Learning the galaxy-environment connection with graph neural networks

John F. Wu [1 2] **Christian Kragh Jespersen** [3]

Abstract

Galaxies co-evolve with their host dark matter halos. Models of the galaxy-halo connection, calibrated using cosmological hydrodynamic simulations, can be used to populate dark matter halo catalogs with galaxies. We present a new method for inferring baryonic properties from dark matter subhalo properties using message-passing graph neural networks (GNNs). After training on subhalo catalog data from the Illustris TNG300-1 hydrodynamic simulation, our GNN can infer stellar mass from the host and neighboring subhalo positions, kinematics, masses, and maximum circular velocities. We find that GNNs can also robustly estimate stellar mass from subhalo properties in $2d$ projection. While other methods typically model the galaxy-halo connection in isolation, our GNN incorporates information from galaxy environments, leading to more accurate stellar mass inference.

1. Introduction

In the current ΛCDM paradigm of hierarchical galaxy formation, the galaxy-halo connection is crucial for understanding how galaxies form and evolve, and for constraining the small-scale clustering of matter (Somerville & Davé, 2015; Wechsler & Tinker, 2018; Vogelsberger et al., 2020). Techniques for modeling the co-evolution of galaxies and dark matter range from simple, non-parametric approaches to full-physics magnetohydrodynamic simulations which require $> 10^8$ CPU hours of computation (e.g., Vale & Ostriker, 2004; Pillepich et al., 2018). Detailed simulations contribute important insights into galaxy formation, but due to their complexity and heavy computational costs, they are hard to analyze and cannot be performed for cosmologically significant volumes. Machine learning (ML) is a natural option for making progress on both of these problems.

We present an equivariant Graph Neural Network (GNN), which takes as its input a graph composed of halos linked on a linking scale of 5 Mpc, and predicts baryonic properties. The GNN incorporates the effects of a galaxy's environment, thereby improving the prediction of its baryonic properties compared to traditional methods. We are also able to train a network on the Illustris TNG300-1 box in 10 minutes on a single NVIDIA A10G GPU; inference takes one second. In this work, we focus on estimating stellar mass from a catalog of subhalo positions, velocities, $M_{\rm halo}$, and $V_{\rm max}$.

2. Related work

The connection between galaxies and their dark matter halos has been characterized via abundance matching or halo occupation distribution (HOD) models of central halos (Berlind & Weinberg, 2002; Wechsler et al., 2002), conditional luminosity or mass functions (Yang et al., 2003; Moster et al., 2010), subhalo abundance matching (Kravtsov et al., 2004; Vale & Ostriker, 2004; Conroy et al., 2006), and empirical models of the galaxy-halo connection (e.g., Reddick et al., 2013; Behroozi et al., 2019). Several works have also attempted to perform abundance matching or paint baryons (i.e., stars) onto dark matter maps by using classical machine learning algorithms (e.g., Kamdar et al., 2016; Agarwal et al., 2018; Calderon & Berlind, 2019) and/or neural networks (e.g., Zhang et al., 2019; Moster et al., 2021; Mohammad et al., 2022).

In general, these previous methods treat halo/galaxy systems as unrelated entities with no formation history. To rectify this, Villanueva-Domingo et al. (2022) construct mathematical graphs to represent group halos, and train a GNN to learn the central halo mass, which was later applied to estimate the halo masses of local Group galaxies (Villanueva-Domingo et al., 2021). GNNs have also been successfully used to model the dependence of galaxy properties on merger history (e.g., Jespersen et al., 2022; Tang & Ting, 2022), and generate synthetic galaxy catalogs (Jagvaral et al., 2022).

In cosmology, several works have already demonstrated the representational power of GNNs, and have used it for simulation-based inference (likelihood-free inference).

[1]Space Telescope Science Institute, 3700 San Martin Dr, Baltimore, MD 21218 [2]Johns Hopkins University, 3400 N. Charles St, Baltimore, MD 21218 [3]Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA. Correspondence to: John F. Wu <jowu@stsci.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

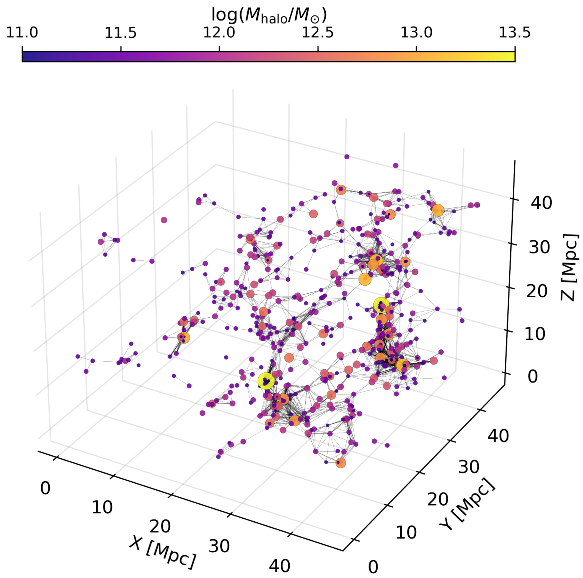

Figure 1. Cosmic graph of a TNG300 subvolume spanning approximately 45 Mpc. Subhalos live on nodes and are colored by their logarithmic subhalo mass. Edges are formed between pairs of subhalos separated by less than the linking length of 5 Mpc.

Villanueva-Domingo & Villaescusa-Navarro (2022) employ GNNs to infer the cosmological parameters Ω_m and σ_8, using $3d$ galaxy positions and stellar properties from the CAMELS simulation suite (Villaescusa-Navarro et al., 2021). Makinen et al. (2022) show that GNNs can optimally extract and compress catalog data for cosmological parameter inference. Shao et al. (2023) and de Santi et al. (2023) train GNNs to infer cosmological parameters from dark matter-only simulations, and then validate their robustness on other N-body and hydrodynamic simulations.

