Predicting dark matter halo masses from simulated galaxy images and environments

Anonymous Authors[1]

Abstract

Galaxies are theorized to form and co-evolve with their dark matter halos, such that their stellar masses and halo masses should be well-correlated. However, it is not known whether other observable galaxy features, such as their morphologies or large-scale environments, can be used to tighten the correlation between galaxy properties and halo masses. In this work, we train a baseline random forest model to predict halo mass using galaxy features from the Illustris TNG50 hydrodynamical simulation, and compare with convolutional neural networks (CNNs) and graph neural networks (GNNs) trained respectively using galaxy image cutouts and galaxy point clouds. The best baseline model has a root mean squared error (RMSE) of 0.310 and mean absolute error (MAE) of 0.220, compared to the CNN (RSME $= 0.359$, MAE $= 0.238$), GNN (RMSE $= 0.248$, MAE $= 0.158$), and a novel combined CNN+GNN (RMSE $= 0.248$, MAE $= 0.144$). The CNN is likely limited by our small data set, and we anticipate that the CNN and CNN+GNN would benefit from training on larger cosmological simulations. We conclude that deep learning models can leverage information from galaxy appearances and environment, beyond commonly used summary statistics, in order to better predict the halo mass.

1. Introduction

Galaxies are theorized to form inside and co-evolve with dark matter halos (Wechsler & Tinker, 2018). 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 (e.g., Somerville & Davé, 2015; Behroozi et al., 2019).

While galaxies are observable in the real Universe, dark matter properties are often determined through indirect measurements. For example, dark matter halo mass distributions can be inferred from the galaxy and galaxy cluster kinematics (e.g., Zwicky, 1933; Rubin et al., 1980) or through rare instances of gravitational lensing (e.g., Clowe et al., 2006). However, these constraints rely on detailed observations that are not widely available; in those cases, M_{halo} can only be assumed by using the SMHMR to assign a halo mass given a galaxy stellar mass.

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 large-scale environment. Using a galaxy sample from the Illustris TNG50-1 cosmological simulation (hereafter TNG50; Nelson et al., 2019b;a; Pillepich et al., 2019), we test whether galaxy morphology and large-scale environment contains information that can improve (lower) the scatter in the SMHMR.

A machine learning (ML) model such as a random forest can easily learn to predict M_{halo} from $M_\star$ (e.g., Kamdar et al., 2016). We gauge the level of improvement in predicting M_{halo} after we include galaxy morphology (i.e., commonly used summary statistics) and/or environmental overdensity as features in ML models. We train deep neural networks to learn the morphological information directly from synthetic galaxy imaging, and to learn the environmental information directly from galaxy point clouds. In all these experiments, we rely on the simulated galaxy sample from TNG50 to calibrate M_{halo} against other galaxy properties.

1.1. Some notes on nomenclature

Before we proceed, we define some terminology used in this paper. The terms *galaxies* and *halos* are used interchangeably, since every TNG50 galaxy studied here resides in a dark matter halo. Halos can be *satellites* (also known as *subhalos*) of a more massive halo, or they can be *centrals*. The latter means that they gravitationally dominate their surroundings.

We test several ML models that take galaxy *features* as inputs. All features are derived from the TNG50 data (see

Section 2). For the baseline models (described in Section 3.3), we employ morphological *summary statistics*, i.e., scalar features commonly computed by astronomers from galaxy images that describe the appearances of galaxies. Environmental overdensity can also be described with summary statistics. Crucially, these summary statistics are lossy descriptions of the underlying data, i.e., morphological summary statistics do not fully describe the information in galaxy images, and the overdensity summary statistic does not fully describe the information in galaxy environments.

We compare baseline models against more sophisticated neural network models. To quantify these comparisons, we introduce several *evaluation metrics* (see Section 3.2). Some metrics penalize large errors (i.e., *outliers*) more severely, while other metrics are insensitive to outliers. For most metrics presented here (with the exception of the R^2 correlation coefficient), lower values are better. We optimize all models using the mean squared error (MSE) loss function, which also serves as a metric for comparison.

1.2. Our contributions

We present two findings that represent novel contributions to our field:

- Simple parameterizations of galaxy morphology and environmental overdensity are informative for predicting galaxy halo masses. Combining these two sets of information yields even better predictions.

- Deep neural networks can learn additional information not contained in the simple features. Although our convolutional neural network requires a larger training sample to optimally and flexibly learn all the information within synthetic galaxy images, we find that our graph neural network dramatically lowers the M_{halo} prediction error by learning environmental relationships from the galaxy point cloud.

2. TNG50 Simulated Data

2.1. Galaxy catalogs

We used simulated galaxy data from TNG50, the highest resolution hydrodynamical simulation in the IllustrisTNG Project. The high spatial resolution of TNG50 is necessary for adequately resolving galaxy morphologies and internal structures, but as a result, the simulation volume is small by cosmological standards (a box with ~ 50 Mpc per side). Pillepich et al. (2019) provide an overview of results from TNG50, and Engler et al. (2021) characterize the SMHMR in the TNG simulations in detail.

We downloaded subhalo catalogs from snapshot 99 ($z = 0$), and we include both central and satellite galaxies in our sample. Every dark matter halo in our sample contains a galaxy. M_{halo} is defined as the total mass of gravitationally bound dark matter particles, and $M_\star$ is defined as the total mass of gravitationally bound star particles in the simulation. After administering quality flag cuts, we select galaxies with stellar mass $\log{(M_\star/M_\odot)} > 9.5$. Our parent sample comprises 1,666 galaxies in the TNG50 volume.

