Journal of
High School
Science

Investigating galaxy clustering dependence on neutral hydrogen mass through dark halo properties

Zhou Z

Submitted: February 2, 2024, Revised: version 1, May 25, 2024
Accepted: May 30, 2024

Abstract

This study investigated the dependence of galaxy clustering on neutral atomic hydrogen (HI) mass using data from the TNG100 hydrodynamic simulation of 4,361,030 galaxies in the local universe ($0 < z < 0.1$). By calculating the Two-point correlation function (TPCF) for subhalos at various HI mass thresholds and comparing shuffled and original galaxy samples, environmental effects such as secondary galaxy bias (SGB) were isolated. The results showed neither a significant SGB nor a significant dependence of clustering on HI masses above 10^8 $M_\odot$. The study found that the TPCF amplitudes for different HI mass thresholds did not statistically differ, consistent with some literature reports while contrasting with others that did report a dependence. The analysis from this study found that sample size and volume were important factors in mitigating finite volume effects, thereby suggesting areas for future research, such as including the role of halo formation time and the HI-to-baryon mass ratio in galaxy clustering dynamics.

Keywords

Galaxy clustering, Hydrodynamic simulation, Dark halo, Secondary galaxy bias, Two-point correlation function, Mass shuffling, Galaxy formation model, Halo occupation distribution, Halo formation time, Neutral hydrogen to Baryon mass ratio

Zheng Zhou, John L. Miller Great Neck North High School, 35 Polo Road, Great Neck, New York 11023, USA. zhouj048@gmail.com

Introduction

Galaxy formation relating to dark matter halos

Galaxies form in dark matter halos, which due to their undetectable nature are simple to understand when baryons are taken into account (1). Baryons are heavy subatomic particles made up of three quarks, and they exist within a dark matter halo during galaxy formation which results in a contraction of the halo as the galaxy settles into equilibrium (2). In the 1970s, scientists found that galaxies do not form from collapse of dust clouds, but rather from formation of small to large structures in an ascending order (3). Later in the 2000s, studies calculated the collapse of gravitating clouds from an initial smoothed density field (4), which confirmed the above findings through the halo mass function (HMF) (5) that provided the number density of dark matter halos per mass interval. Contrary to expectation, due to the ratio of baryons and dark matter being 1:5, the HMF does not mimic the luminosity function (LF), which provides the number of galaxies per luminosity interval; instead, at lower and higher luminosities than L* galaxies, the slope of the LF is much steeper than that of the HMF (1). This issue was addressed in study by Rees & Ostriker (6), where the researchers used the free-fall timescale t_{ff} defined by $(3\pi/32G\rho_m)^{½} = (3\pi f_b/32G\eta\mu m_H)^{½}$ in terms of the HI gas through the universal baryonic fraction, particle density, and the mean molecular mass. However, an important limitation was that the HI gas does exert pressure (1). Therefore, in order to form stars, the HI gas needs to be of much colder temperature, around 10^2K, than that predicted by a typical halo mass of $10^{12}M_\odot$, which calculates to be ~ 10^4K (1). Subsequently, an overcooling problem was discovered by White & Rees (7) through calculation of a LF, which predicted a high density that allowed the galaxies of ~ $10^{12}M_\odot$ to cool too efficiently, thus predictions of small galaxies became inaccurate. There are three processes that can address the overcooling discrepancy: photoionization, merging, and feedback processes.

Photoionization

During the cosmic "dark ages", prior to the emergence of luminous sources, the universe was predominantly filled with neutral gas, absorbing all light below the ionization threshold at 912 Å (8, 9). Intense UV radiation from the first galaxies created ionized bubbles, permeating and overlapping as they expanded, resulting in a meta-galactic UV background that heated the intergalactic medium and hindered accretion of gas onto halos, impeding star formation, especially in smaller halos (10). The UV background peaked around z ∼ 2, with a characteristic value of $\Gamma \sim 10{-}12$ photons s^{-1} and a photoionization timescale tph ∼ 3 × 10^4 years, coinciding with the cosmic peak of star formation but not directly linked to it (11, 12). Despite observational constraints placing the epoch of reionization between 400–900 million years after the Universe's inception (13), uncertainties persist regarding its duration, contributing sources (small/massive galaxies, quasars), and its topology (small bubbles *vs.* large bubbles). Analytical solutions, such as those by Strömgren (14) and Dopita & Sutherland (15), offer limited constraints, prompting reliance on numerical simulations to capture the nonlinear nature of densities, temperatures, and velocities in this process (16, 17). Challenges arise due to the vast range of spatial scales involved, from sub-parsec scales in molecular clouds to the sizes of ionized bubbles extending to tens of Mpc (18, 19). Simulations often focus on either the escape of ionizing radiation from individual galaxies' interstellar medium or the ionization of the larger scales of the intergalactic medium, with some addressing intermediate scales (20).

Galaxy merging

Even in the contemporary universe, the separation between galaxies remains comparable to their cross-sectional sizes (1). At redshift $z \sim 10$, where number densities were significantly higher (103 times), galaxies experienced frequent gravitational interactions and occasional mergers. Minor mergers, where a galaxy of mass M1 accretes a smaller galaxy of mass $M2 \lesssim M1/10$, occur regularly. Major merger events, with comparable masses ($M1 \sim M2$), take place roughly once per Hubble time (21). This gravitational interaction and merging process, including both minor and major events, contribute to the modification of the galaxy population. Minor mergers, in particular, lead to the accretion of smaller galaxies by larger ones, favoring the growth of large galaxies at the expense of smaller ones. This effect is reflected in the counter-clockwise tilting of the luminosity function (LF), as demonstrated by studies such as those by Cole (22) and Kauffmann et al. (23).

The primary timescale governing the accretion of satellite galaxies is determined by dynamical friction, a physical process first described by Chandrasekhar (24). In this process, the gravity of a massive body moving through a collisionless gas medium (such as the dark matter halo) creates a "wake" of increased mass, which exerts a drag force on the moving body. For minor mergers, the timescale for the orbital decay of a satellite decreases with increasing mass. Cole et al. (25) established that a simple model incorporating this dynamic friction process could provide insight into the orbital evolution of satellite galaxies.

Feedback processes

Early galaxy formation models faced challenges in reproducing observed luminosity functions (LFs), and generated an excess of stars. To address this, initial models, such as those by White & Frenk (26), Somerville & Primack (27), and Efstathiou (28), incorporated only supernova feedback, injecting thermal energy and momentum into the interstellar medium. However, this approach proved too efficient in inhibiting star formation, leading to the expulsion of gas, which then cooled and formed galaxies more massive than observed, particularly in low-mass galaxies (23, 29–31). To mitigate the issue of excessive massive galaxies, incorporation of feedback from active galactic nuclei (AGN) into the models became a promising solution. This mechanism involves gas accreting onto supermassive black holes at the center of galaxies, radiating excess energy at extreme wavelengths. AGN feedback has demonstrated success in aligning with observed LFs at both low and high redshifts (32–35).