3. Cosmic graphs

3.1. Simulation data

We use `SUBFIND` $z = 0$ subhalo catalogs (Springel et al., 2001) derived from the Illustris TNG300-1 hydrodynamic simulation (Nelson et al., 2019b; Pillepich et al., 2019). We split the full cosmological box into $6^3 = 216$ subvolumes in order to fit into 16 GB of memory, such that each subvolume is about $(50\ \mathrm{Mpc})^3$. For consistency with the TNG simulations, we adopt the Planck Collaboration et al. (2016) cosmology and set $H_0 = 67.74\ \mathrm{km\ s^{-1}\ Mpc^{-1}}$.

We select unflagged subhalos that have more than 50 star particles, $\log(M_\star/M_\odot) > 9$, and $\log(M_{\rm halo}/M_\odot) > 10$. Due to cosmic variance, some subvolumes only have a few hundred subhalos, while others have thousands. In Figure 1, we show an example of a typical subvolume.

3.2. Equivariant graph neural networks

We construct a mathematical graph for each TNG300 subvolume, such as the one depicted in Figure 1. We designate $\mathcal{V}_i = (\boldsymbol{x}_i, \boldsymbol{v}_i, M_{\mathrm{halo},i}, V_{\mathrm{max},i})$ as the eight node features. Subhalos within a linking length of $L = 5$ Mpc are connected with edges. Subvolumes are padded by 2.5 Mpc on each side, such that subvolumes do not share connections that would be relevant for the linking length. We allow nodes to be connected to themselves (i.e., self-loops). On each edge $\mathcal{E}_{ij}$, we compute three features: the squared Euclidean distance $d_{ij} \equiv ||\boldsymbol{x}_i - \boldsymbol{x}_j||$, the inner product between unit vectors $\boldsymbol{e}_i \cdot \boldsymbol{e}_j$, and the inner product between unit vectors $\boldsymbol{e}_i \cdot \boldsymbol{e}_{i-j}$, where unit vectors $\boldsymbol{e}_i \equiv (\boldsymbol{x}_i - \bar{\boldsymbol{x}})/||\boldsymbol{x}_i - \bar{\boldsymbol{x}}||)$ are defined using positions $\boldsymbol{x}_i$ relative to the centroid of the point cloud distribution $\bar{\boldsymbol{x}}$, and $\boldsymbol{e}_{i-j}$ is the unit vector in the direction of $\boldsymbol{x}_i - \boldsymbol{x}_j$.

We use a message-passing GNN based on interaction networks (Battaglia et al., 2016; 2018), similar to the model used by Villanueva-Domingo et al. (2022). By design, the GNN is equivariant to permutations and invariant under the $E(3)$ group action, i.e., invariant to rotations, reflections, and translations. For more details about equivariant GNNs, see the appendices of Garcia Satorras et al. (2021) and Sections 3.1 and 3.2 of Villanueva-Domingo & Villaescusa-Navarro (2022). We aggregate layer inputs at each node by max pooling over information from neighboring nodes.[1] Our GNN has one set of fully connected layers with 256 latent channels and 128 hidden channels. We predict two quantities for each node, which correspond to the logarithmic stellar mass $y_i \equiv \log(M_{\star,i}/M_\odot)$ and the logarithmic variance, $\log \Sigma_i$ (i.e., the logarithm of the squared uncertainty on stellar mass).

3.3. Optimization

Our loss function is composed of two terms: the mean squared error on the logarithmic stellar mass $||\hat{\boldsymbol{y}} - \boldsymbol{y}||^2$, and the squared difference between the predicted and measured variance $||\hat{\boldsymbol{\Sigma}} - (\hat{\boldsymbol{y}} - \boldsymbol{y})^2||^2$. The latter term ensures that the variance is appropriately estimated (see Moment Networks, described in Section 2 of Jeffrey & Wandelt, 2020). We stabilize training by taking the logarithm of each loss term before summing them. We monitor the loss as well as the root mean squared error (RMSE) on $\log(M_\star/M_\odot)$.

We perform $k = 6$-fold cross-validation. For each fold, we train on 180 subvolumes and validate on 36 subvolumes, such that the validation set forms a $\sim 50 \times 300 \times 300\ \mathrm{Mpc}^3$ subbox. We augment the training data set by adding random noise, sampled from a normal distribution with 10^{-5} times

[1]We do not find significant improvements by using a concatenation of sum, max, mean, and variance aggregations, or by using learnable aggregation functions.

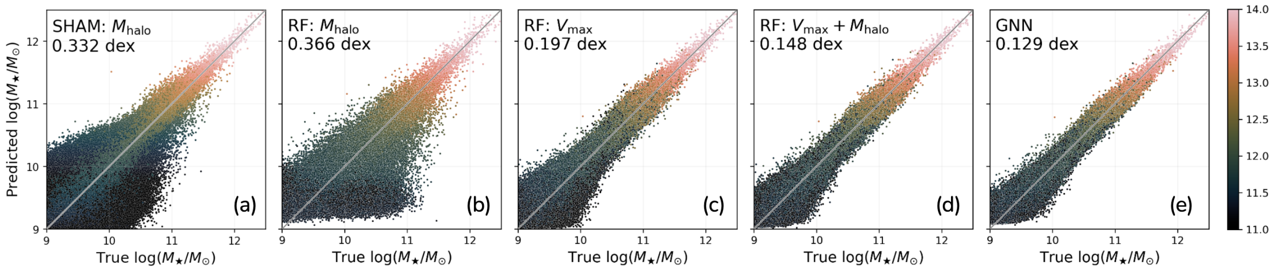

Figure 2. Predicted stellar mass versus true stellar mass for the TNG300 data. From left to right, we show results for a subhalo abundance matching (SHAM) model, three random forest (RF) models, and our $3d$ GNN trained using $\boldsymbol{x}$, $\boldsymbol{v}$, $M_{\rm halo}$, and $V_{\rm max}$. We also report the scatter in the reconstructed $\log(M_\star/M_\odot)$.

the standard deviation, for each node variable. Based on a preliminary hyperparameter search, we implement a simple optimization schedule over a total of 1000 epochs using the `AdamW` optimizer (Kingma & Ba, 2014; Loshchilov & Hutter, 2017) and a batch size of 36. We begin with a learning rate of 10^{-2} and weight decay of 10^{-4}, and then decrease both by a factor of 5 at 500 epochs, and again decrease both by a factor of 5 at 750 epochs. We inspect the training and validation losses to ensure that the optimization is converged and does not overfit the training data.