2.2. Image cutouts

Rodriguez-Gomez et al. (2019) generate gri-band synthetic image cutouts for all galaxies with $\log{(M_{halo}/M_\odot)} > 9$ by using SKIRT radiative transfer code (Baes et al., 2011; Camps & Baes, 2015) and Bruzual & Charlot (2003) stellar population synthesis models designed to match observations from the Pan-STARRS 3π Steradian Survey (Chambers et al., 2016).

We crop or zero-pad the image cutouts to ensure that they are all the same size $3 \times 224 \times 224$, which is a common image size used in machine learning (Krizhevsky et al., 2012). Each image must be further processed to add realistic observational effects. Following recommendations from Rodriguez-Gomez et al. (2019), we convolve the images with an azimuthally symmetric Gaussian profile to match the Pan-STARRS survey imaging. We adopt a point spread function with 1.11, 1.21, and 1.31 arcsec full width at half-maximum (FWHM) in *i*, *r*, and *g* bands, respectively.

2.3. Morphological features

In addition to generating synthetic galaxy images, Rodriguez-Gomez et al. (2019) provide morphological summary statistics derived from the image cutouts using the `statmorph` library. These morphological features include the radii at 20%, 50%, and 80% of the galaxy light in circular apertures (r_{20}, r_{50}, r_{80}); Petrosian, Sersic, and half-light radii using best-fit elliptical apertures (r_{Petro}, r_{Sersic}, r_{half}); Sersic index (n_{Sersic}); Gini and M_{20} statistics (Lotz et al., 2004); concentration, asymmetry, and smoothness (CAS) statistics (Conselice, 2003); and multimode, intensity, and deviation (MID) statistics (Freeman et al., 2013). We refer the reader to the original works for details on these morphological measurements (see also Section 4 of Rodriguez-Gomez et al., 2019). We remove 117 galaxies that have unreliable morphological features based on a flag in the catalog, leaving 1549 galaxies in our sample.

2.4. Galaxy overdensity

We compute a summary statistic for environmental overdensity, Δ_G, by counting the number of galaxies within some R_{max}:

$$\Delta_G = |\{H : d\,(G, H) < R_{max}\}| \tag{1}$$

We choose $R_{max} = 3$ Mpc, a scale that is most informative for describing the large-scale environment for the IllustrisTNG galaxy-halo connection (Wu et al., 2024). After we apply our selection criteria, the average galaxy density in the TNG50 catalogs is 2.8×10^{-3} Mpc^{-3}.

3. Experimental Design

We aim to estimate the halo mass using progressively more complex techniques. For each experiment, we split the data into the same train, validation, and test split (Section 3.1). We minimize the validation loss in order to select optimization hyperparameters. After we choose these hyperparameters and train our models, we "unblind" the test set and report the results. In Section 3.2, we define our evaluation metrics.

In Section 3.3, we describe the procedure for training baseline models to predict M_{halo} using galaxy properties such as $M_\star$, morphological features, and overdensity parameter. These morphological and overdensity baseline models inform us whether there is extra information (beyond the simple baseline model) that is helpful for predicting halo mass.

We further test whether morphological information can be learned empirically using CNNs (Section 3.4), and whether galaxy environmental information can be learned empirically using GNNs (Section 3.5). Finally, we present a novel combined CNN and GNN model that predicts M_{halo} from the pixel-level image cutouts and galaxy point cloud (Section 3.6).

3.1. Data Split

As is typical in ML problems, the available data are divided into training, validation and testing splits. In many cases, this is sampled randomly to reduce biases, but such a division is not appropriate for spatially correlated data such as graphs; GNNs use information from adjacent galaxies. Thus, a random split of the galaxy sample would cause information from the validation and testing sets to leak into the training set.

To completely separate these subsets of data, all galaxies (i.e., graph nodes) in the training, validation, and testing sets must be well-separated (i.e., they cannot be connected by graph edges). We set a constant linking length (R_{max}) of 3 Mpc during graph creation. We split the data into three contiguous sub-volumes, with a 6 Mpc partition in between each set to guarantee that the training, validation, and testing sets are independent. Figure 1a shows this data split in a two-dimensional projection. The resulting split leaves 1,011 galaxies for training, 112 galaxies for validation and 259 galaxies reserved for testing.

Table 1. Summary of TNG50 features used in our baseline models.

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

Due to our data set split, there is an increased risk that different subsets probe different cosmic environments. In other words, it is possible that the validation (or testing) sets comprise galaxies in a non-representative region of the Universe. This risk is exacerbated by the relatively small cosmic volume in the TNG50 simulation. To mitigate this risk, we check for differences between the halo mass distributions for the three data subsets. In Figure 1b, we show that the M_{halo} distributions for each data set are quite similar. Thus, we proceed with this split and discuss the potential issue further in Section 5.

3.2. Evaluation Metrics

All models are trained to minimize Mean Squared Error (MSE). The test sets are evaluated on MSE, along with Mean Absolute Error (MAE), Root Mean Square Error (RMSE), linear correlation coefficient (R^2), Normalized Median Absolute Deviation (NMAD), bias, and outlier fraction ($f_{outlier}$).

NMAD is an outlier-insensitive metric for the prediction error (or scatter), normalized so that the NMAD of a Gaussian distribution is equal to a standard deviation.

$$\mathrm{NMAD} \approx 1.4826 \times \mathrm{Median}\left(|y - \hat{y} - \mathrm{Median}\left(y - \hat{y}\right)|\right). \tag{2}$$

Bias is defined as the mean offset, and describes whether the parameter is generally overestimated or underestimated.