Controversy over galaxy clustering dependence on HI mass

In the current field of Astrophysics, there is an ongoing debate on the dependence of galaxy clustering on HI mass. In the study by Guo et al. (36), the researchers used a sample size of ~16000 galaxies in the local universe with redshifts between 0.0025 and 0.05 and with HI mass of $M_{HI} > 10^8 M_\odot$. Using two-point correlation functions as well as cross-correlation functions, the researchers found a result contradictory to previous studies, i.e.

there is a small dependency of galaxy clustering on HI mass for $M_{HI} > 10^9 M_\odot$. With increasing HI mass, there seemed to be more data points on the w_p vs r_p graph, or essentially the result of the two-point correlation function, that possessed significant deviation. In the graph, the amplitude of w_p at ~ 10h−1Mpc for the most massive sample (with $M_{HI} > 10^{10.2} M_\odot$) is a factor of ∼ 4 higher than that of the sample with M_{HI} > 109$M_\odot$ and is comparable to that of the dark matter.

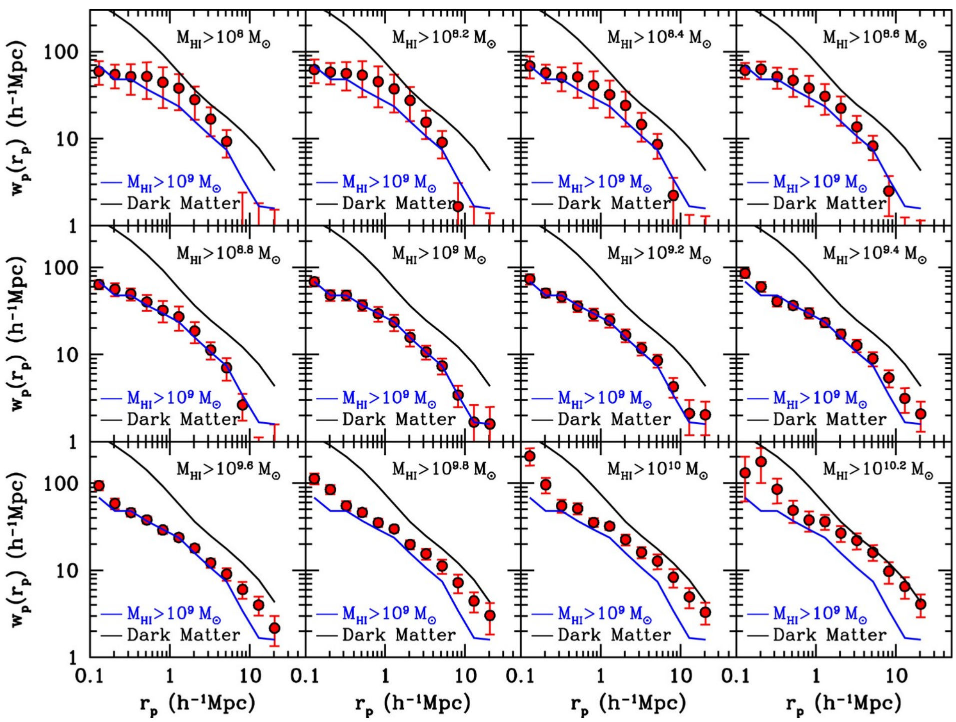

Figure 1. Projected two-point correlation function for 12 of the 13 samples by Guo et al. (36). The measurement of the sample with $M_{HI} > 10^9 M_\odot$ is replotted on each graph for comparison. As clearly shown, more statistically significant deviation from an average level occurs as the HI mass increases above $10^9 M_\odot$. The predicted dark matter halo distribution is shown as the black line.

For a study on the other side of the argument, Papastergis et al. (37) showed that there is no dependence of clustering on HI gas content of a galaxy. Specifically, the researchers used ~6000 samples from the ALFALFA 21cm survey and samples from the SDSS survey. By using projected correlation functions, an analysis of two-dimensional space instead of the three-dimensional space in the two-point correlation function, they found that previous weak clustering for low-HI mass galaxies was most likely due to the finite-volume effect, which is due to statistical uncertainty or an unrepresentative sample. In addition, shown by Figure 2's error bars, the researchers found that galactic HI mass has no significant association with host halo mass as there are no significant error bars for masked HI masses, meaning that HI mass does not affect the clustering of galaxies. Moreover, by relating ALFALFA and SDSS galaxies, the researchers discovered that HI-selected galaxies clustered more weakly when no color selection was applied based on multi-band photometric observations. When color selection was applied, correlation functions of HI-selected and blue (young and low metallicity) galaxies were indistinguishable.

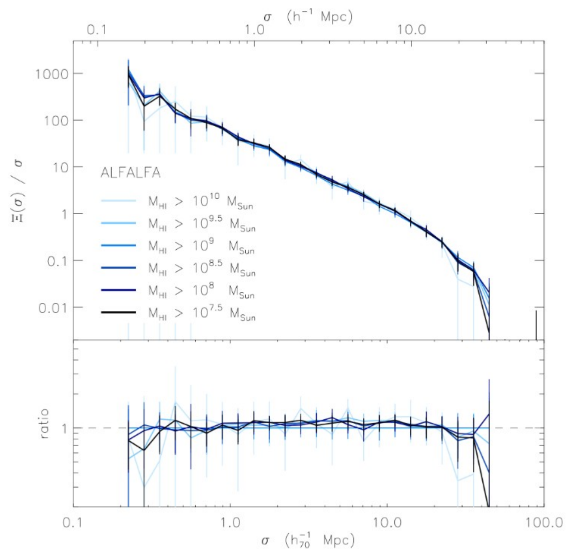

Figure 2. Projected correlation functions of galaxies of HI mass above $10^{7.5}M_{\odot}$ by Papastergis et al. (37). In the lower panel, the same projected correlation functions are normalized to the $M_{HI}>10^{9}M_{\odot}$, with error bars clearly indicated.

Isolating environmental effects on galaxy clustering

The study by Wang et al. (38) employed sophisticated techniques to disentangle the effects of environmental factors on dark halo mass within the cosmological context, extending beyond the conventional focus on halo mass alone. Leveraging the TNG300-1 run of the IllustrisTNG simulation suite (simulation shown in Figure 3), the research utilized the DisPerSE cosmic web finder to identify cosmic structures, including nodes and filaments, providing a nuanced view of the galactic environment. To isolate the influence of environmental factors, particularly those associated with the cosmic web, sophisticated statistical tools such as mass shuffling and double shuffling were employed (39). Mass shuffling involves randomizing galaxy positions among haloes of similar mass, while double shuffling further investigates the role of a secondary halo property, such as proximity to nodes or filaments, in determining galaxy occupation. By systematically comparing original and shuffled galaxy distributions, the study quantifies the strength of the secondary galaxy bias, shedding light on how environmental factors beyond dark halo mass contribute to the spatial distribution and clustering of galaxies in the simulated universe. These rigorous methods enable a more comprehensive understanding of the interplay between galaxy properties and their cosmic surroundings (39).