4. Results

Overall, we find that the GNN can infer the stellar mass from subhalo properties with remarkable accuracy. We recover the galaxy stellar mass to within RMSE $= 0.129$ dex of its simulated value by using a GNN. The predictions are largely unbiased as a function of mass.

4.1. Comparisons against baseline models

In Figure 2, we compare the performance of different models trained and cross-validated on the same TNG300 data set. The panels show, from left to right: (a) a subhalo abundance matching (SHAM) model, (b) a random forest (RF) trained using $M_{\rm halo}$ as input, (c) a RF trained using $V_{\rm max}$, (d) a RF trained using both $M_{\rm halo}$ and $V_{\rm max}$, and (e) a GNN trained using $3d$ positions, $3d$ velocities, $M_{\rm halo}$, and $V_{\rm max}$. In Table 1, we list performance metrics for various RF and GNN models, including the RMSE, mean average error (MAE), normalized median absolute deviation (NMAD),[2] Pearson correlation coefficient (ρ), correlation of determination (R^2), bias, and outlier fraction ($> 3\times$ NMAD).

The SHAM model constructs separate monotonic relationships between $M_{\rm halo}$ or $V_{\rm max}$ and $M_\star$ for centrals and satellites. Another difference between the SHAM model and other approaches considered here is the former's explicit treatment of subhalo centrality. In order to facilitate an apples-to-apples comparison, we also train an abundance matching (AM) model that does not distinguish between satellites and centrals; however the AM model performs considerably worse than the SHAM counterpart. We note that the AM and SHAM models are trained and evaluated on the same data set, so their performance metrics may be overinflated.

We also train several RF models, which serve as reasonable proxies for AM or conditional luminosity function models (Calderon & Berlind, 2019). By comparing panels (b) and (c), we observe that $V_{\rm max}$ is more physically connected to $M_\star$ than $M_{\rm halo}$, in agreement with previous findings (i.e., Conroy et al. 2006; Reddick et al. 2013; we find this to be true for the RF, AM, and SHAM models). A RF trained on both $M_{\rm halo}$ and $V_{\rm max}$ provides an even better reconstruction (RMSE = 0.148 dex).

Ultimately, we find that the GNN strongly outperforms all baseline models. While the GNN does not distinguish between centrals and satellites, it may be able to learn whether a given subhalo is a central based on surrounding subhalo properties (see Section 5.2).

4.2. Centrals versus satellites

Satellite dark matter halos are preferentially stripped relative to stars in a host halo's tidal field (Smith et al., 2016). In Appendix A, we show the stellar mass-halo mass relation for satellite and central galaxies in TNG300 (Figure 3). Indeed, we observe that satellite galaxies exhibit significantly more dispersion than centrals $M_\star$–$M_{\rm halo}$ relation. Our $3d$ GNN is also worse at predicting $\log(M_\star/M_\odot)$ for satellites than for centrals (see bottom two rows of Table 1), but this is due to the inherently larger scatter in the satellite-halo relation. We find that there is an overall negative bias for satellites and and positive bias for centrals, because the GNN must learn separate offset relations for both centrals and satellites.

[2]We define $\text{NMAD}(\boldsymbol{x}) \equiv k \cdot \text{median}(|x - \text{median}(\boldsymbol{x})|)$, where $k \approx 1.4826$ ensures that the NMAD and standard deviation are equal for a normally distributed $\boldsymbol{x}$.

Table 1. Cross-validation metrics for the AM, SHAM, RF, and GNN models discussed in the text. The GNN trained using positions, velocities, $M_{\rm halo}$, and $V_{\rm max}$ achieves the best metrics (shown in bold) in nearly every category. The last two rows report metrics for the $3d$ GNN model, except that only central and satellite subhalos are selected from the cross-validation set. We note that the AM/SHAM models are trained and evaluated on the same data set. For RF and GNN models, we repeat the entire training and cross-validation experiment three times; the scatter is too small to be shown in the displayed significant figures for all columns except the bias and outlier fraction.

Model	**RMSE** (dex)	**MAE** (dex)	**NMAD** (dex)	**Pearson ρ**	R^2	**Bias** (10^{-3} dex)	**Outlier fraction** (%)
AM - $M_{\rm halo}$	0.424	0.327	0.323	0.736	0.472	0.1	3.73
AM - $V_{\rm max}$	0.173	0.150	0.132	0.956	0.912	**0.0**	1.91
SHAM - $M_{\rm halo}$	0.332	0.231	0.235	0.838	0.677	0.1	6.20
SHAM - $V_{\rm max}$	0.151	0.133	0.115	0.966	0.933	**0.0**	1.75
RF - $M_{\rm halo}$	0.366	0.308	0.277	0.780	0.606	$\mathbf{-0.0 \pm 0.1}$	2.53 ± 0.01
RF - $V_{\rm max}$	0.197	0.177	0.152	0.942	0.886	-0.3 ± 0.0	1.44 ± 0.01
RF - $M_{\rm halo} + V_{\rm max}$	0.148	0.135	0.114	0.967	0.936	0.3 ± 0.0	1.31 ± 0.00
GNN ($2d$ projection)	0.135	0.131	0.106	0.973	0.946	-3.9 ± 2.2	0.68 ± 0.01
GNN ($3d$)	**0.129**	**0.125**	**0.102**	**0.975**	**0.951**	0.8 ± 0.6	$\mathbf{0.68 \pm 0.00}$
GNN ($3d$) - centrals	0.123	0.119	0.097	0.979	0.959	4.6 ± 0.7	0.67 ± 0.01
GNN ($3d$) - satellites	0.138	0.136	0.109	0.968	0.936	-5.0 ± 0.6	0.58 ± 0.01

