Building Novel Neural Networks to Unveil the Connection between Galaxies, their Dark Matter Halos, and their Environment

Abstract

According to Λ Cold Dark Matter (ΛCDM) theory, the physical processes governing a galaxy's evolution are regulated by its host dark matter halo, which, in turn, is regulated by the larger-scale environment. In particular, simulations reveal correlations between environment and specific halo properties that dictate the galaxy's evolution, such as halo mass, density distribution, angular momentum, accretion rate, and merger history. However, traditional models struggle to capture the complexity of the connection between a galaxy, its dark matter halo, and the larger-scale environment. **We propose to create the Nearest-Neighbors Neural Network (NN^2) to extract observable information embedded in a galaxy's environment (e.g., the locations and stellar masses of a galaxy's nearest neighbors) and predict the properties of the host halo of a given galaxy.** The network will be trained on mock galaxy-catalogs from UNIVERSEMACHINE. We will then verify the network's accuracy by comparing its predictions to the known properties of the simulated halos. The trained network will be applied to the 3D-HST survey to infer the properties of the halos within. Through this, the network will: 1) reveal new connections between galaxy, halo, and environment; 2) serve as a powerful tool for placing galaxies into halos in future cosmological simulations; and 3) be a framework for inferring the properties of real halos from next-generation survey data, allowing for direct comparison between observational statistics and theory.

- **Scientific Justification**

The Large-Scale Environment Traces Dark Matter Halo Properties

In modern cold dark matter models, all galaxies form in dark matter overdensities called halos. Within this framework, the process of a galaxy's formation and evolution is tied to the halo in which it forms. This galaxy-halo connection defines many properties of the galaxy (see Somerville & Davé, 2015 for a review). Observations reveal a strong correlation between galaxy mass and host halo mass (Tasitsiomi et al., 2004). However, there is significant diversity in galaxy properties like shape and size at fixed halo mass (Naab & Ostriker, 2017), suggesting secondary halo properties such as density profile, angular momentum, mass accretion history, and merger history also play an important role in determining a galaxy's properties. For example, galaxy size has been theorized to depend on many halo properties, including: 1) virial radius (Hearin et al., 2017); 2) angular momentum (Hearin et al., 2019); 3) growth rate (Behroozi et al., 2022); and 4) density profile (Jiang et al. 2019). Other galaxy properties are similarly predicted to correlate with secondary halo properties (Wecshler & Tinker, 2018 and references therein).

While the nature of dark matter makes it challenging to constrain secondary halo properties directly, the large-scale environment is a useful a proxy. Secondary halo properties are known to correlate with the spatial distribution of halos (Mao et al., 2018). Within simulations, grouping halos by different properties produces unique clustering biases, known as assembly biases. For example, halos with higher concentrations generally occur in denser environments (Behroozi et al., 2022). As halos contain galaxies, halo clustering biases are observable through the resultant galaxy clustering (Wecshler and Tinker 2018). *Hence, different secondary halo properties are strongly suggested to leave unique and observationally measurable imprints on their environment.*

Including the available information in the environment is essential for physically accurate galaxy modeling. However, the interplay between galaxy, halo, and environment is highly complex and not fully captured by modern galaxy evolution models. There are no existing models that are both *flexible* enough to capture the true complexity of these connections and *not overly prescriptive* of the physical processes regulating the connection. Excitingly, a machine learning presents an avenue for exploring these connections with great flexibility and no requirements for assumptions about the underlying physics.

We propose to create the Nearest-Neighbors Neural Network (NN^2), a machine learning algorithm that will predict the host-halo properties of galaxies in the 3D-HST survey. The network will extract halo masses, density profiles, angular momenta, growth rates, and merger histories from observable information embedded in the galactic environment. The predicted halo properties will provide constraints on the galaxy-halo connection, addressing the long-standing problem of the role of the host-halo in the formation and evolution of a galaxy. NN^2 will be easily expanded to cover additional observational surveys to provide additional constraints as new data becomes available.