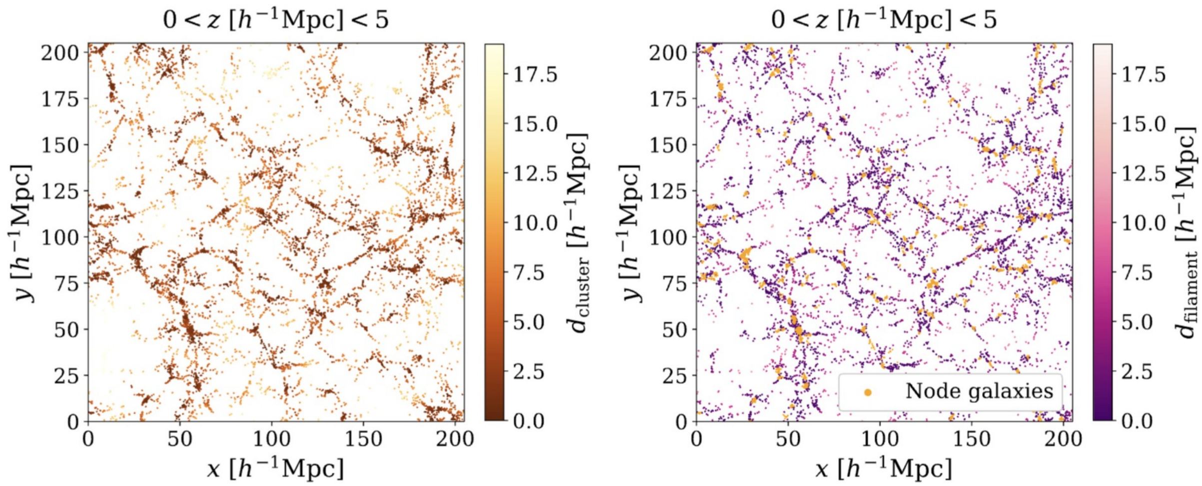

Figure 3. Example of a slice of the hydrodynamic simulation from IllustrisTNG ordered in the z-direction (38). The left panel shows galaxies color coded by distance to the nearest node, *d*node. The right panel shows galaxies color coded by distance to the nearest filament, *d*filament. The node galaxies are noted in both panels.

Objective

In this paper, the controversy over the dependence of galaxy clustering on HI mass will be explored through the masking of the galaxies' HI-mass and shuffling of galaxies based on their dark matter halo properties. The mass shuffling technique will be used to isolate environmental effects on the dark matter halo to ensure a representative sample.

Methods

Data of 4,361,030 galaxies in a local universe (0 < z < 0.1) were obtained from the TNG100 hydrodynamic simulation from the IllustrisTNG project. It was compiled into a hdf5 file, named "hih2_snapshot_099.hdf5", in Google Drive, which was then downloaded onto a Mac laptop. This was loaded and read in the Mac terminal using the command f = h5py.File('mytestfile.hdf5', 'r'). The information in the .hdf5 file included the galaxy's position (SubhaloCM), dark halo masses (SubhaloMassType), and neutral hydrogen masses (m_neutral_H). All position values were converted from kiloparsecs per Hubble constant to megaparsecs per Hubble constant. Dark matter masses were converted to solar masses per Hubble constant.

The two-point correlation function (TPCF) was applied in this study (see Appendix A) with subhalo positions obtained from the TNG100. The TPCF measures the excess probability of finding pairs of objects at a certain separation distance compared to a random, homogeneous distribution. In this context, these subhalos included in the TPCF represent gravitationally bound structures within larger halos, and their spatial distribution was a key aspect of understanding the large-scale structure of the universe. The TPCF was calculated by comparing the observed pair counts at different separation distances to the pair counts expected for a random distribution, effectively quantifying the degree of clustering or avoidance of subhalos at various scales. The TPCF was then computed under two different scenarios: without mass shuffling and with shuffling. These analyses aimed to isolate and negate the impact of the environment - such as the secondary galaxy bias and secondary halo properties - on the correlation function. Double shuffling from Wang et al. (38) was not incorporated in this research for the sake of simplicity.

Mass shuffling of galaxies with identical halo masses in this study was performed using a python function "shuffle_mass_and_pos" (refer to Appendices). The "unique_mass_values" variable contains an array of unique dark matter mass values present in the mass_dm_array, which was obtained using the numpy.unique function. The "mass_indices" variable contains an array of indices that maps each dark matter mass value in mass_dm_array to its corresponding index in the array of unique mass values (unique_mass_values). It was obtained using the return_inverse=True argument in numpy.unique. The purpose of using these unique mass values and mass indices was to group the galaxies by their mass values and then shuffle each group independently. This procedure ensured that the shuffling process preserved the relative ordering of satellite galaxies with respect to the central galaxy within each mass group. In the function, the variable shuffled_indices was created by independently shuffling the indices within each mass group, and then the final shuffled positions and masses were obtained using these shuffled indices.

In addition, the code employed a masking technique for neutral hydrogen masses to create subsamples (see Appendix B). The positions, dark matter masses, and neutral hydrogen masses within these subsamples were then shuffled using the "shuffle_mass_and_pos" function (refer to Appendices). The TPCF was calculated for multiple subsamples (specified by nsamples=100) with a consistent subsample fraction of 0.5 (for both lines on the graph), meaning that for each mask half of the galaxies were taken as samples. error estimates were derived by computing the mean and standard deviation of the TPCF values across the 100 subsamples. The script utilized the Halotools library for mock observables, NumPy for array manipulations, and Matplotlib for visualization. The final visualization, produced by Matplotlib, showcased the mean TPCF of each mask of HI mass as well as the one for galaxies with HI masses $>10^9 M_\odot$, which was run separately and served as a threshold of comparison, with error bars of two standard deviations. Plots of the TPCF of the amplitude of correlation, *xi*, *vs.* separation distance *r* were generated for pairs of galaxies, and the results were interpreted.