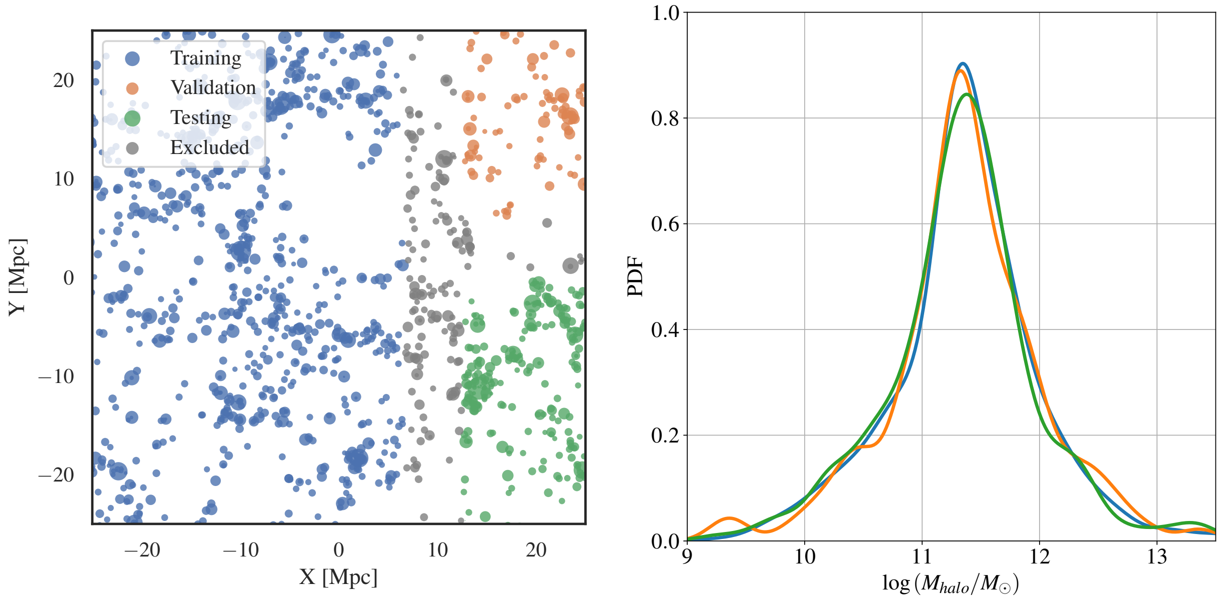
$$\mathrm{bias} = \sum_{i=1}^{N}(\hat{y}_i - y_i) \tag{3}$$

The Outlier Fraction is the ratio of predictions with residuals $> 3\times$ NMAD. This metric determines the rate of catastrophic outliers.

$$f_{outlier} = \frac{1}{N} \cdot \left|\left\{y_i : |\hat{y} - y| > 3 * \mathrm{NMAD}\right\}\right| \tag{4}$$

3.3. Baseline Models

We train several baseline models to facilitate comparisons with deep neural networks; these baseline are not meant to achieve the best possible performance. Instead, they predict halo mass from commonly used galaxy features, such as stellar mass ($M_\star$), morphological parameters (R_{Petro}, A, S),



(a) Each split and excluded galaxies shown in a $2d$ projected point cloud. The marker size is proportional to M_{halo}.

(b) Distributions of galaxy halo masses for data split.

Figure 1. Training, validation, and testing data split figures.

and overdensity (Δ_G). For each baseline model, we train a random forest with 100 estimators using the input features described below. Table 1 summarizes how different baseline models employ the features described in Section 2.

3.3.1. Simple Baseline Model

The simple baseline model is trained using stellar mass as the only feature. The galaxy-halo connection is often expressed as a one-to-one map between galaxy $M_\star$ and M_{halo}, so we expect that our simple baseline model should have some predictive power. However, there are secondary correlations between galaxy and dark matter halo properties (Wechsler & Tinker, 2018), and the simple baseline model will fail to capture those dependencies.

3.3.2. Morphological Baseline Model

The morphological baseline model is trained using several galaxy morphology features in conjunction with stellar mass. We use $k = 5$-fold cross-validation to evaluate which morphological features are most critical to incorporate in our baseline model. Although we consider random forest models with the many morphological features described in Section 2.3, we find no substantial improvement after including Petrosian radius, Smoothness, and Asymmetry. Therefore, we elect to use those three morphological parameters in addition to stellar mass for our morphological baseline model.

3.3.3. Overdensity Baseline Model

There is evidence that galaxy overdensity can tighten the scatter in the SMHMR (e.g., Blanton et al., 2006). Our overdensity model is a random forest that predicts M_{halo} from $M_\star$ and Δ_G.

3.3.4. Combined Baseline Model

We construct a combined baseline model using all features described above: $M_\star$, R_{Petro}, A, S, and Δ_G. Any improvement in prediction for this model, relative to the other baseline models, can be interpreted as evidence that the galaxy morphology and large-scale environment contribute distinct information for estimating the halo mass.

3.4. CNN

3.4.1. Model Architecture

The backbone of the CNN is a ResNet18 (He et al., 2015) pretrained on ImageNet data (Paszke et al., 2019), with the final prediction layer replaced with 100 output features. The CNN output is concatenated with the galaxy's stellar mass, which is input into a 3 layer network to output the M_{halo} estimate. All layers are trainable.

This CNN is trained for 1000 epochs with the AdamW optimizer (Kingma & Ba, 2014; Loshchilov & Hutter, 2019) at a learning rate of $\gamma = 5 \times 10{-}4$, a weight decay of $\lambda = 1 \times 10^{-3}$, and a batch size of 64.

