Predicting dark matter halo masses from simulated galaxy images and environments

Austin Larson
Department of Computer Science,
Johns Hopkins University

John F Wu
Space Telescope Science Institute
Department of Physics & Astronomy,
Johns Hopkins University

Craig Jones
Department of Computer Science,
Johns Hopkins University

1. Introduction

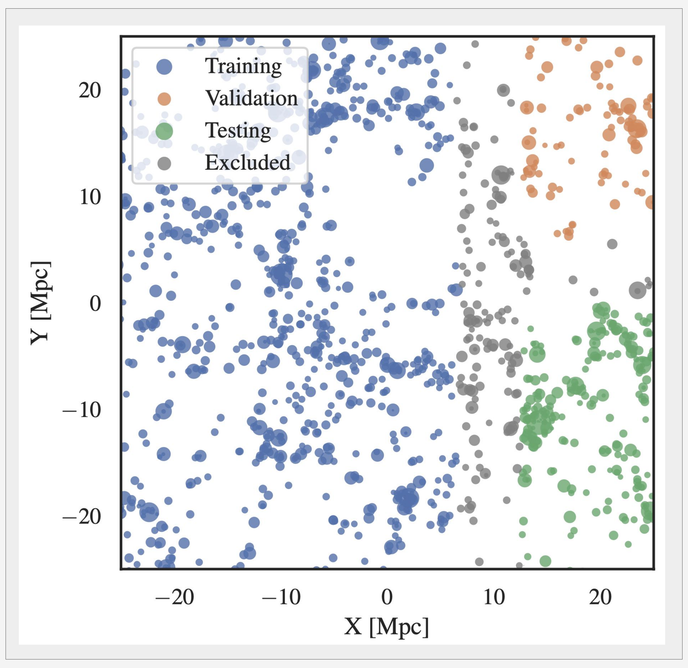
Galaxies are theorized to form inside and co-evolve with dark matter halos. The close relationship between galaxies and their halos has led to a tight relationship between galaxy stellar mass $M_\star$ and dark matter halo mass M_{halo}, which is known as the stellar mass–halo mass relation (SMHMR). The SMHMR can be calibrated by galaxy and halo properties derived from cosmological hydrodynamic simulations, or from other approaches such as semi- analytic models or empirical models [1].

However, it is likely that M_{halo} depends on galaxy properties other than $M_\star$. We present an exploration of how M_{halo} might be predicted from not just the stellar mass, but also galaxy morphology and the spatial distributions of galaxies (i.e., environment). Using a galaxy sample from the Illustris TNG50 cosmological simulation[2], we test whether galaxy morphology and large-scale environment contain information that can improve (lower) the scatter in the SMHMR.

We train simple machine learning (ML) models such as random forests to predict M_{halo} from $M_\star$. We also evaluate the level of improvement in predicting M_{halo} after including galaxy morphology and galaxy environmental overdensity as ML model features.

We also consider deep learning methods: *(a)* we train a convolutional neural network (CNN) that learns morphological information directly from synthetic galaxy image cutouts, and *(b)*, we train a graph neural network (GNN) that learns environmental information directly from galaxy point clouds. Finally, we combine the CNN and GNN into a joint model that encodes information from both the surroundings and appearances of galaxies to predict their dark matter content.

2. Method: Comparing Scalar methods, CNNs, and GNNs



We used simulated galaxy data from the z = 0 snapshot from TNG50, the highest resolution hydrodynamical simulation in the IllustrisTNG Project. Our selection of $M_\star > 10^{9.5}$ M_{sun} galaxies limits the sample size to N = 1,666. We download SKIRT-processed galaxy images in three optical-wavelength bands, and postprocess them to imitate Pan-STARRS survey imaging [3].

Baseline Model	$M_\star$	R_{Petro}	S	A	Δ_G
Simple	√				
Morphological	√	√	√	√	
Overdensity	√				√
Combined	√	√	√	√	√