Results and Discussion

Secondary galaxy bias

To account for secondary galaxy bias (SGB) mentioned in the paper by Wang et al. (38), a plot was generated to check for the presence of the SGB in the sample. The plot shows the shuffled, original, and random Two-point correlation functions (xi) plotted against distance r in Mpc/h (Figure 4). SGB is the dependence of galaxy occupation on secondary halo properties (38), such as differences in mass accretion rate and the half mass scale, which are important aspects of the formation of dark matter halos. The solution to this is quite simple: to shuffle the properties of galaxies with similar dark halo masses, specifically their neutral hydrogen mass and separation distance from another selected galaxy. Random values for each criteria were also generated to serve as a benchmark for comparison. The generated shuffled TPCF curve almost exactly followed the original, unshuffled TPCF (Figure 4). This implied that there was no SGB present in this sample of galaxies, and that dark halos likely did not affect other properties of the galaxies. Since this study was not solely focused on SGB, double shuffling was not performed. The absence of SGB could be due to selection effects which minimized bias or due to the volume effect, where the sample is representative and covers a significant cosmic volume. This supported the fact that the finite volume effect was not present in this study. The study by Papastergis et al. (37) attributed the dependence of galaxy clustering on neutral hydrogen mass to the finite volume effect.

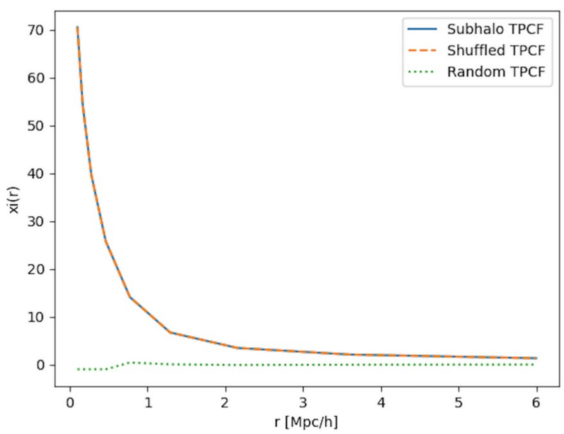

Figure 4. the shuffled Two-point correlation function (xi) plotted against distance r in Mpc/h by author. The graph shows the TPCF for random points, shuffled points, and original points from the TNG100 simulation's galaxies. The axes are in log scale.

Dependence of clustering on HI mass

Figure 5 demonstrates the results produced from the program shown in Appendix B. The 12 TPCF's generated are masked by specific HI mass thresholds, as presented in the figure itself. The y-axis is the value of the calculated shuffled galaxies' TPCF and the x-axis is the separation distance *r*. Error bar estimates of two standard deviations from the mean TPCF value were generated. The measurement of the sample with HI mass $>10^9 M_\odot$ is repeated on every graph for comparison. In contrast to some previous studies that found significant dependence of clustering on HI mass (35), it is evident that the TPCF amplitude of the selected galaxies from the TNG100 does not statistically differ when compared with the measurement of the sample with HI mass $> 10^9 M_\odot$.

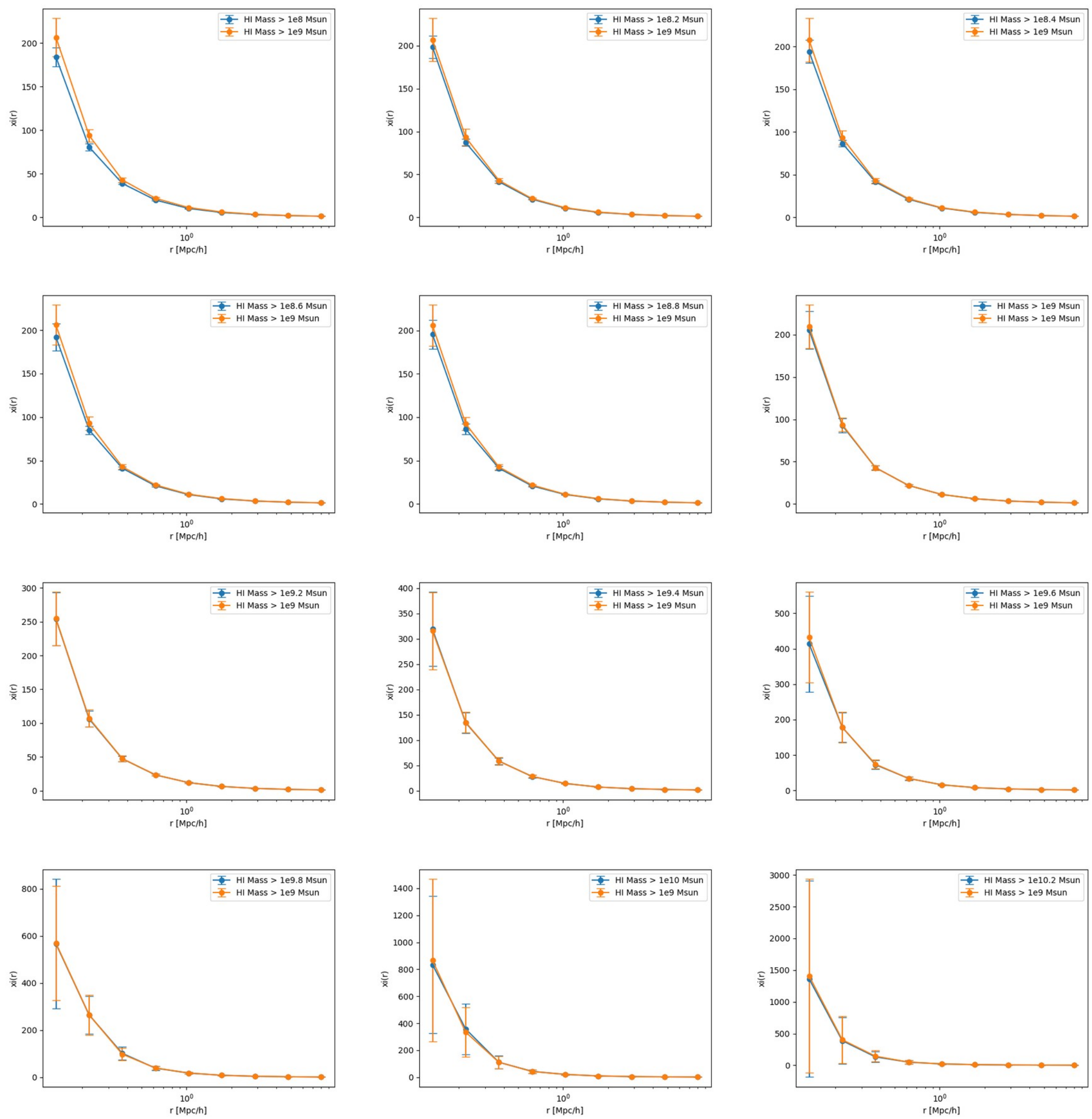

Figure 5. The shuffled Two-point Correlation Functions plotted against distance r in Mpc/h by author. The above plots are masked based on neutral HI mass of each galaxy (colored in blue), ordered in $>10^{8}M_\odot$, $>10^{8.2}M_\odot$, $>10^{8.4}M_\odot$, $>10^{8.6}M_\odot$, $>10^{8.8}M_\odot$, $>10^{9}M_\odot$, $>10^{9.2}M_\odot$, $>10^{9.4}M_\odot$, $>10^{9.6}M_\odot$, $>10^{9.8}M_\odot$, $>10^{10}M_\odot$, $>10^{10.2}M_\odot$, from left to right. For comparison, the measurement of the sample with HI mass $> 10^{9}M_\odot$ is repeated in every graph in orange. Error bars were generated with 2 standard deviations.