4.3. Cosmic substructure in projection

We also construct cosmic graphs in projection, i.e. projected coordinates x_1 and x_2, and radial velocity v_3, instead of the full phase space information (see Appendix B). This $2d$ GNN model achieves RMSE = 0.135 dex scatter, which still exceeds the performance of the best RF estimator (see Table 1). Because the $2d$ GNN encode projected large scale structure information, it outperforms the RF models that can only learn isolated subhalo information.

5. Discussion

We have presented a novel method for populating dark matter subhalos with galaxy stellar masses. Mathematical graphs combine individual halo properties and environmental parameters in an equivariant representation, resulting in robust predictions for both central and satellite galaxies. As shown in Table 1 and Figure 2, the cosmic graphs outperform random forests trained on $V_{\rm max}$ and $M_{\rm halo}$. For galaxies with $\log(M_\star/M_\odot) \geq 9$ and $\log(M_{\rm halo}/M_\odot) \geq 10$, we recover the logarithmic stellar mass to within a root mean squared error (RMSE) of 0.129 dex.

5.1. Inductive biases of GNNs

We note that previous works have employed convolutional neural networks (CNNs) for painting stars onto dark matter maps (Zhang et al., 2019; Mohammad et al., 2022). Unlike abundance matching models and RFs, CNNs are able to represent local spatial information. However, CNNs and GNNs have different inductive biases: CNNs are well-suited for representing fields discretized onto a Cartesian grid, while GNNs are well-suited for representing objects and relationships between them. Galaxies have small sizes (∼kpc) relative to their typical separations (∼Mpc), and they interact with each other (and their surrounding media) through multiple physical mechanisms (e.g., gravitational attraction, tides, ram pressure, etc). Therefore, cosmic structures naturally conform to a graphical representation, motivating our use of GNNs in this work.

5.2. Galaxy environments

We note that a GNN with no edges except self-loops would essentially model the galaxy-halo connection in isolation; all environmental information is contained and passed along the edges. However, if we remove self-loops from the GNN, then the GNN is still able to infer $\log(M_\star/M_\odot)$ to within RMSE ~ 0.145 dex. A GNN without self-loops must estimate galaxy stellar mass *solely* from neighboring halo information, which demonstrates that galaxy environments are informative for modeling the galaxy-halo connection.

We find that the GNN with max-pooling aggregation function achieves 0.001 dex lower RMSE than a GNN with sum-pooling. This result suggests that the GNN selects the largest value for some combination of $M_{\rm halo}$, $V_{\rm max}$, and distance to neighboring subhalos in order to best make predictions. We can speculatively interpret this as evidence that the largest and most nearby subhalo is most informative to a GNN. The largest subhalo might dominate environmental effects (e.g. tides and ram pressure) and control a given subhalo's stellar mass. Meanwhile, the summed information should capture *all* of the forces, and we expect it to be more robust or transferable across domains. This interpretation requires addition testing and an exhaustive hyperparameter search over GNN architecture and optimization procedures, which we aim to do in a follow-up work.[3]

5.3. Applications to observations

The strong performance of $2d$ GNNs (§4.3) is promising for facilitating comparisons to observations beyond the Local Group, where we can only reliably measure projected positions and line-of-sight velocities rather than full phase space information. Our method can be used to quickly estimate galaxy properties of constrained N-body (McAlpine et al., 2022) and Gpc-scale N-body simulated volumes (Garrison et al., 2018; Maksimova et al., 2021) for comparison with wide-area galaxy surveys in the low-redshift Universe (Ruiz-Macias et al., 2021; Carlsten et al., 2022; Darragh-Ford et al., 2022; Driver et al., 2022; Wu et al., 2022).

5.4. Limitations and caveats

While we have shown that the GNN outperforms other methods, this demonstration does not definitively prove that GNNs are exploiting environmental information. Indeed, we have used a linking length of 5 Mpc, but this hyperparameter may be suboptimal and should be tuned. It is also possible that intrinsic scatter imposes a RMSE floor (i.e., due to the "butterfly effect" in cosmological simulations Genel et al., 2019), although GNN results using merger trees have shown that galaxy stellar mass can be recovered to even lower scatter (Jespersen et al., 2022). Finally, it may be that merger history is more important than environmental information, and that the clustering information learned by a GNN only incrementally improves performance relative to other approaches.

Our results will depend on choice of halo finder, i.e. if we were to use an alternative to the SUBFIND algorithm (e.g. ROCKSTAR; Behroozi et al. 2013). We have not tested our results using different halo finding tools, and it is unclear whether a GNN trained using one halo finder catalog will properly generalize to another catalog produced by a different halo finder. We also note that our results, while promising, must be tested on dark matter only simulations with halo catalogs matched to the hydrodynamic simulation catalogs before we can rely on GNNs to paint galaxies onto dark matter subhalos.