The complete, multi-variate connections between galaxy, halo, and environment are not constrained

Historically, models of the galaxy-halo connection focused on describing the relationship between one galaxy property (typically mass or luminosity) and one halo property (e.g., Conroy et al. 2006, Behroozi et al. 2010). These simple empirical models described the basic relationship well but provided inadequate descriptions of the detailed spatial distribution of galaxies as shown

by increasingly precise measurements and the diversity of galaxy properties at fixed mass (Wechsler & Tinker 2018). In reality, the galaxy-halo connection is far more complicated with multi-variate connections between galaxy properties, halo properties, and environment (e.g., Salcedo et al. 2018). For example, galaxy size likely depends on several halo properties and environmental factors, as illustrated in Figure 1. Determining the influence of the host-halo and the environment on galaxy properties, like size and shape, is essential for fully understanding the mechanics of galaxy formation and for making realistic galaxy models.

However, modern empirical and physical models have limited flexibility for capturing the details of the galaxy-halo connection and environmental biases due to requiring assumptions about the functional forms and underlying physics of these relationships, respectively. As observational constraints on galaxy clustering and observable properties continue to improve, there is a need for sufficiently complex models that can take advantage of new constraints to better understand and simulate galaxies. A highly flexible model, such as that proposed here, would take full advantage of available observational constraints to capture both simple and complex connections, including connections that have not yet been considered.

Understanding the galaxy-halo connection requires connecting observations to simulations

3D-HST is a survey of ~10,000 galaxies over $1<z<3.5$, designed to study galaxy evolution (Suess et al., 2019a,b). Using near-infrared spectroscopic observations, the survey provides galaxy redshifts – the essential third dimension needed to measure the spatial distribution of galaxies. In addition, the survey has the necessary resolution to measure galaxy properties like shape and size (Brammer et al., 2012). While the survey seeks to study the physical processes shaping galaxies, including the role of environment, the dearth of information about the dark matter halos that host these galaxies is apparent. Yet, the presence of detailed spatial information makes it possible to assess halo properties.

In order to learn about these halo properties, we need to connect observations to theoretical models in which halo properties are known. One such model is UNIVERSEMACHINE (Behroozi et al., 2019), an empirical model of the galaxy-halo connection. Given an underlying dark-matter only simulation, UNIVERSEMACHINE self-consistently follows the evolution of halos and galaxies over cosmic time and produces a catalog of galaxies assigned to halos. The parametrization of the empirical model is constrained using global observational statistics, including stellar mass functions and the cosmic star formation rate density. This method has been successful in reproducing the observed stellar mass-halo mass relation for peak halo masses between $10^8\ M_\odot$ and $10^{15}\ M_\odot$ (Yunchong et al., 2021). There is, however, a challenge in connecting observations and mock catalogs, mainly that we do not know the dark matter halo properties of the galaxies in the observational sample. We propose to train a neural network to recognize the halo-environment connection in a mock catalog produced by UNIVERSEMACHINE and apply this relationship to the galaxies in 3D-HST to learn about their host dark matter halos.

A neural network can capture the complexity of the relationship between halo properties and environment

A machine learning approach to connecting observational (3D-HST) and theoretical (UNIVERSEMACHINE galaxy catalogs) data is more flexibile than traditional empirical modeling approaches. Empirical models may iterate over many possible formulations to best fit global observational statistics, but this is an inefficient process that is not guaranteed to find a

good fit within the model's constraints (Moster et al., 2021). Hence, we can use neural networks to more aptly constrain the connection between galaxies, halos, and their environment.

To train a neural network to recover halo properties from available observational data, we first need a dataset where halo properties are known. NN^2 will originally be trained on existing mock galaxy catalogs, with a focus on using observable clustering information (as simulated) to predict halo properties. To avoid inheriting the mock catalogs assumptions about the galaxy-halo connection, the network will not be provided any information about secondary galaxy properties from UNIVERSEMACHINE and will instead only rely on the well-supported spatial and stellar mass data.