3.4.2. Image Data

We apply point-wise Gaussian noise necessary to model sky background noise, with σ values of $1/15$, $1/19$, and $1/25$ $e^{-}s^{-1}$pixel^{-1} for the i, r, and g band images. To reduce overfitting of the model, multiple simple augmentation techniques are used during model training: horizontal flip, vertical flip, rotation up to $90°$. The image is padded by 5 pixels in every direction and randomly cropped, resulting in a random jitter. While a rigorous ablation study on data augmentation choices is outside the scope of our work, we ran simple validation experiments to determine that these augmentation techniques are helpful for efficiently training our model and reducing overfitting. We finally rescale the pixel values to take on a mean of 0 and standard deviation of 1 in each channel based on our (post-processed) image statistics; the same rescaling is applied to the training, validation, and testing data.

3.5. GNN

3.5.1. Graph Construction

Graphs are well-suited for modeling galaxies in cosmological volumes. Each node in the graph represents a galaxy, with stellar mass as the only node feature. We create edges between two galaxies if they are separated by less than the $R_{max} = 3$ used in the overdensity estimate; this is defined as the linking length of the graph. The Euclidean distance between connected nodes is used as the only edge feature.

After splitting the data into training, validation, and testing sets (see Section 3.1), we batch the training data set into 24 clusters using the METIS algorithm to reduce memory usage (see, e.g., the `ClusterLoader` class in Pytorch-geometric Chiang et al., 2019; Fey & Lenssen, 2019).

3.5.2. Model Architecture

We use a GNN that can pass messages between edges from neighboring nodes, which enables node and edge states to be updated. This architecture permits the GNN to process information from each pair of neighboring nodes as well as the edge. A pooling layer then aggregates all of the information back to each node in order to simultaneously make predictions for every node in the batch.

Our GNN is based on the architecture described in Wu & Jespersen (2023). Our GNN uses 4 parallel networks of fully-connected layers with 16 latent channels and 16 hidden channels, followed by `max` pooling. Each node's output is then concatenated with the original node feature (stellar mass), which is fed into a final three-layer neural network to predict M_{halo}.

We train the GNN for 500 epochs using the AdamW optimizer with a learning rate of $\gamma = 1 \times 10^{-3}$ and weight decay $\lambda = 1 \times 10^{-4}$. The entire validation and testing sets can each fit into a single batch, so we make predictions using the entire galaxy subgraph.

3.6. CNN + GNN

One final model combines the learnable parameters from both the CNN and the GNN. We construct the model using the (initially) frozen pretrained GNN backbone as the GNN component for this combined network. We also initialize the ImageNet-pretrained ResNet18 model for the CNN component, and attach a linear layer with 64 neurons. The GNN and CNN outputs are then concatenated with the stellar mass, which are passed through the same set of final linear layers used for the GNN.[1]

We batch examples using the same clustering algorithm that was used for the GNN model, which requires collating the same image examples for the CNN portion for model. This combined CNN+GNN is trained for 500 epochs using the AdamW optimizer with a learning rate of $\gamma = 5 \times 10^{-4}$ and $\lambda = 1 \times 10^{-3}$, but because the CNN and GNN are initialized to pretrained weights, we use an early stopping criterion to refine the optimization procedure. If the validation MSE loss rises 0.005 above the minimum validation loss five times without finding a new minimum validation loss, then we re-enable learning on (i.e., unfreeze) the GNN backbone and reduce the learning rate by a factor of 5. If a new minimum validation loss is achieved, then the counter is reset to 0, and we proceed with training just the CNN. Unfreezing the GNN and reducing the learning rate only occurs once; afterwards, we continue training for the remainder of the 500 epochs.