Additionally, domain adaptation will likely be needed to ensure simulated results can transfer to other simulations (e.g., while varying cosmological parameters; Villaescusa-Navarro et al. 2021) or to observations (e.g, Ciprijanovic et al., 2023). As a preliminary test, we repeat our experiment by training on TNG300 and validating on TNG50 data, and vice versa; in both cases the results are poor (> 0.2 dex). However, by training on a subset both simulations, we can recover $\log(M_\star/M_\odot)$ to ~ 0.13 dex for TNG300 and ~ 0.14 dex for TNG50 (Nelson et al., 2019a;b). This test suggests that cross-domain applications, such as transferring GNN results from simulations to observations, will necessitate some form of domain adaptation.

[3] The linking length is a particularly important hyperparameter. In our preliminary tests, we have found 5 Mpc to give good results.

Software and Data

Our code is completely public on Github: https://github.com/jwuphysics/halo-gnns/tree/halos-to-stars. We have used the following software and tools: numpy (Harris et al., 2020), matplotlib (Hunter, 2007), pandas (Wes McKinney, 2010), pytorch (Paszke et al., 2019), and pytorch-geometric (Fey & Lenssen, 2019).

We only use public simulation data from Illustris, which can be downloaded from https://www.tng-project.org/data/.

Acknowledgments

JFW and CKJ thank Peter Behroozi, Haley Bowden, Francisco Villaescua-Navarro, Tjitske Starkenburg, and Risa Wechsler for valuable discussions that sharpened this work. We also thank the two anonymous reviewers who provided excellent comments and suggestions that improved this manuscript. This research has made use of NASA's Astrophysics Data System Bibliographic Services. The authors are grateful to the Kavli Institute for Theoretical Physics "Building a Physical Understanding of Galaxy Evolution with Data-driven Astronomy" program, where this work began. This research was supported in part by the National Science Foundation under Grant No. NSF PHY-1748958.

References

Agarwal, S., Davé, R., and Bassett, B. A. Painting galaxies into dark matter haloes using machine learning. *MNRAS*, 478(3):3410–3422, August 2018. doi: 10.1093/mnras/sty1169.

Battaglia, P. W., Pascanu, R., Lai, M., Rezende, D., and Kavukcuoglu, K. Interaction Networks for Learning about Objects, Relations and Physics. *arXiv e-prints*, art. arXiv:1612.00222, December 2016. doi: 10.48550/arXiv.1612.00222.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., and Pascanu, R. Relational inductive biases, deep learning, and graph networks. *arXiv e-prints*, art. arXiv:1806.01261, June 2018. doi: 10.48550/arXiv.1806.01261.

Behroozi, P., Wechsler, R. H., Hearin, A. P., and Conroy, C. UNIVERSEMACHINE: The correlation between galaxy growth and dark matter halo assembly from z = 0-10. *MNRAS*, 488(3):3143–3194, September 2019. doi: 10.1093/mnras/stz1182.

Behroozi, P. S., Wechsler, R. H., and Wu, H.-Y. The ROCKSTAR Phase-space Temporal Halo Finder and the Velocity Offsets of Cluster Cores. *ApJ*, 762(2):109, January 2013. doi: 10.1088/0004-637X/762/2/109.

Berlind, A. A. and Weinberg, D. H. The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass. *ApJ*, 575(2):587–616, August 2002. doi: 10.1086/341469.

Calderon, V. F. and Berlind, A. A. Prediction of galaxy halo masses in SDSS DR7 via a machine learning approach. *MNRAS*, 490(2):2367–2379, December 2019. doi: 10.1093/mnras/stz2775.

Carlsten, S. G., Greene, J. E., Beaton, R. L., Danieli, S., and Greco, J. P. The Exploration of Local VolumE Satellites (ELVES) Survey: A Nearly Volume-limited Sample of Nearby Dwarf Satellite Systems. *ApJ*, 933(1):47, July 2022. doi: 10.3847/1538-4357/ac6fd7.

Ciprijanovic, A., Lewis, A., Pedro, K., Madireddy, S., Nord, B., Perdue, G., and Wild, S. M. Deepastrouda: Semi-supervised universal domain adaptation for cross-survey galaxy morphology classification and anomaly detection. *Machine Learning: Science and Technology*, 2023. URL http://iopscience.iop.org/article/10.1088/2632-2153/acca5f.

Conroy, C., Wechsler, R. H., and Kravtsov, A. V. Modeling Luminosity-dependent Galaxy Clustering through Cosmic Time. *ApJ*, 647(1):201–214, August 2006. doi: 10.1086/503602.

Darragh-Ford, E., Wu, J. F., Mao, Y.-Y., Wechsler, R. H., Geha, M., Forero-Romero, J. E., Hahn, C., Kallivayalil, N., Moustakas, J., Nadler, E. O., Nowotka, M., Peek, J. E. G., Tollerud, E. J., Weiner, B., Aguilar, J., Ahlen, S., Brooks, D., Cooper, A. P., de la Macorra, A., Dey, A., Fanning, K., Font-Ribera, A., Gontcho, S. G. A., Honscheid, K., Kisner, T., Kremin, A., Landriau, M., Levi, M. E., Martini, P., Meisner, A. M., Miquel, R., Myers, A. D., Nie, J., Palanque-Delabrouille, N., Percival, W. J., Prada, F., Schlegel, D., Schubnell, M., Tarlé, G., Vargas-Magaña, M., Zhou, Z., and Zou, H. Target Selection and Sample Characterization for the DESI LOW-Z Secondary Target Program. *arXiv e-prints*, art. arXiv:2212.07433, December 2022. doi: 10.48550/arXiv.2212.07433.

de Santi, N. S. M., Shao, H., Villaescusa-Navarro, F., Abramo, L. R., Teyssier, R., Villanueva-Domingo, P., Ni, Y., Anglés-Alcázar, D., Genel, S., Hernandez-Martinez, E., Steinwandel, U. P., Lovell, C. C., Dolag, K., Castro, T., and Vogelsberger, M. Robust field-level likelihood-free inference with galaxies. *arXiv e-prints*, art. arXiv:2302.14101, February 2023. doi: 10.48550/arXiv.2302.14101.