A portion of the catalogs will be set aside for validation and testing purposes, which will be used to determine how well the network reproduces the mock catalog. Figure 3 shows preliminary results comparing the predictions of NN^2 against the mock catalog. The finalized network will be easily generalizable to simulations beyond UNIVERSEMACHINE and be applicable to observational data. This is an efficient method for approximating halo properties, where they are unknown (and likely cannot be measured).

After training, the network will be applied to archival survey data (3D-HST). Given the rudimentary inputs required for the network, and the high-precision three-dimensional data from the survey, the network is expected to produce reasonable halo property estimates for a majority of the ~10,000 galaxies in the sample. The survey data also contains information about galaxy shapes and sizes, among other properties, so that the application of the neural network will constrain the galaxy-halo connection (e.g., does halo angular momentum correlate with galaxy size?) in addition to the halo-environment correlations modeled by the network itself.

Applications of halo property predictions to galaxy modeling and beyond

The main products of this proposal are halo property predictions for galaxies in 3D-HST and the network that predicts them, which will be easily generalizable to similar observational or theoretical galaxy catalogs. Currently, halo properties are a critical missing component in analyzing the 3D-HST survey, which is focused on constraining galaxy physics. Together the predicted halo properties and observed galaxy properties, will put constraints on the galaxy-halo connection. This proposal's approach opens a whole new avenue for exploring the effect of dark matter halo assembly on galaxy properties, and ultimately, improving future galaxy models.

Beyond the study of the GHC, halo property data has implications for the physics of galaxy formation and evolution, evaluating the role of the cosmic ecosystem at different scales, and even jointly constraining galaxy growth and cosmology. Hence, adding halo property data to 3D-HST will serve the community at large by incorporating an important element in understanding galaxy physics and extending the applications of the survey data outside the direct study of galaxy properties.

References:
Behroozi, P.+2019, MNRAS, 488, 3143
Behroozi, P.+2022, MNRAS, 509, 2800
Brammer, G.+2012, ApJS, 200, 13
Crain, R.A.+ 2015, MNRAS, 450, 1937
Hearin, A.P.+2017, MNRAS, 435, 1313
Hearin, A.+2019, MNRAS, 489, 1805
Henriques, B.M.B.+2015, MNRAS, 451, 2663
Jiang, F.+2017, MNRAS, 472, 657
Mao, Y.Y.+2018, MNRAS, 474, 5143
Moster, B.+2021, MNRAS, 507, 2215
Naab, T., Ostriker, J.P. 2017, ARA&A, 55, 59
Somerville, R.S., Davé, R. 2015. ARA&A, 53,51
Suess, K.+2019a, ApJ, 877, 103
Suess, K.+2019b, ApJL, 885, L22
Tasitsiomi, A.+2004, ApJ, 614, 533
Weschler, R. H., Tinker J. L., 2018 ARA&A, 56, 435
Yunchong, W.+2021, ApJ, 915, 116

