-up work. In future work, we also plan to explore the possibility of creating samples of larger cosmological volumes and at different, lower redshifts. These, along with extensions of the model to predict HI velocities to allow a redshift-space analysis and extensions to different cosmological parameter values will, in our opinion, pave the way for machine learning methods to extend the capabilities of the small and expensive hydrodynamical simulations in the analysis of up coming survey data.

References

[1] Dylan Nelson et al. The IllustrisTNG Simulations: Public Data Release. 2018.

[2] Annalisa Pillepich et al. First results from the IllustrisTNG simulations: the stellar mass content of groups and clusters of galaxies. *Mon. Not. Roy. Astron. Soc.*, 475:648, 2018.

[3] Volker Springel et al. First results from the IllustrisTNG simulations: matter and galaxy clustering. *Mon. Not. Roy. Astron. Soc.*, 475:676, 2018.

[4] Dylan Nelson et al. First results from the IllustrisTNG simulations: the galaxy color bimodality. *Mon. Not. Roy. Astron. Soc.*, 475:624, 2018.

[5] Jill P. Naiman, Annalisa Pillepich, Volker Springel, Enrico Ramirez-Ruiz, Paul Torrey, Mark Vogelsberger, Rüdiger Pakmor, Dylan Nelson, Federico Marinacci, Lars Hernquist, Rainer Weinberger, and Shy Genel. First results from the IllustrisTNG simulations: a tale of two elements - chemical evolution of magnesium and europium. , 477(1):1206–1224, Jun 2018.

[6] Federico Marinacci et al. First results from the IllustrisTNG simulations: radio haloes and magnetic fields. *Mon. Not. Roy. Astron. Soc.*, 480(4):5113–5139, 2018.

[7] Annalisa Pillepich et al. Simulating Galaxy Formation with the IllustrisTNG Model. *Mon. Not. Roy. Astron. Soc.*, 473(3):4077–4106, 2018.

[8] Rüdiger Pakmor, Rainer Weinberger, Volker Springel, Jill Naiman, Lars Hernquist, Annalisa Pillepich, Federico Marinacci, Mark Vogelsberger, Paul Torrey, Dylan Nelson, and Shy Genel. Simulating galaxy formation with black hole driven thermal and kinetic feedback. *Monthly Notices of the Royal Astronomical Society*, 465(3):3291–3308, 11 2016.

[9] Annalisa Pillepich, Jill P Naiman, Lars Hernquist, Dylan Nelson, Rainer Weinberger, Rüdiger Pakmor, Volker Springel, Federico Marinacci, Mark Vogelsberger, Paul Torrey, and Shy Genel. First results from the IllustrisTNG simulations: the stellar mass content of groups and clusters of galaxies. *Monthly Notices of the Royal Astronomical Society*, 475(1):648–675, 12 2017.

[10] Guinevere Kauffmann, Dylan Nelson, Annalisa Pillepich, Rainer Weinberger, Rüdiger Pakmor, Volker Springel, Jill Naiman, Lars Hernquist, Shy Genel, Federico Marinacci, Mark Vogelsberger, and Paul Torrey. First results from the IllustrisTNG simulations: the galaxy colour bimodality. *Monthly Notices of the Royal Astronomical Society*, 475(1):624–647, 11 2017.

[11] Rainer Weinberger, Rüdiger Pakmor, Volker Springel, Annalisa Pillepich, Dylan Nelson, Jill Naiman, Lars Hernquist, Federico Marinacci, Mark Vogelsberger, Paul Torrey, and Shy Genel. First results from the IllustrisTNG simulations: matter and galaxy clustering. *Monthly Notices of the Royal Astronomical Society*, 475(1):676–698, 12 2017.

[12] Mark Vogelsberger, Paul Torrey, Federico Marinacci, Rainer Weinberger, Rüdiger Pakmor, Volker Springel, Jill Naiman, Lars Hernquist, Dylan Nelson, Annalisa Pillepich, and Shy Genel. First results from the IllustrisTNG simulations: radio haloes and magnetic fields. *Monthly Notices of the Royal Astronomical Society*, 480(4):5113–5139, 08 2018.

[13] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *CoRR*, abs/1701.07875, 2017.

[14] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017.

[15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[16] Francisco Villaescusa-Navarro, Shy Genel, Emanuele Castorina, Andrej Obuljen, David N. Spergel, Lars Hernquist, Dylan Nelson, Isabella P. Carucci, Annalisa Pillepich, Federico Marinacci, Benedikt Diemer, Mark Vogelsberger, Rainer Weinberger, and Rüdiger Pakmor. Ingredients for 21 cm intensity mapping. *The Astrophysical Journal*, 866(2):135, oct 2018.