Driver, S. P., Bellstedt, S., Robotham, A. S. G., Baldry, I. K., Davies, L. J., Liske, J., Obreschkow, D., Taylor, E. N., Wright, A. H., Alpaslan, M., Bamford, S. P., Bauer, A. E., Bland-Hawthorn, J., Bilicki, M., Bravo, M., Brough, S., Casura, S., Cluver, M. E., Colless, M., Conselice, C. J., Croom, S. M., de Jong, J., D'Eugenio, F., De Propris, R., Dogruel, B., Drinkwater, M. J., Dvornik, A., Farrow, D. J., Frenk, C. S., Giblin, B., Graham, A. W., Grootes, M. W., Gunawardhana, M. L. P., Hashemizadeh, A., Häußler, B., Heymans, C., Hildebrandt, H., Holwerda, B. W., Hopkins, A. M., Jarrett, T. H., Heath Jones, D., Kelvin, L. S., Koushan, S., Kuijken, K., Lara-López, M. A., Lange, R., López-Sánchez, Á. R., Loveday, J., Mahajan, S., Meyer, M., Moffett, A. J., Napolitano, N. R., Norberg, P., Owers, M. S., Radovich, M., Raouf, M., Peacock, J. A., Phillipps, S., Pimbblet, K. A., Popescu, C., Said, K., Sansom, A. E., Seibert, M., Sutherland, W. J., Thorne, J. E., Tuffs, R. J., Turner, R., van der Wel, A., van Kampen, E., and Wilkins, S. M. Galaxy And Mass Assembly (GAMA): Data Release 4 and the z ¡ 0.1 total and z ¡ 0.08 morphological galaxy stellar mass functions. *MNRAS*, 513(1):439–467, June 2022. doi: 10.1093/mnras/stac472.

Fey, M. and Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric, 5 2019. URL https://github.com/pyg-team/pytorch_geometric.

Garcia Satorras, V., Hoogeboom, E., and Welling, M. E(n) Equivariant Graph Neural Networks. *arXiv e-prints*, art. arXiv:2102.09844, February 2021. doi: 10.48550/arXiv.2102.09844.

Garrison, L. H., Eisenstein, D. J., Ferrer, D., Tinker, J. L., Pinto, P. A., and Weinberg, D. H. The Abacus Cosmos: A Suite of Cosmological N-body Simulations. *ApJS*, 236(2):43, June 2018. doi: 10.3847/1538-4365/aabfd3.

Genel, S., Bryan, G. L., Springel, V., Hernquist, L., Nelson, D., Pillepich, A., Weinberger, R., Pakmor, R., Marinacci, F., and Vogelsberger, M. A Quantification of the Butterfly Effect in Cosmological Simulations and Implications for Galaxy Scaling Relations. *ApJ*, 871(1):21, January 2019. doi: 10.3847/1538-4357/aaf4bb.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

Jagvaral, Y., Lanusse, F., Singh, S., Mandelbaum, R., Ravanbakhsh, S., and Campbell, D. Galaxies on graph neural networks: towards robust synthetic galaxy catalogs with deep generative models. *arXiv e-prints*, art. arXiv:2212.05596, December 2022. doi: 10.48550/arXiv.2212.05596.

Jeffrey, N. and Wandelt, B. D. Solving high-dimensional parameter inference: marginal posterior densities & Moment Networks. *arXiv e-prints*, art. arXiv:2011.05991, November 2020. doi: 10.48550/arXiv.2011.05991.

Jespersen, C. K., Cranmer, M., Melchior, P., Ho, S., Somerville, R. S., and Gabrielpillai, A. Mangrove: Learning Galaxy Properties from Merger Trees. *ApJ*, 941(1):7, December 2022. doi: 10.3847/1538-4357/ac9b18.

Kamdar, H. M., Turk, M. J., and Brunner, R. J. Machine learning and cosmological simulations - II. Hydrodynamical simulations. *MNRAS*, 457(2):1162–1179, April 2016. doi: 10.1093/mnras/stv2981.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, art. arXiv:1412.6980, December 2014. doi: 10.48550/arXiv.1412.6980.

Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., Klypin, A. A., Gottlöber, S., Allgood, B., and Primack, J. R. The Dark Side of the Halo Occupation Distribution. *ApJ*, 609 (1):35–49, July 2004. doi: 10.1086/420959.

Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization. *arXiv e-prints*, art. arXiv:1711.05101, November 2017. doi: 10.48550/arXiv.1711.05101.

Makinen, T. L., Charnock, T., Lemos, P., Porqueres, N., Heavens, A. F., and Wandelt, B. D. The Cosmic Graph: Optimal Information Extraction from Large-Scale Structure using Catalogues. *The Open Journal of Astrophysics*, 5(1):18, December 2022. doi: 10.21105/astro.2207.05202.

Maksimova, N. A., Garrison, L. H., Eisenstein, D. J., Hadzhiyska, B., Bose, S., and Satterthwaite, T. P. AbacusSummit: a massive set of high-accuracy, high-resolution N-body simulations. *Monthly Notices of the Royal Astronomical Society*, 508(3):4017–4037, 09 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab2484. URL https://doi.org/10.1093/mnras/stab2484.

McAlpine, S., Helly, J. C., Schaller, M., Sawala, T., Lavaux, G., Jasche, J., Frenk, C. S., Jenkins, A., Lucey, J. R., and Johansson, P. H. SIBELIUS-DARK: a galaxy catalogue of the local volume from a constrained realization simulation. *MNRAS*, 512(4):5823–5847, June 2022. doi: 10.1093/mnras/stac295.