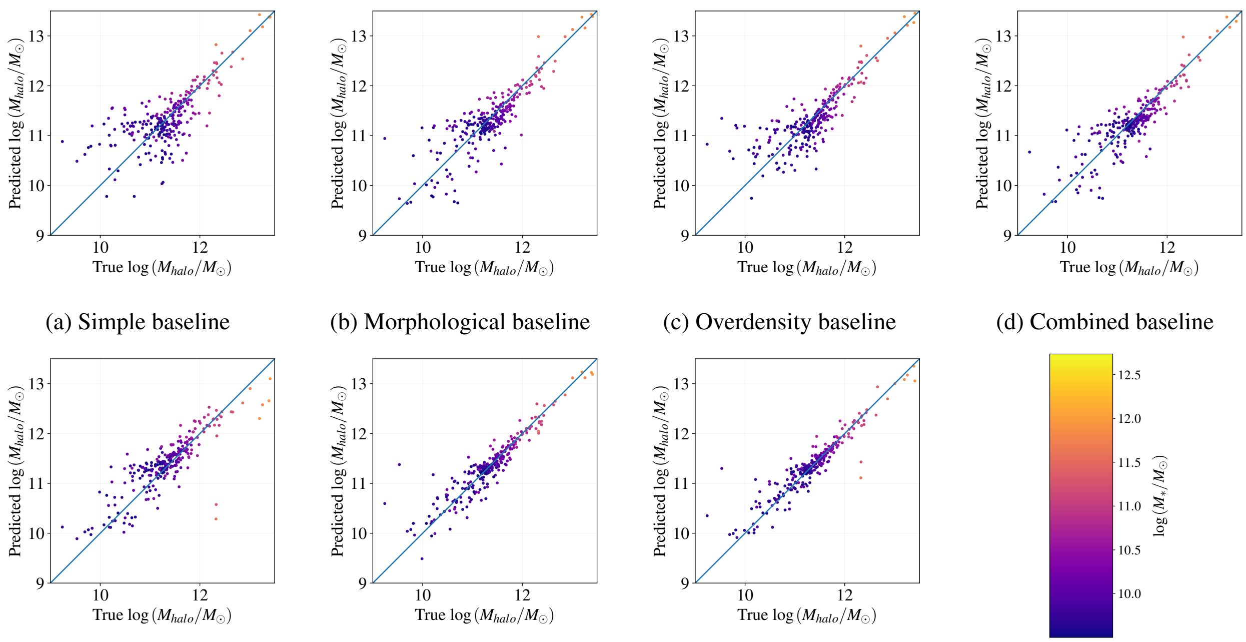
4. Results

Our test set results are summarized in Figure 2 and Table 2. The figure shows predicted versus true halo masses (colored by the stellar mass) for all models. The table compares the evaluation metrics (described in Section 3.2) for all models. Additional details about the neural network training and validation loss curves are presented in Appendix A.

4.1. Baseline Models

The simple baseline—a Random Forest using only stellar mass—performed the worst (RMSE $= 0.455$), as expected. The morphological baseline model performs significantly better, with RMSE $= 0.345$. There is clearly information encoded in these morphological features that cannot be captured by stellar mass alone. The overdensity baseline model

[1]The CNN outputs are not passed in as graph node features, as this was too computationally prohibitive for our experiments. Such a model may permit such a model to learn "interactions" between neighboring galaxies' image features.



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

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

Figure 2. True vs. predicted M_{halo} plots for all models, with color scale for $M_{\star}$.

performs better than the simple baseline (RMSE $= 0.374$), but worse than the morphological baseline model. From Figure 2c, we can see that it overpredicts the halo mass for most lower-mass systems, i.e., at true $\log(M_{halo}/M_{\odot}) < 10.5$. Finally, including both overdensity and the morphological features in the combined baseline model further reduces losses (RMSE $= 0.310$).

4.2. CNN

Surprisingly, we find that the CNN results in a poor RMSE (0.359) compared to the morphological baseline model (0.345), and suffers from a very high outlier fraction ($f_{outlier} = 6.18\%$). These outliers are visually apparent in Figure 2e, wherein the CNN fails to predict M_{halo} for a few galaxies in the test set (and fails spectacularly for two galaxies in particular). However, the MAE and NMAD is lower for the CNN than than for the morphological baseline model, suggesting that the non-outlier predictions by the CNN are still reasonably accurate.

4.3. GNN

The GNN predicts with much higher accuracy than the overdensity baseline, as evident in the metrics. This is likely the highest performing model in this experiment, even though ResNet18 has hundreds of times more trainable parameters than the GNN. Relative to the previous models, the GNN has very few outliers and a low bias, indicating that the data it received does not differ largely from the training set.

4.4. CNN + GNN

The combined CNN+GNN model performs better than either the CNN and GNN separately, although the RMSE and R^2 values are very similar to those from the GNN. Again, we see the same two strong outliers (Figure 2g) as we did in the CNN-only model. However, the CNN+GNN still manages to achieve the best $f_{outlier}$, indicating that most of its predictions are very accurate. This is also supported by its low MAE and NMAD.

5. Discussion

5.1. Comparing neural networks against baseline models

Results from our baseline models demonstrate that galaxy morphology and environment are valuable for improving M_{halo} predictions for TNG50 galaxies. Moreover, these features can be combined to achieve even btter results. The baseline models use “handcrafted” features that are commonly used in astronomy (e.g., Conselice 2004), but these features may not summarize all the information in galaxy images or environments that are relevant for constraining M_{halo}. The successes of deep learning offers a hint that CNNs and GNNs may be able to extract more useful information directly from images and point clouds. Here, we interpret our results from more sophisticated CNNs and GNNs.

Table 2. Test results from baseline and neural network models as characterized by the evaluation metrics described in Section 3.2. The best metrics are underlined.

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%

We find that the neural network models generally perform better than the baseline models (Table 2), confirming our previous hypothesis. This is despite the fact that highly overparameterized neural networks can struggle to learn from small data sets (see Section 5.3 for more discussion). Nonetheless, the CNN still 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.[2] The CNN+GNN model achieves the best performance metrics across the board—but it could probably perform even better if not for the CNN's catastrophic failures.

5.2. Successes and failures of neural networks

The CNN primarily fails in two different ways. First, the halo masses of high-mass ($M_\star > 10^{12.5} M_\odot$) galaxies are consistently underpredicted. The latter failure mode can be explained by small TNG50 volume, which inhibits the CNN—with its $\mathcal{O}(10^7)$ trainable parameters—from adequately learning how to handle these rare massive galaxies. Second, there are two galaxy samples in the test set that have dramatically underpredicted halo masses (see, e.g., Fig 2e). We have visually inspected these galaxy images, and found that one of them is a pair of interacting galaxies. As we note in Appendix A (see Figure 3), the validation loss achieves a minimum value after only 116 epochs of training, which suggests that the CNN is undertrained. The validation loss of MSE $= 0.0894$ is much better than test loss MSE $= 0.129$; thus we surmise that the CNN did not fully converge, and would benefit from more training examples.

It is interesting that the GNN demonstrates such strong performance in predicting M_{halo}. One interpretation of this result is that the large-scale environment is more informative for predicting M_{halo} than the detailed galaxy appearances. However, the morphological baseline model *does* outperform the overdensity baseline model, which may imply that our overdensity parameter Δ_G is too simplistic to describe the overall environment. Indeed, Wu et al. (2024) find that GNNs are better suited than overdensity for describing environmental dependence on the relationship between galaxies and their dark matter halos, which reinforces the idea that GNNs are better equipped to extract environmental information from galaxy point cloud data.