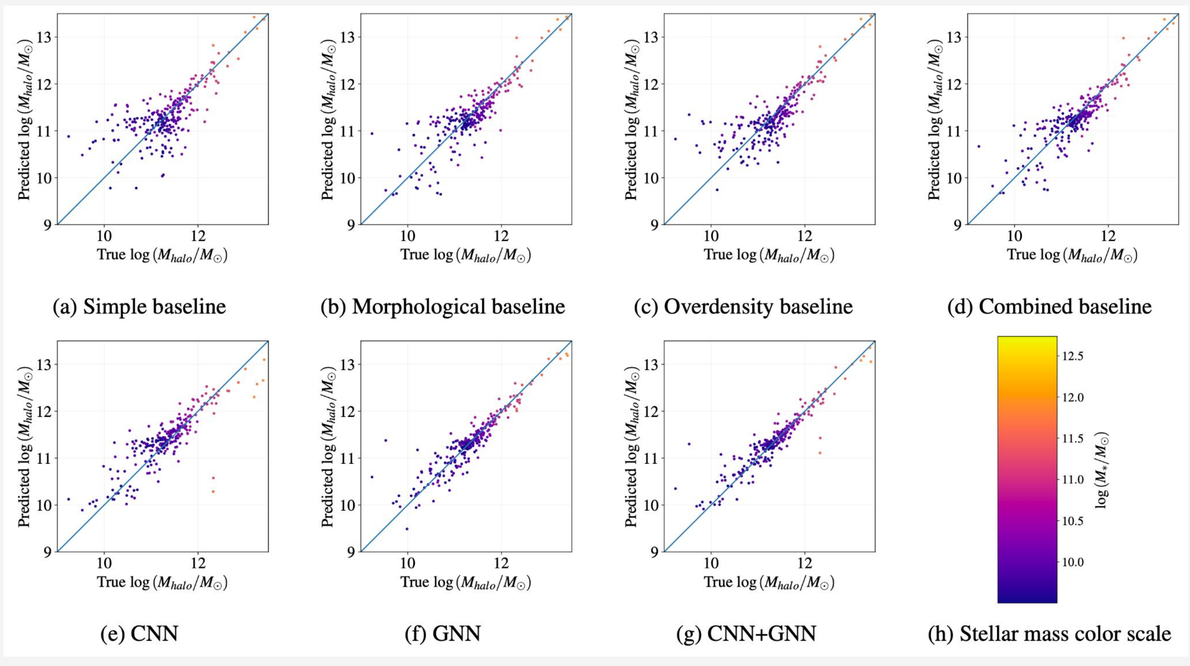
The data is split into subsets with 6 Mpc separation to ensure that the GNN is unable to learn information from the test set. Our baseline methods use random forests with a variety of scalar features, including stellar mass ($M_\star$), petrosian radius (r_{pet}), smoothness (S), asymmetry (A), and overdensity (Δ_g). Here we compute overdensity by counting the number of other galaxies within 3 Mpc.

The morphological baseline is compared to the CNN using galaxy images. The GNN, based on prior work [4,5], uses stellar mass and positions to compare with the overdensity baseline. A combined CNN+GNN is compared with the combined baseline model.

3. The cosmic graph structure is critical for mapping dark matter from galaxies!

The CNN performs better than the morphological baseline model in terms of outlier-insensitive metrics such as the MAE and NMAD. The GNN learns far more environmental information than is encoded in the simple overdensity parameter Δ_g, leading to far lower prediction errors. The CNN+GNN model achieves the best performance metrics across the board, and can benefit from even more training.

Model Name	MSE	MAE	RMSE	R^2	Bias	NMAD	$f_{outlier}$
Simple Baseline	0.207	0.328	0.455	0.524	+0.009	0.313	2.27%
Morphological Baseline	0.119	0.246	0.345	0.726	−0.023	0.234	2.43%
Overdensity Basline	0.140	0.260	0.374	0.680	−0.008	0.261	2.09%
Combined Baseline	0.096	0.220	0.310	0.779	−0.034	0.220	2.26%
CNN	0.129	0.238	0.359	0.705	+0.062	0.218	6.18%
GNN	0.062	0.158	0.248	0.858	+0.007	0.146	1.78%
CNN + GNN	0.062	0.144	0.248	0.859	+0.020	0.123	1.77%



(a) Simple baseline (b) Morphological baseline (c) Overdensity baseline (d) Combined baseline

(e) CNN (f) GNN (g) CNN+GNN (h) Stellar mass color scale

4. Summary and discussion

The GNN and GNN+CNN models perform best. These GNNs are trained using cosmic graphs connected by a 3 Mpc linking length, suggesting that galaxy environment is particularly important for predicting galaxies' halo masses. Surprisingly, we find that the CNN-only model results in a poor RMSE and suffers from numerous outliers (which are visually apparent in the figure above). Due to the small size of our dataset, we suspect that the CNN is undertrained. Additionally, the small test set may not represent the full simulation volume.

We caution that our models may have learned the specific characteristics of the TNG50 simulation, and may not generalize well to other data. If we seek to robustly apply ML predictions outside of the TNG data set, we should employ techniques such as domain adaptation [6] or train on multiple simulations that vary astrophysical or cosmological parameters [7].

References and acknowledgments

[1] Wechsler & Tinker 2018, 56, 435
[2] Pillipech et al., 2019, *MNRAS*, 490, 3196
[3] Rodriguez-Gomez et al. 2019, 483, 4140
[4] Wu & Jespersen 2023, *ICML ML4Astro Workshop*
[5] Wu et al., 2024, *arXiv*, 2402.07995
[6] Ciprijanovic et al. 2023, *MLST*, 4, 025013
[7] Villaescusa-Navarro et al., 2023, *ApJS*, 265, 54

The authors are grateful for the KITP galevo23 program, which inspired this work. This research was supported in part by the National Science Foundation under Grant No. NSF PHY-1748958.