Mohammad, F. G., Villaescusa-Navarro, F., Genel, S., Anglés-Alcázar, D., and Vogelsberger, M. Inpainting Hydrodynamical Maps with Deep Learning. *ApJ*, 941(2): 132, December 2022. doi: 10.3847/1538-4357/ac9f14.

Moster, B. P., Somerville, R. S., Maulbetsch, C., van den Bosch, F. C., Macciò, A. V., Naab, T., and Oser, L. Constraints on the Relationship between Stellar Mass and Halo Mass at Low and High Redshift. *ApJ*, 710(2):903–923, February 2010. doi: 10.1088/0004-637X/710/2/903.

Moster, B. P., Naab, T., Lindström, M., and O'Leary, J. A. GalaxyNet: connecting galaxies and dark matter haloes with deep neural networks and reinforcement learning in large volumes. *MNRAS*, 507(2):2115–2136, October 2021. doi: 10.1093/mnras/stab1449.

Nelson, D., Pillepich, A., Springel, V., Pakmor, R., Weinberger, R., Genel, S., Torrey, P., Vogelsberger, M., Marinacci, F., and Hernquist, L. First results from the TNG50 simulation: galactic outflows driven by supernovae and black hole feedback. *MNRAS*, 490(3):3234–3261, December 2019a. doi: 10.1093/mnras/stz2306.

Nelson, D., Springel, V., Pillepich, A., Rodriguez-Gomez, V., Torrey, P., Genel, S., Vogelsberger, M., Pakmor, R., Marinacci, F., Weinberger, R., Kelley, L., Lovell, M., Diemer, B., and Hernquist, L. The IllustrisTNG simulations: public data release. *Computational Astrophysics and Cosmology*, 6(1):2, May 2019b. doi: 10.1186/s40668-019-0028-x.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Pillepich, A., Springel, V., Nelson, D., Genel, S., Naiman, J., Pakmor, R., Hernquist, L., Torrey, P., Vogelsberger, M., Weinberger, R., and Marinacci, F. Simulating galaxy formation with the IllustrisTNG model. *MNRAS*, 473(3): 4077–4106, January 2018. doi: 10.1093/mnras/stx2656.

Pillepich, A., Nelson, D., Springel, V., Pakmor, R., Torrey, P., Weinberger, R., Vogelsberger, M., Marinacci, F., Genel, S., van der Wel, A., and Hernquist, L. First results from the TNG50 simulation: the evolution of stellar and gaseous discs across cosmic time. *MNRAS*, 490(3):3196–3233, December 2019. doi: 10.1093/mnras/stz2338.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., Baccigalupi, C., Banday, A. J., Barreiro, R. B., Bartlett, J. G., and et al. Planck 2015 results. XIII. Cosmological parameters. *A&A*, 594:A13, September 2016. doi: 10.1051/0004-6361/201525830.

Reddick, R. M., Wechsler, R. H., Tinker, J. L., and Behroozi, P. S. The Connection between Galaxies and Dark Matter Structures in the Local Universe. *ApJ*, 771(1):30, July 2013. doi: 10.1088/0004-637X/771/1/30.

Ruiz-Macias, O., Zarrouk, P., Cole, S., Baugh, C. M., Norberg, P., Lucey, J., Dey, A., Eisenstein, D. J., Doel, P., Gaztañaga, E., Hahn, C., Kehoe, R., Kitanidis, E., Landriau, M., Lang, D., Moustakas, J., Myers, A. D., Prada, F., Schubnell, M., Weinberg, D. H., and Wilson, M. J. Characterizing the target selection pipeline for the Dark Energy Spectroscopic Instrument Bright Galaxy Survey. *MNRAS*, 502(3):4328–4349, April 2021. doi: 10.1093/mnras/stab292.

Shao, H., Villaescusa-Navarro, F., Villanueva-Domingo, P., Teyssier, R., Garrison, L. H., Gatti, M., Inman, D., Ni, Y., Steinwandel, U. P., Kulkarni, M., Visbal, E., Bryan, G. L., Anglés-Alcázar, D., Castro, T., Hernández-Martínez, E., and Dolag, K. Robust Field-level Inference of Cosmological Parameters with Dark Matter Halos. *ApJ*, 944(1):27, February 2023. doi: 10.3847/1538-4357/acac7a.

Smith, R., Choi, H., Lee, J., Rhee, J., Sanchez-Janssen, R., and Yi, S. K. The Preferential Tidal Stripping of Dark Matter versus Stars in Galaxies. *ApJ*, 833(1):109, December 2016. doi: 10.3847/1538-4357/833/1/109.

Somerville, R. S. and Davé, R. Physical Models of Galaxy Formation in a Cosmological Framework. *ARA&A*, 53:51–113, August 2015. doi: 10.1146/annurev-astro-082812-140951.

Springel, V., White, S. D. M., Tormen, G., and Kauffmann, G. Populating a cluster of galaxies - I. Results at [formmu2]z=0. *MNRAS*, 328(3):726–750, December 2001. doi: 10.1046/j.1365-8711.2001.04912.x.

Tang, K. S. and Ting, Y.-S. Galaxy Merger Reconstruction with Equivariant Graph Normalizing Flows. *arXiv e-prints*, art. arXiv:2207.02786, July 2022. doi: 10.48550/arXiv.2207.02786.

Vale, A. and Ostriker, J. P. Linking halo mass to galaxy luminosity. *MNRAS*, 353(1):189–200, September 2004. doi: 10.1111/j.1365-2966.2004.08059.x.

Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., Spergel, D. N., Somerville, R. S., Dave, R., Pillepich, A., Hernquist, L., Nelson, D., Torrey, P., Narayanan, D., Li, Y., Philcox, O., La Torre, V., Maria Delgado, A., Ho, S., Hassan, S., Burkhart, B., Wadekar, D., Battaglia, N., Contardo, G., and Bryan, G. L. The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations. *ApJ*, 915(1):71, July 2021. doi: 10.3847/1538-4357/abf7ba.