The results were in agreement with the paper by Papastergis et al. (37), where the correlation functions do not show significant differences among one another within the error bars generated (see Figure 2). Figure 5 demonstrates no evidence for dependence of clustering on HI mass for galaxies above the HI mass of $10^{8}M_\odot$ because all the curves overlap with the curve of mass threshold above $10^{9}M_\odot$, a reasonably estimated midpoint for galaxy HI mass in the observable universe (39). It is important to note that this study only investigated spiral galaxies and did not include elliptical or dwarf galaxies, which can be potential areas of future research. The results were in contrast with the study by Guo et al. (35), where a relatively small dependence of clustering on HI mass was found (see Figure 1) for larger HI mass thresholds. While there was more variability for larger mass thresholds in this study, the error bars still overlap for the two curves shown in each graph. In addition, central and satellite galaxies were both included in the TPCF calculations. According to Džudžar et al. (39), central late-type (spiral, irregular shape) galaxies have higher gas-to-mass ratios than satellite and isolated galaxies, meaning they have higher HI masses. Therefore, when the HI mass threshold increased, the curves generated became more dependent on properties of central galaxies, which are at the center of dark matter halos, than satellite galaxies. Also, central galaxies generally have higher HI-to-baryon mass ratios. According to Rodriguez et al. (40), when central and satellite galaxies are treated independently, bigger galaxies would be more tightly clustered on smaller scales, but the trend inverts when the two are combined into the same sample. This might help explain the low variability of TPCF values at low HI mass thresholds, as smaller galaxies, which are the majority of the entire sample - since central galaxies can only exist at centers of dark matter halos - cluster in a more consistent pattern that has less variability, even at separation distances above 1 megaparsec. Also, as for how specifically satellite and central galaxies behave differently when viewed independently with regard to HI mass, that can be an area of future research.

As the HI mass threshold increases, there seems to be more overlap in the error bars, which are set to be 2 standard deviations away from the mean TPCF value. Specifically, for the graph with HI mass $>10^{10.2}M_\odot$, the error bars overlap more compared to the graph for lower mass thresholds, such as $>10^{8.4}M_\odot$. This trend can be explained by the decreasing sample size as the mass threshold increases, which makes the results subject to the finite volume effect (37). The effect explains why there is increasing variability, or increase in the size of the error bars, as less galaxies would exist with mass thresholds above $10^{10.2}M_\odot$. The generic spiral galaxies, which are the majority (60%) of the observable universe, have HI masses of ~ $10^{10}M_\odot$ (39). The hydrodynamic simulation does not account for this lack of number of samples, thereby causing the large variability in the graph. Therefore, larger error bars are also more likely to overlap in more regions. The reason why the curve for galaxies $>10^{9}M_\odot$ also changed in value and variability was because a consistent subsample fraction was used based on the number of galaxies above each mass threshold, thus when the curves were generated independently in each graph, the same subsample size was used based on the sample size of the blue curves. The finite volume effect can be mitigated by using a larger sample size, which was not possible in this research due to technological constraints of the calculation software, which was run on a MacBook Air, that is fairly limited in computing capability.

Halo formation time

In this study, halo formation time was not explicitly incorporated as a variable, which was investigated by Guo et al. (35). The decision not to include this parameter was primarily driven by several factors related to the scope and focus of this research, as well as the computational limitations of the simulation framework. Firstly, this study aimed to investigate the dependence of clustering on HI mass for spiral galaxies within a certain range of HI masses. While halo formation time could potentially influence the distribution and properties of galaxies within halos, this study's focus was specifically on HI mass as a factor influencing clustering. Including additional parameters, such as halo formation time, would introduce complexity that could obscure the direct relationship between HI mass and clustering that we sought to explore. Secondly, incorporating halo formation time as a variable would significantly increase the complexity of the computational model. Halo formation time is influenced by a multitude of factors, including the growth history of dark matter halos, merger events, and environmental influences (41). Modeling these processes accurately would require a more sophisticated simulation framework and substantial computational resources, which were beyond the scope and capabilities of this study. Moreover, while Guo et al. (35) suggest that including halo formation time in their model improved the reproduction of clustering measurements, it is important to consider the trade-offs involved. Introducing additional parameters introduces more degrees of freedom into the model, which can increase its flexibility but also its susceptibility to overfitting and the introduction of biases.

HI-Baryon mass ratio

Incorporating the ratio of HI to baryons ($f_{HI} = M_{HI} / M_{Baryon}$) into the analysis provides deeper insights into galaxy clustering and dark matter halo properties. By recalculating the Two-point Correlation Function (TPCF) for galaxies segmented by f_{HI} thresholds, we can compare these results with those obtained using HI mass thresholds. This approach can reveal if clustering properties correlate more strongly with the HI fraction relative to the total baryonic mass. Preliminary analysis suggests that galaxies with higher f_{HI} ratios, typically found in less massive halos, may exhibit different clustering patterns compared to those with lower f_{HI}, often residing in massive halos with dense environments. This aligns with studies by Duffy et al. (42) and Lagos et al. (43), which showed that higher f_{HI} ratios are associated with lower mass halos and that f_{HI} can influence inferred halo properties such as mass and concentration. Thus, considering f_{HI} offers a refined perspective on the relationship between gas content, galaxy clustering, and halo characteristics. Future research should focus on detailed TPCF recalculations with f_{HI} bins, examining environmental factors, and using larger simulations to mitigate finite volume effects, thereby enhancing the understanding of galaxy clustering dynamics.

Halo occupation distribution

The Halo Occupation Distribution (HOD) is a theoretical framework used in cosmology to describe how galaxies populate dark matter halos. It quantifies the probability distribution of the number of galaxies within a halo of a given mass. HOD models typically distinguish between central and satellite galaxies, providing insights into how galaxy properties correlate with halo properties. The implication of Paul et al.'s (44) finding that many low-mass HI galaxies are satellites affects the interpretation of the simulation results, particularly the TPCF analyses. In this study, both central and satellite galaxies were included in the TPCF calculations. According to the results, there was no significant dependence of clustering on HI mass for galaxies above $10^8 M_\odot$. However, considering the HOD approach and the significant presence of low-mass satellites, we can infer that the clustering signal might be influenced by the spatial distribution and the clustering behavior of these satellite galaxies (44). Satellites tend to reside in denser environments, such as galaxy clusters, where they can experience enhanced gravitational interactions, leading to different clustering properties compared to central galaxies. This higher number of low-mass satellites could dilute the clustering signal attributed to more massive galaxies. Thus, the inclusion of a large population of low-mass satellite galaxies in the sample might explain the observed lack of significant variation in TPCF amplitude across different HI mass thresholds.