5.3. The challenge of small data sets

We have mentioned that our data split reduces the galaxy data set to only a about a thousand examples for training, few hundred examples each for validation and testing. The limited data sets raise several challenges. First, neural networks typically benefit from more training data, so it is likely that our deep learning models are not fully converged. Second, the small subsets of data for validation or testing can increase the variance in prediction results; it is conceivable that a nested cross-validation can help ameliorate this issue. Third, due to our choice to separate galaxies into contiguous subvolumes, the validation or testing data sets can exhibit different galaxy properties on average than the ones seen in the training data set.

We anticipate that our neural networks would perform even better or larger data volumes. However, increasing the simulation box size often comes at the cost of lowering the resolution, which can have an adverse effect on simulating realistic galaxy appearances at all.

5.4. Application to real data and domain adaptation

Our successful experiments indicate that we may be able to better predict galaxy properties by leveraging galaxy imaging and their cosmic surroundings. Can we immediately apply this to real observations? Unfortunately, the answer is likely not. Our models have learned the very specific characteristics of simulated data from TNG50, which differs from the real Universe in myriad ways. In general, this problem is known as "domain shift" or "data set shift" and it

[2] We also note that the overdensity baseline model performs very poorly for low-mass galaxies, which may indicate challenges with predicting M_{halo} for satellite galaxies residing in high-density environments.

applies to any kind of shift in the high-dimensional inputs to deep learning models. Not only does domain shift prevent us from successfully applying a model trained on TNG50 to the real Universe, but it also prevents us from applying this model trained on TNG50 to another simulation. There exist proposed solutions for mitigating this effect ("domain adaptation", see e.g., Csurka, 2017), which have been applied to astronomical data (Ćiprijanović et al., 2023), but the problem is far from solved. Therefore, we caution against naively using these ML models to predict M_{halo} for real galaxies.

6. Conclusions

We have estimated the dark matter halo mass, M_{halo}, using galaxy data from the Illustris TNG50-1 hydrodynamic simulations. We split our sample of well-resolved galaxies with reliable morphological measurements into spatially separated regions for the training ($N = 1011$), validation ($N = 112$), and testing ($N = 259$) data subsets. All results are presented using metrics from the testing data set.

We first train baseline random forest models that use stellar mass ($M_\star$), morphological features (Petrosian radius, asymmetry, and smoothness), and galaxy overdensity (Δ_G) over 3 Mpc scales. From our baseline model investigation, we find that galaxy morphology and overdensity are useful features for accurately estimating M_{halo}. Moreover, the combined set of features produces an even more performant model (Table 2).

After we confirm that morphology and overdensity are important for predicting dark matter halo mass, we use synthetic galaxy image cutouts and galaxy $3d$ point cloud as inputs to neural network models. In other words, we eschew summary statistics in favor of directly learning from the pixel-level and point cloud data (in addition to the stellar mass). Our conclusions for the neural network models are listed below:

1. While a deep convolutional neural network (CNN) can achieve strong performance on the training and validation set (RMSE $= 0.299$), its performance on the test set suggests that it has been overfit to the small data set (RMSE $= 0.359$, which is worse than the morphological baseline model result—RMSE $= 0.345$). This interpretation is also supported by its extremely poor performance on two dramatic outliers in the test set (see Figure 2e) and its large outlier fraction ($f_{outlier} = 6.18\%$). Nevertheless, the CNN outperforms the morphological baseline well in terms of outlier-insensitive metrics like mean absolute error (MAE $= 0.238$) and normalized median absolute deviation (NMAD $= 0.218$).
2. The graph neural network (GNN), trained on cosmic graphs connected by a 3 Mpc linking length, achieves very low prediction error (RMSE $= 0.248$) relative to the overdensity baseline model (RMSE $= 0.374$) and combined baseline model (RMSE $= 0.310$), which suggests that galaxy environment is particularly important for constraining the halo mass.
3. We have trained a novel CNN+GNN joint model that achieves the best performance overall. Because the CNN component of the model suffers from the overfitting issue described above (see also Figure 2g), the CNN+GNN model is comparable to the GNN model in terms of RMSE and R^2; however, its outlier-insensitive metrics (MAE $= 0.144$, NMAD $= 0.123$) are far superior to any other models' performance.

Our results demonstrate that a CNN+GNN model is capable of jointly extracting detailed information from galaxy appearances and large-scale environments. These results are promising for future data-driven approaches to predicting dark matter properties from galaxies, particularly if they can be trained on much larger galaxy samples spanning larger cosmic volumes (see Section 5.3). However, we also caution against applying these models trained on simulation data to real observations without first accounting for domain shift (Section 5.4).

References

Baes, M., Verstappen, J., De Looze, I., Fritz, J., Saftly, W., Vidal Pérez, E., Stalevski, M., and Valcke, S. Efficient Three-dimensional NLTE Dust Radiative Transfer with SKIRT. ApJS, 196(2):22, October 2011. doi: 10.1088/0067-0049/196/2/22.

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.

Blanton, M. R., Eisenstein, D., Hogg, D. W., and Zehavi, I. The Scale Dependence of Relative Galaxy Bias: Encouragement for the "Halo Model" Description. ApJ, 645(2):977–985, July 2006. doi: 10.1086/500918.

Bruzual, G. and Charlot, S. Stellar population synthesis at the resolution of 2003. MNRAS, 344(4):1000–1028, October 2003. doi: 10.1046/j.1365-8711.2003.06897.x.