Villanueva-Domingo, P. and Villaescusa-Navarro, F. Learning Cosmology and Clustering with Cosmic Graphs. *ApJ*, 937(2):115, October 2022. doi: 10.3847/1538-4357/ac8930.

Villanueva-Domingo, P., Villaescusa-Navarro, F., Genel, S., Anglés-Alcázar, D., Hernquist, L., Marinacci, F., Spergel, D. N., Vogelsberger, M., and Narayanan, D. Weighing the Milky Way and Andromeda with Artificial Intelligence. *arXiv e-prints*, art. arXiv:2111.14874, November 2021. doi: 10.48550/arXiv.2111.14874.

Villanueva-Domingo, P., Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., Marinacci, F., Spergel, D. N., Hernquist, L., Vogelsberger, M., Dave, R., and Narayanan, D. Inferring Halo Masses with Graph Neural Networks. *ApJ*, 935(1):30, August 2022. doi: 10.3847/1538-4357/ac7aa3.

Vogelsberger, M., Marinacci, F., Torrey, P., and Puchwein, E. Cosmological simulations of galaxy formation. *Nature Reviews Physics*, 2(1):42–66, January 2020. doi: 10.1038/s42254-019-0127-2.

Wechsler, R. H. and Tinker, J. L. The Connection Between Galaxies and Their Dark Matter Halos. *ARA&A*, 56:435–487, September 2018. doi: 10.1146/annurev-astro-081817-051756.

Wechsler, R. H., Bullock, J. S., Primack, J. R., Kravtsov, A. V., and Dekel, A. Concentrations of Dark Halos from Their Assembly Histories. *ApJ*, 568(1):52–70, March 2002. doi: 10.1086/338765.

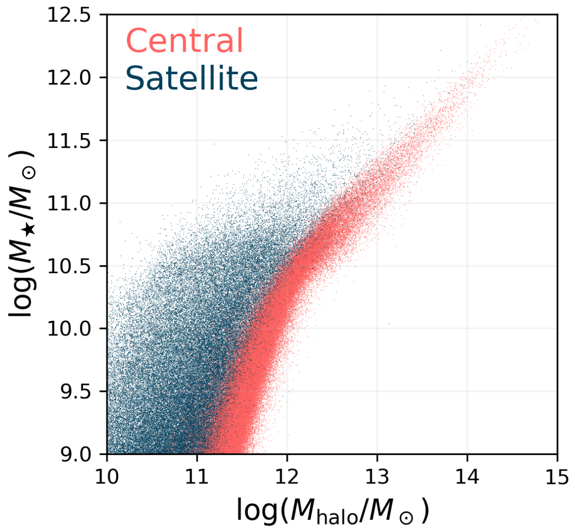

Figure 3. The stellar mass-halo mass relation in TNG300 for satellite and central galaxies.

Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman (eds.), *Proceedings of the 9th Python in Science Conference*, pp. 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

Wu, J. F., Peek, J. E. G., Tollerud, E. J., Mao, Y.-Y., Nadler, E. O., Geha, M., Wechsler, R. H., Kallivayalil, N., and Weiner, B. J. Extending the SAGA Survey (xSAGA). I. Satellite Radial Profiles as a Function of Host-galaxy Properties. *ApJ*, 927(1):121, March 2022. doi: 10.3847/1538-4357/ac4eea.

Yang, X., Mo, H. J., and van den Bosch, F. C. Constraining galaxy formation and cosmology with the conditional luminosity function of galaxies. *MNRAS*, 339(4): 1057–1080, March 2003. doi: 10.1046/j.1365-8711.2003.06254.x.

Zhang, X., Wang, Y., Zhang, W., Sun, Y., He, S., Contardo, G., Villaescusa-Navarro, F., and Ho, S. From Dark Matter to Galaxies with Convolutional Networks. *arXiv e-prints*, art. arXiv:1902.05965, February 2019. doi: 10.48550/arXiv.1902.05965.

A. The stellar mass-halo mass relation for satellites and centrals

In Figure 3, we show halo masses and stellar masses for central galaxies (red) and satellites (blue) from the TNG300 `SUBFIND` catalogs. Our GNN is able to learn the offset relationships for both central and satellite subhalos.

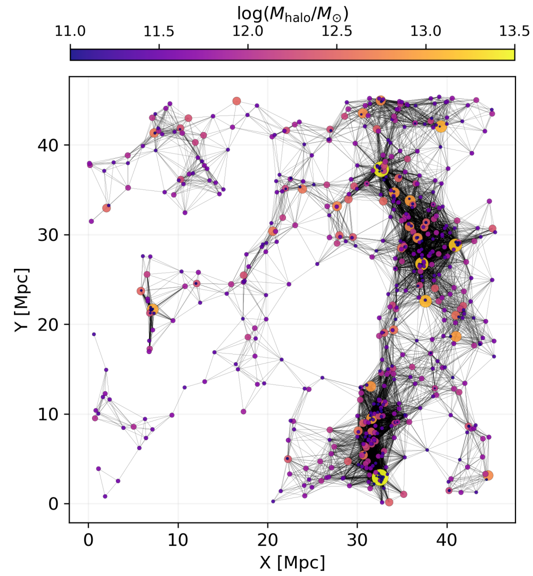

Figure 4. A graph of galaxies in projection, analogous to Figure 1. Subhalos now connected with edges if their projected distances are less than 5 Mpc.

B. Cosmic graphs in projected coordinates

In §4.3, we trained a GNN to learn the galaxy-halo connection using projected positions and radial velocity, in addition to $M_{\rm halo}$ and $V_{\rm max}$. In Figure 4, we show a projected version of the subvolume that appeared in Figure 1.