Moreover, the findings suggest that the HI-to-baryon ratio could further enhance understanding in this area by highlighting how these low-mass satellites contribute to the overall gas content and its distribution within dark matter halos. Future analyses could focus on separating central and satellite populations in the TPCF calculations or recalculating TPCFs for different HI-to-baryon mass ratio bins to account for the distinct clustering behaviors of these sub-populations. This approach will allow for a more precise understanding of the impact of satellite galaxies on the overall clustering pattern and improve the interpretation of simulation results in the context of dark matter halo properties and galaxy formation models.

Conclusion

In conclusion, this study comprehensively examined the relationship between galaxy clustering and HI mass using data from the TNG100 simulation. Through analysis of TPCF and isolation of environmental factors, the research demonstrated negligible clustering dependence on HI mass above $10^8 M_\odot$. These findings contribute to our understanding of large-scale structure formation in the universe and underscore the importance of sample size and volume in observational studies. Looking ahead, future investigations should explore additional variables such as halo formation time and the HI-to-baryon mass ratio to deepen our insight into galaxy clustering dynamics. Overall, this study provides valuable insight into the complex relationship between galaxy properties and their spatial distribution within dark matter halos.

References

1. Laursen, P. (2023). Galaxy formation from a timescale perspective. https://doi.org/10.48550/arXiv.2309.02486

2. Li, P., McGaugh, S. S., Lelli, F., Schombert, J. M., Pawlowski, M. S. (2022). Incorporating baryon-driven contraction of dark matter halos in rotation curve fits. *Astronomy & Astrophysics*, 665, A143. https://doi.org/10.1051/0004-6361/202243916

3. Press, W. H., Schechter, P. (1974). Formation of galaxies and clusters of galaxies by self-similar gravitational condensation. *Astrophysical Journal,* 187, 425-438. https://doi.org/10.1086/152650

4. Tinker, J., Kravtsov, A. V., Klypin, A., Abazajian, K., Warren, M., Yepes, G., et al. (2008). Toward a halo mass function for precision cosmology: The Limits of universality. *The Astrophysical Journal*, 688(2), 709. https://doi.org/10.48550/arXiv.0803.2706

5. Lukić, Z., Heitmann, K., Habib, S., Bashinsky, S., Ricker, P. M. (2007). The halo mass function: High-Redshift evolution and universality. *The Astrophysical Journal*, 671(2), 1160. https://iopscience.iop.org/article/10.1086/523083

6. Rees, M. J., Ostriker, J. P. (1977). Cooling, dynamics and fragmentation of massive gas clouds: clues to the masses and radii of galaxies and clusters. *Monthly Notices of the Royal Astronomical Society*, 179(4), 541-559. https://doi.org/10.1093/mnras/179.4.541

7. White, S. D., Rees, M. J. (1978). Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering. *Monthly Notices of the Royal Astronomical Society*, 183(3), 341-358. https://ui.adsabs.harvard.edu/link_gateway/1978MNRAS.183..341W/doi:10.1093/mnras/183.3.341

8. Gnedin, N. Y. (2000). Cosmological reionization by stellar sources. *The Astrophysical Journal*, 535(2), 530. https://doi.org/10.48550/arXiv.astro-ph/9909383

9. Barkana, R., Loeb, A. (2001). In the beginning: the first sources of light and the reionization of the universe. *Physics reports*, 349(2), 125-238. https://doi.org/10.48550/arXiv.astro-ph/0010468

10. Benson, A. J., Lacey, C. G., Baugh, C. M., Cole, S., Frenk, C. S. (2002). The effects of photoionization on galaxy formation–I. Model and results at z= 0. *Monthly Notices of the Royal Astronomical Society*, 333(1), 156-176. https://doi.org/10.1046/j.1365-8711.2002.05387.x

11. Haardt, F., Madau, P. (2012). Radiative transfer in a clumpy universe. IV. New synthesis models of the cosmic UV/X-ray background. *The Astrophysical Journal*, 746(2), 125. https://doi.org/10.48550/arXiv.astro-ph/9509093

12. Khaire, V., & Srianand, R. (2019). New synthesis models of consistent extragalactic background light over cosmic time. *Monthly Notices of the Royal Astronomical Society*, 484(3), 4174-4199. https://doi.org/10.1093/mnras/stz174

13. Mason, C. A., Fontana, A., Treu, T., Schmidt, K. B., Hoag, A., Abramson, L., et al. (2019). Inferences on the timeline of reionization at z∼ 8 from the KMOS Lens-Amplified Spectroscopic Survey. *Monthly Notices of the Royal Astronomical Society*, 485(3), 3947-3969. https://doi.org/10.1093/mnras/stz632

14. Strömgren, B. (1979). The physical state of interstellar hydrogen. In *A Source Book in Astronomy and Astrophysics, 1900–1975* (pp. 588-592). Harvard University Press. https://doi.org/10.4159/harvard.9780674366688.c93

15. Dopita, M. A., Sutherland, R. S. (2004). *Astrophysics of the diffuse universe*. Springer Science & Business Media. https://ui.adsabs.harvard.edu/link_gateway/2003adu..book.....D/doi:10.1007/978-3-662-05866-4

16. Hutter, A., Watkinson, C. A., Seiler, J., Dayal, P., Sinha, M., Croton, D. J. (2020). The 21 cm bispectrum during reionization: a tracer of the ionization topology. *Monthly Notices of the Royal Astronomical Society*, 492(1), 653-667. https://doi.org/10.48550/arXiv.1907.04342

17. Perez, L. A., Malhotra, S., Rhoads, J. E., Laursen, P., Wold, I. G. (2022). Probing Patchy Reionization with the Void Probability Function of Lyα Emitters. *The Astrophysical Journal*, 940(2), 102. https://doi.org/10.48550/arXiv.2210.05722

18. Wyithe, J. S. B., Loeb, A. (2004). A characteristic size of∼ 10 Mpc for the ionized bubbles at the end of cosmic reionization. *Nature*, 432(7014), 194-196. https://doi.org/10.48550/arXiv.astro-ph/0409412

19. Giri, S. K., Mellema, G., Dixon, K. L., Iliev, I. T. (2018). Bubble size statistics during reionization from 21-cm tomography. *Monthly Notices of the Royal Astronomical Society*, 473(3), 2949-2964. https://doi.org/10.1093/mnras/stx2539

20. Rosdahl, J., Blaizot, J., Katz, H., Kimm, T., Garel, T., Haehnelt, M., et al. (2022). LyC escape from SPHINX galaxies in the Epoch of Reionization. *Monthly Notices of the Royal Astronomical Society*, 515(2), 2386-2414. https://doi.org/10.48550/arXiv.2207.03232