Camps, P. and Baes, M. SKIRT: An advanced dust radiative transfer code with a user-friendly architecture. *Astronomy and Computing*, 9:20–33, March 2015. doi: 10.1016/j.ascom.2014.10.004.

Chambers, K. C., Magnier, E. A., Metcalfe, N., Flewelling, H. A., Huber, M. E., Waters, C. Z., Denneau, L., Draper,

P. W., Farrow, D., Finkbeiner, D. P., Holmberg, C., Koppenhoefer, J., Price, P. A., Rest, A., Saglia, R. P., Schlafly, E. F., Smartt, S. J., Sweeney, W., Wainscoat, R. J., Burgett, W. S., Chastel, S., Grav, T., Heasley, J. N., Hodapp, K. W., Jedicke, R., Kaiser, N., Kudritzki, R. P., Luppino, G. A., Lupton, R. H., Monet, D. G., Morgan, J. S., Onaka, P. M., Shiao, B., Stubbs, C. W., Tonry, J. L., White, R., Bañados, E., Bell, E. F., Bender, R., Bernard, E. J., Boegner, M., Boffi, F., Botticella, M. T., Calamida, A., Casertano, S., Chen, W. P., Chen, X., Cole, S., Deacon, N., Frenk, C., Fitzsimmons, A., Gezari, S., Gibbs, V., Goessl, C., Goggia, T., Gourgue, R., Goldman, B., Grant, P., Grebel, E. K., Hambly, N. C., Hasinger, G., Heavens, A. F., Heckman, T. M., Henderson, R., Henning, T., Holman, M., Hopp, U., Ip, W. H., Isani, S., Jackson, M., Keyes, C. D., Koekemoer, A. M., Kotak, R., Le, D., Liska, D., Long, K. S., Lucey, J. R., Liu, M., Martin, N. F., Masci, G., McLean, B., Mindel, E., Misra, P., Morganson, E., Murphy, D. N. A., Obaika, A., Narayan, G., Nieto-Santisteban, M. A., Norberg, P., Peacock, J. A., Pier, E. A., Postman, M., Primak, N., Rae, C., Rai, A., Riess, A., Riffeser, A., Rix, H. W., Röser, S., Russel, R., Rutz, L., Schilbach, E., Schultz, A. S. B., Scolnic, D., Strolger, L., Szalay, A., Seitz, S., Small, E., Smith, K. W., Soderblom, D. R., Taylor, P., Thomson, R., Taylor, A. N., Thakar, A. R., Thiel, J., Thilker, D., Unger, D., Urata, Y., Valenti, J., Wagner, J., Walder, T., Walter, F., Watters, S. P., Werner, S., Wood-Vasey, W. M., and Wyse, R. The Pan-STARRS1 Surveys. *arXiv e-prints*, art. arXiv:1612.05560, December 2016. doi: 10.48550/arXiv.1612.05560.

Chiang, W., Liu, X., Si, S., Li, Y., Bengio, S., and Hsieh, C. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. *CoRR*, abs/1905.07953, 2019. URL http://arxiv.org/abs/1905.07953.

Ćiprijanović, A., Lewis, A., Pedro, K., Madireddy, S., Nord, B., Perdue, G. N., and Wild, S. M. DeepAstroUDA: semi-supervised universal domain adaptation for cross-survey galaxy morphology classification and anomaly detection. *Machine Learning: Science and Technology*, 4 (2):025013, June 2023. doi: 10.1088/2632-2153/acca5f.

Clowe, D., Bradač, M., Gonzalez, A. H., Markevitch, M., Randall, S. W., Jones, C., and Zaritsky, D. A Direct Empirical Proof of the Existence of Dark Matter. ApJ, 648 (2):L109–L113, September 2006. doi: 10.1086/508162.

Conselice, C. J. The relationship between stellar light distributions of galaxies and their formation histories. *The Astrophysical Journal Supplement Series*, 147(1):1–28, July 2003. ISSN 1538-4365. doi: 10.1086/375001. URL http://dx.doi.org/10.1086/375001.

Csurka, G. *A Comprehensive Survey on Domain Adaptation for Visual Applications*, pp. 1–35. Springer International Publishing, Cham, 2017. ISBN 978-3-319-58347-1. doi: 10.1007/978-3-319-58347-1_1. URL https://doi.org/10.1007/978-3-319-58347-1_1.

Engler, C., Pillepich, A., Joshi, G. D., Nelson, D., Pasquali, A., Grebel, E. K., Lisker, T., Zinger, E., Donnari, M., Marinacci, F., Vogelsberger, M., and Hernquist, L. The distinct stellar-to-halo mass relations of satellite and central galaxies: insights from the IllustrisTNG simulations. MNRAS, 500(3):3957–3975, January 2021. doi: 10.1093/mnras/staa3505.

Fey, M. and Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Freeman, P. E., Izbicki, R., Lee, A. B., Newman, J. A., Conselice, C. J., Koekemoer, A. M., Lotz, J. M., and Mozena, M. New image statistics for detecting disturbed galaxy morphologies at high redshift. MNRAS, 434(1): 282–295, September 2013. doi: 10.1093/mnras/stt1016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

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, 2014.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019.

Lotz, J. M., Primack, J., and Madau, P. A New Nonparametric Approach to Galaxy Morphological Classification. AJ, 128(1):163–182, July 2004. doi: 10.1086/421849.

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 *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., 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.