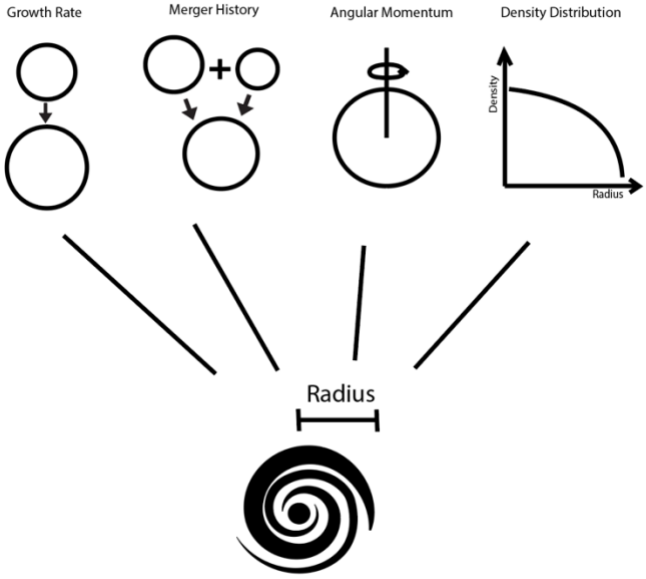

Fig. 1 The Galaxy-Halo Connection: There are complicated connections between galaxy size (i.e., radius) and secondary halo properties. These secondary halo properties include: 1) growth rate – the rate at which halo accretes mass; 2) merger history – particularly when the last major merger occurred; 3) angular momentum; and 4) density distribution. The physical nature and functional form of the relationship between these properties and galaxy size are largely unknown. Hence, the connections between galaxy and halo properties have not been adequately constrained.

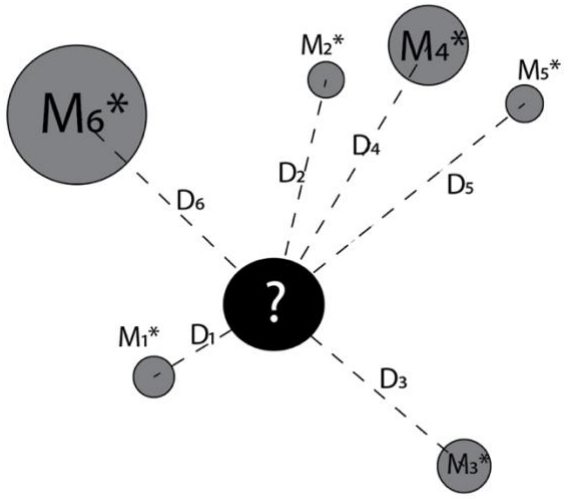

Fig. 2 The Galactic Environment: The properties of a given halo are correlated with the density of halos in their environment. Given the projected 2D-distances to a halo's neighbors and the stellar masses of the galaxies they host, we can predict the secondary properties of the mystery halo. A neural network can be trained to find these environmental correlations in mock catalogs where halo properties are known. *Once trained, the neural network would be able to predict halo properties for galaxies from 3D-HST with only the available information in the environment.*

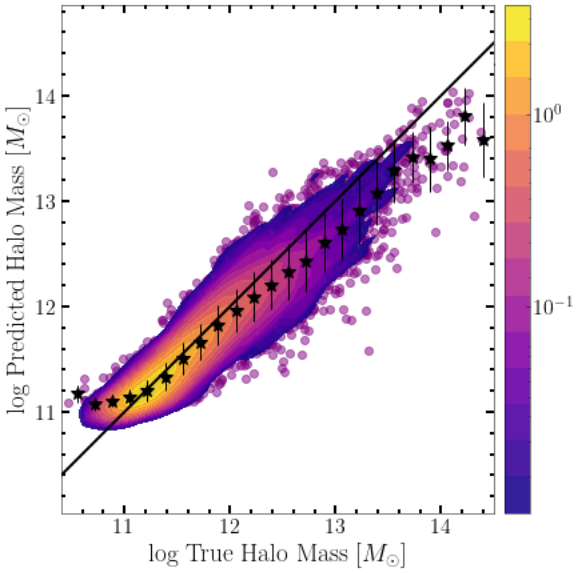

Fig. 3 Halo Mass Predictions: Halo properties predicted by the network are compared against values from the simulated catalog. This figure shows preliminary results for a network trained to predict halo mass evaluated on a sample of data it has not seen before. The mass predicted by the network is plotted against the true halo mass from the catalog. The black line shows a perfect match between prediction and label, the color bar represents the density of points at a given location, and black stars show average values grouped in bins of predicted halo mass. From this figure, we can see that the preliminary network can reproduce simulated halo masses to high accuracy.