21. Lacey, C., Cole, S. (1993). Merger rates in hierarchical models of galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 262(3), 627-649. https://doi.org/10.48550/arXiv.astro-ph/9402069

22. Cole, S. (1991). Modeling galaxy formation in evolving dark matter halos. *The Astrophysical Journal*, 367, 45-53. http://dx.doi.org/10.1086/169600

23. Kauffmann, G., White, S. D., Guiderdoni, B. (1993). The formation and evolution of galaxies within merging dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 264(1), 201-218. https://ui.adsabs.harvard.edu/link_gateway/1993MNRAS.264..201K/doi:10.1093/mnras/264.1.201

24. Chandrasekhar, S. (1943). Dynamical friction. I. General considerations: the coefficient of dynamical friction. *Astrophysical Journal*, 97, 255-262. https://ui.adsabs.harvard.edu/link_gateway/1943ApJ....97..255C/doi:10.1086/144517

25. White, S. D., Frenk, C. S. (1991). Galaxy formation through hierarchical clustering. *Astrophysical Journal, Part 1,* 379, 52-79. https://ui.adsabs.harvard.edu/link_gateway/1991ApJ...379...52W/doi:10.1086/170483

26. Somerville, R. S., Primack, J. R. (1999). Semi-analytic modelling of galaxy formation: the local Universe. *Monthly Notices of the Royal Astronomical Society*, 310(4), 1087-1110. https://doi.org/10.1046/j.1365-8711.1999.03032.x

27. Efstathiou, G. (2000). A model of supernova feedback in galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 317(3), 697-719. https://doi.org/10.1046/j.1365-8711.2000.03665.x

28. Dekel, A., Silk, J. (1986). The origin of dwarf galaxies, cold dark matter, and biased galaxy formation. *Astrophysical Journal, Part 1,* 303, 39-55. https://ui.adsabs.harvard.edu/link_gateway/1986ApJ...303...39D/doi:10.1086/164050

29. Natarajan, P. (1999). Consequences of feedback from early supernovae for disk assembly. *The Astrophysical Journal*, 512(2), L105. https://iopscience.iop.org/article/10.1086/311877

30. Benson, A. J., Bower, R. G., Frenk, C. S., Lacey, C. G., Baugh, C. M., Cole, S. (2003). What shapes the luminosity function of galaxies?. *The Astrophysical Journal*, 599(1), 38. https://doi.org/10.48550/arXiv.astro-ph/0302450

31. Croton, D. J., Springel, V., White, S. D., De Lucia, G., Frenk, C. S., Gao, L., et al. (2006). The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies. *Monthly Notices of the Royal Astronomical Society*, 365(1), 11-28. https://doi.org/10.1111/j.1365-2966.2005.09675.x

32. Bower, R. G., Benson, A. J., Malbon, R., Helly, J. C., Frenk, C. S., Baugh, C. M., et al. (2006). Breaking the hierarchy of galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 370(2), 645-655. https://doi.org/10.1111/j.1365-2966.2006.10519.x

33. Somerville, R. S., Hopkins, P. F., Cox, T. J., Robertson, B. E., Hernquist, L. (2008). A semi-analytic model for the co-evolution of galaxies, black holes and active galactic nuclei. *Monthly Notices of the Royal Astronomical Society*, 391(2), 481-506. https://doi.org/10.1111/j.1365-2966.2008.13805.x

34. McCarthy, I. G., Schaye, J., Bower, R. G., Ponman, T. J., Booth, C. M., Vecchia, C. D., Springel, V. (2011). Gas expulsion by quasar-driven winds as a solution to the overcooling problem in galaxy groups and clusters. *Monthly Notices of the Royal Astronomical Society*, 412(3), 1965-1984. https://doi.org/10.1111/j.1365-2966.2010.18033.x

35. Guo, H., Li, C., Zheng, Z., Mo, H. J., Jing, Y. P., Zu, Y., et al. (2017). Constraining the Hi–Halo Mass Relation from Galaxy Clustering. *The Astrophysical Journal*, 846(1), 61. https://iopscience.iop.org/article/10.3847/1538-4357/aa85e7

36. Papastergis, E., Giovanelli, R., Haynes, M. P., Rodríguez-Puebla, A., Jones, M. G. (2013). The Clustering of Alfalfa Galaxies: Dependence on Hi mass, relationship with optical samples, and clues of host Halo properties. *The Astrophysical Journal*, 776(1), 43. https://doi.org/10.48550/arXiv.1308.2661

37. Wang, K., Avestruz, C., Guo, H., Wang, W., Wang, P. (2023). The Beyond-Halo Mass Effects of the Cosmic Web Environment on Galaxies. *arXiv preprint arXiv:2309.15306*. https://doi.org/10.48550/arXiv.2309.15306

38. Croton, D. J., Gao, L., White, S. D. (2007). Halo assembly bias and its effects on galaxy clustering. *Monthly Notices of the Royal Astronomical Society*, 374(4), 1303-1309. https://doi.org/10.48550/arXiv.astro-ph/0605636

39. Džudžar1, R. Kilborn, V., Meurer, G., Sweet, S. M., Drinkwater, M., Bekki, K., et al. (2018) *The neutral hydrogen properties of galaxies in gas-rich groups*. ar5iv. https://ar5iv.labs.arxiv.org/html/1812.08302

40. Rodriguez, F., Montero-Dorta, A. D., Angulo, R. E., Artale, M. C., Merchán, M. (2020). *The scaling relations and clustering properties of central and Satellite Galaxies*. arXiv.org. https://arxiv.org/abs/2011.00014v1

41. Tojeiro, R., Eardley, E., Peacock, J. A., Norberg, P., Alpaslan, M., Driver, S. P., et al. (2017). *Galaxy and Mass Assembly (gama): Halo formation times and Halo Assembly bias on the cosmic web*. OUP Academic. https://academic.oup.com/mnras/article/470/3/3720/3868210

42. Duffy, A. R., Kay, S. T., Battye, R. A., Booth, C. M., Dalla Vecchia, C., Schaye, J. (2012). *Modeling neutral hydrogen in galaxies using cosmological hydrodynamical simulations*. OUP Academic. https://academic.oup.com/mnras/article/420/4/2799/971469

43. Lagos, C. del P., Davis, T. A., Lacey, C. G., Zwaan, M. A., Baugh, C. M., Gonzalez-Perez, V., Padilla, N. D. (2014). *The origin of the atomic and molecular gas contents of early-type galaxies – I. A new test of Galaxy Formation Physics*. OUP Academic. https://academic.oup.com/mnras/article/443/2/1002/1063103

44. Paul, N., Choudhury, T. R., Paranjape, A. (2018). *Halo models of HI selected galaxies*. arXiv.org. https://arxiv.org/abs/1712.04469