Rodriguez-Gomez, V., Snyder, G. F., Lotz, J. M., Nelson, D., Pillepich, A., Springel, V., Genel, S., Weinberger, R., Tacchella, S., Pakmor, R., Torrey, P., Marinacci, F., Vogelsberger, M., Hernquist, L., and Thilker, D. A. The optical morphologies of galaxies in the IllustrisTNG simulation: a comparison to Pan-STARRS observations. MNRAS, 483(3):4140–4159, March 2019. doi: 10.1093/mnras/sty3345.

Rubin, V. C., Ford, W. K., J., and Thonnard, N. Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 (R=4kpc) to UGC 2885 (R=122kpc). ApJ, 238:471–487, June 1980. doi: 10.1086/158003.

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.

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.

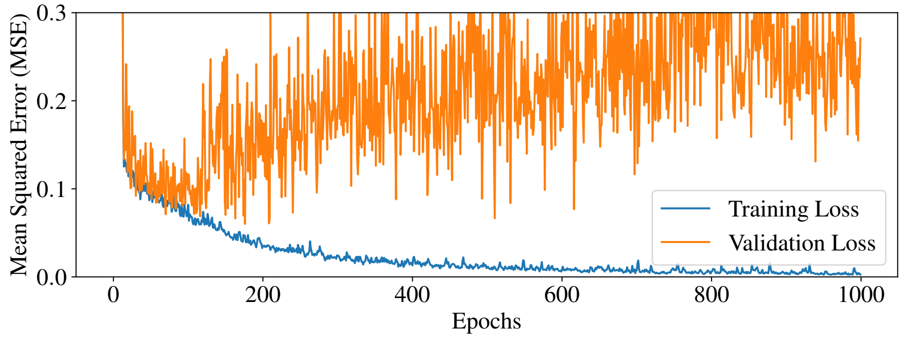
Wu, J. F. and Jespersen, C. K. Learning the galaxy-environment connection with graph neural networks, 2023.

Wu, J. F., Jespersen, C. K., and Wechsler, R. H. How the galaxy-halo connection depends on large-scale environment, 2024.

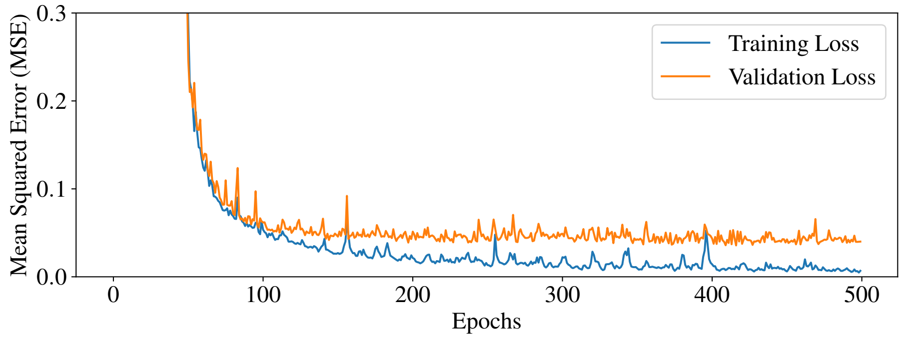
Zwicky, F. Die Rotverschiebung von extragalaktischen Nebeln. *Helvetica Physica Acta*, 6:110–127, January 1933.

A. Training and validation loss curves

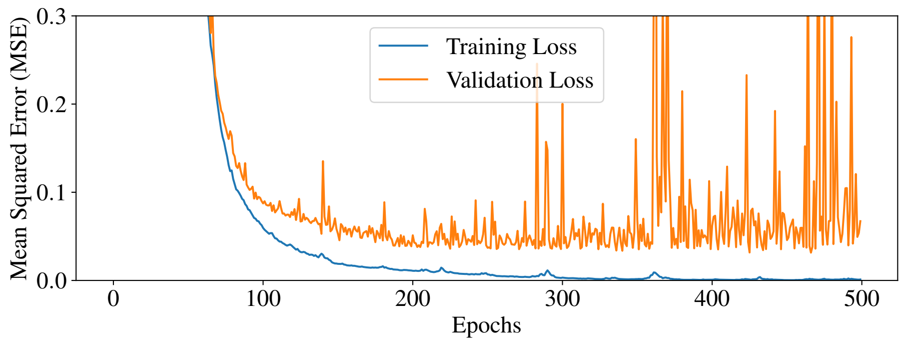
In Figure 3, we show the training and validation loss curves from optimizing the CNN, GNN, and CNN+GNN models via the training procedures described in Section 3. To make predictions, we save the model checkpoints that achieve lowest validation losses, but here we show the entire loss curves (out to 500 epochs) in order to gauge the level of over- or under-fitting.



(a) CNN: minimum validation loss during epoch 116.



(b) GNN: minimum validation loss during epoch 416.



(c) CNN+GNN: After epoch 144, we unfroze the GNN component and lowered the CNN learning rate (see description of optimization procedure in Section 3.6). Minimum validation loss during epoch 466.

Figure 3. Training and validation loss curves for our deep learning models. The subfigure caption indicates which epoch minimized the validation loss; that checkpointed model is then used to make final test set predictions.

Figure 3a shows that the CNN validation loss curve begins to increase shortly after the minimum validation epoch (116) while the training loss continues to decrease. The validation loss is quite noisy, which suggests that it may not be a reliable indicator of model convergence. Thus, the CNN's modest results are unsurprising (e.g., Section 5.2).

The GNN and CNN+GNN loss curves (in panels b and c) seem to indicate that there is a substantial gap between training and validation losses. However, when optimizing GNNs, it is common to find that the training and validation loss curves plateau rather than diverge—which would signify a problem with overfitting. This is likely because GNNs have far fewer trainable parameters than CNNs, and are therefore less susceptible to overfitting.