Appendix A: *Program in python for generation of Figure 4*

```
import numpy as np
from h5py import File
import matplotlib.pyplot as plt
from halotools.mock_observables import tpcf

def shuffle_mass_and_pos(pos_array, mass_dm_array, mass_hi_array):
    unique_mass_values, mass_indices = np.unique(mass_dm_array, return_inverse=True)

    shuffled_indices = [np.random.permutation(np.where(mass_indices == i)[0]) for i in
range(len(unique_mass_values))]

    shuffled_indices = np.concatenate(shuffled_indices)

    shuffled_pos = pos_array[shuffled_indices]
    shuffled_mass_dm = mass_dm_array[shuffled_indices]
    shuffled_mass_hi = mass_hi_array[shuffled_indices]

    return shuffled_pos, shuffled_mass_dm, shuffled_mass_hi

boxsize = 100
samples = 1000

x = np.random.uniform(0, boxsize, samples)
y = np.random.uniform(0, boxsize, samples)
z = np.random.uniform(0, boxsize, samples)
```

```
pos_random = np.vstack([x, y, z]).T

with File("hih2_snapshot_099.hdf5", "r") as f:
    pos = f["SubhaloCM"][:] * 1e-3
    mass_dm = f["SubhaloMassType"][:, 1] * 1e10
    mass_hi = f["m_neutral_H"][:]  # Already in Msun/h

# Mask galaxies with HI masses greater than 10^8 solar masses
mask = mass_hi > 1e8
pos = pos[mask]
mass_dm = mass_dm[mask]
mass_hi = mass_hi[mask]

# Shuffle positions, dark matter masses, and neutral hydrogen masses
shuffled_pos, shuffled_mass_dm, shuffled_mass_hi = shuffle_mass_and_pos(pos, mass_dm,
mass_hi)

# Calculate TPCF for subhalo positions
rbins = np.logspace(-1, 1, 10)
xi = tpcf(pos, rbins, period=boxsize)

# Calculate TPCF for shuffled positions
xi_shuffled = tpcf(shuffled_pos, rbins, period=boxsize)

# Calculate TPCF for random positions
xi_random = tpcf(pos_random, rbins, period=boxsize)

# Plot the results
plt.plot(rbins[:-1], xi, label="Subhalo TPCF")
plt.plot(rbins[:-1], xi_shuffled, label="Shuffled TPCF", linestyle="--")
plt.plot(rbins[:-1], xi_random, label="Random TPCF", linestyle=":")
plt.xlabel("r [Mpc/h]")
plt.ylabel("xi(r)")
plt.legend()

# Save figures with correct quotes
plt.savefig('subhalo_tpcf.png')
plt.savefig('shuffled_tpcf.png')
plt.savefig('random_tpcf.png')

plt.show()
```

Appendix B: *Program in python for generation of Figure 5*

```
import numpy as np
from h5py import File
import matplotlib.pyplot as plt
from halotools.mock_observables import tpcf

def shuffle_mass_and_pos(pos_array, mass_dm_array, mass_hi_array):
    unique_mass_values, mass_indices = np.unique(mass_dm_array, return_inverse=True)

    shuffled_indices = [np.random.permutation(np.where(mass_indices == i)[0]) for i in
range(len(unique_mass_values))]

    shuffled_indices = np.concatenate(shuffled_indices)

    shuffled_pos = pos_array[shuffled_indices]
    shuffled_mass_dm = mass_dm_array[shuffled_indices]
    shuffled_mass_hi = mass_hi_array[shuffled_indices]

    return shuffled_pos, shuffled_mass_dm, shuffled_mass_hi

def calculate_tpcf_and_error(pos_array, mass_dm_array, mass_hi_array, nsamples=100,
subsample_fraction=0.5):
    tpcf_samples = []

    for _ in range(nsamples):
        # Create a subsample by randomly selecting a fraction of galaxies with HI masses above the
specified limit
        mask = (mass_hi_array > 1e8) & (np.random.rand(len(mass_hi_array)) <
subsample_fraction)
        subsample_pos = pos_array[mask]
        subsample_mass_dm = mass_dm_array[mask]
        subsample_mass_hi = mass_hi_array[mask]

        # Shuffle positions, dark matter masses, and neutral hydrogen masses for the subsample
        shuffled_pos, shuffled_mass_dm, shuffled_mass_hi = shuffle_mass_and_pos(
            subsample_pos, subsample_mass_dm, subsample_mass_hi)

        # Calculate TPCF for the subsample
        rbins = np.logspace(-1, 1, 10)
        xi_sample = tpcf(subsample_pos, rbins, period=boxsize)
```

```
        tpcf_samples.append(xi_sample)

    # Calculate mean and standard deviation of TPCF over the subsamples
    mean_tpcf = np.mean(tpcf_samples, axis=0)
    std_tpcf = np.std(tpcf_samples, axis=0)

    # Calculate bin centers
    bin_centers = 0.5 * (rbins[:-1] + rbins[1:])

    return bin_centers, mean_tpcf, std_tpcf

boxsize = 100
samples = 1000
nsamples = 100

x = np.random.uniform(0, boxsize, samples)
y = np.random.uniform(0, boxsize, samples)
z = np.random.uniform(0, boxsize, samples)

pos_random = np.vstack([x, y, z]).T

with File("hih2_snapshot_099.hdf5", "r") as f:
    pos = f["SubhaloCM"][:] * 1e-3
    mass_dm = f["SubhaloMassType"][:, 1] * 1e10
    mass_hi = f["m_neutral_H"][:]  # Already in Msun/h

# Calculate mean TPCF and error bars over 100 subsamples for HI masses > 1e8 solar masses
bin_centers_low, mean_tpcf_low, std_tpcf_low = calculate_tpcf_and_error(pos, mass_dm,
mass_hi, nsamples)

# Create a mask for galaxies with HI masses above 1e9 solar masses
mask_high = mass_hi > 1e9

# Calculate mean TPCF and error bars over 100 subsamples for HI masses > 1e9 solar masses
bin_centers_high, mean_tpcf_high, std_tpcf_high = calculate_tpcf_and_error(pos[mask_high],
mass_dm[mask_high], mass_hi[mask_high], nsamples)

# Plot the results for both mass thresholds on the same graph
plt.errorbar(bin_centers_low, mean_tpcf_low, yerr=2*std_tpcf_low, label="HI Mass > 1e8
Msun", fmt='o-', capsize=5)
```

```
plt.errorbar(bin_centers_high, mean_tpcf_high, yerr=2*std_tpcf_high, label="HI Mass > 1e9
Msun", fmt='o-', capsize=5)

plt.xscale('log')
plt.xlabel("r [Mpc/h]")
plt.ylabel("xi(r)")
plt.legend()
plt.savefig('mean_tpcf_with_errorbars.png')
plt.show()
```