Draft version May 24, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

Predicting Halo Mass Function Proxies from Merger Tree Distributions using a Hybrid GNN and Gaussian Mixture Model

AstroPilot[1]

[1] *Anthropic, Gemini & OpenAI servers. Planet Earth.*

ABSTRACT

The dark matter halo mass function (HMF) is a fundamental cosmological probe, reflecting the number density of dark matter halos as a function of mass, and is intimately linked to the hierarchical assembly histories encoded in merger trees. This work presents a novel machine learning approach to predict a proxy for the HMF directly from the distribution of merger trees, leveraging the power of graph neural networks (GNNs) and Gaussian mixture models (GMMs). We train a GNN to generate latent embeddings of individual merger trees, capturing their structural and nodal properties, using a dataset of 1000 trees derived from cosmological N-body simulations. Each tree is represented as a graph with node features including halo mass, concentration, maximum circular velocity, and scale factor. The distribution of these embeddings is then modeled using a GMM to cluster the trees into distinct populations. Subsequently, a feedforward neural network (FFNN) is trained to predict an HMF proxy, specifically, a histogram of halo masses within each tree, from the posterior probabilities of the GMM components. Our results demonstrate that the GNN embeddings effectively capture cosmologically relevant information, as evidenced by their ability to predict cosmological parameters in a pretext task. Furthermore, the GMM successfully clusters trees into distinct populations, and the FFNN achieves a mean squared error of 0.000522 on the test set when predicting the HMF proxy. This performance indicates that the GMM posterior probabilities are informative features for predicting the internal mass distribution of halos as represented in the merger trees. This hybrid approach provides a promising avenue for extracting complex information from merger trees and linking it to halo properties, offering a computationally efficient way to emulate aspects of halo populations.

Keywords: Intergalactic medium, Stellar physics, Gravitational lensing, Nucleosynthesis, Galaxy mergers

1. INTRODUCTION

The dark matter halo mass function (HMF), which quantifies the number density of dark matter halos as a function of their mass, is a fundamental prediction of cosmological models. As such, it serves as a critical link between theoretical frameworks and observational data concerning galaxy formation and the large-scale structure of the universe. The HMF is intrinsically tied to the hierarchical assembly of dark matter halos, a process elegantly captured by merger trees that trace the formation history of individual halos. Establishing a robust connection between merger tree characteristics and the HMF is crucial for interpreting astronomical observations and refining our understanding of the underlying cosmological parameters. However, directly predicting the HMF from the complex and stochastic nature of halo merger histories presents a significant challenge. Traditional methods often rely on computationally expensive N-body simulations or analytical approximations that may oversimplify the intricate details of halo formation. This difficulty stems from the high dimensionality and non-linearity inherent in merger tree data, making it difficult to extract meaningful information and relate it to the global properties of halo populations.

In this paper, we introduce a novel machine-learning approach to predict a proxy for the HMF directly from the *distribution* of merger trees. Instead of focusing on individual trees in isolation, our method aims to leverage the statistical properties of an ensemble of trees to infer the HMF. We hypothesize that the distribution of merger trees contains valuable information about the overall mass distribution of halos. Our approach combines the strengths of graph neural networks (GNNs) and Gaussian mixture models (GMMs). First, we train a GNN to generate a low-dimensional latent embedding for each merger tree, effectively capturing

its structural and nodal properties. Each tree is represented as a graph, where nodes correspond to halos and edges represent merger events. The node features include key physical characteristics such as halo mass, concentration, maximum circular velocity (v_{max}), and scale factor, all of which describe the halos at different stages of their evolution. The distribution of these embeddings is then modeled using a GMM, which aims to cluster the trees into distinct populations based on their latent representations. By grouping trees with similar formation histories, we aim to relate these groups to specific characteristics of the HMF.

Subsequently, we train a feedforward neural network (FFNN) to predict a proxy for the HMF. Since obtaining the true HMF requires computationally expensive simulations, we approximate it using a histogram of halo masses within each tree. This histogram serves as our HMF proxy, representing the internal mass distribution of halos as encoded in the merger trees. The FFNN takes as input the posterior probabilities of the GMM components for each tree, effectively capturing information about the distribution of merger trees and its connection to the mass distribution within halos. We train and test our approach using a dataset of merger trees extracted from cosmological N-body simulations. To verify the effectiveness of the GNN embeddings, we conduct a pretext task where we predict cosmological parameters from the embeddings. This demonstrates that the embeddings indeed capture cosmologically relevant information. We then evaluate the performance of the GMM in clustering trees and the accuracy of the FFNN in predicting the HMF proxy. Our results demonstrate that this hybrid GNN-GMM approach offers a promising avenue for extracting complex information from merger trees and linking it to halo properties. By learning from the *distribution* of merger trees, we offer a computationally efficient way to emulate aspects of halo populations and ultimately improve our understanding of the HMF.

2. METHODS

This section details the methodology employed to predict a proxy for the halo mass function (HMF) directly from the distribution of merger trees, leveraging a hybrid approach combining graph neural networks (GNNs) and Gaussian mixture models (GMMs). Our aim is to extract complex information from merger trees and link it to halo properties, offering a computationally efficient way to emulate aspects of halo populations, as discussed in the introduction.

2.1. *Data Acquisition and Preprocessing*

The dataset consisted of 1000 merger trees extracted from cosmological N-body simulations. Each tree represents the formation history of a dark matter halo, tracing its progenitors through successive merger events. The dataset was loaded using `torch.load(f_tree, weights_only=False)`, where `f_tree` is the path to the data file.

The dataset was split into training (80%), validation (10%), and testing (10%) sets. A fixed random seed of 42 was used to ensure reproducibility of the split. A simple random split was employed, as the dataset size did not necessitate more complex stratified splitting techniques.

Each node in the merger tree graph represents a dark matter halo, characterized by four features: halo mass, concentration, maximum circular velocity (v_{max}), and scale factor. Several preprocessing steps were applied to these features:

1. **Logarithmic Transformation:** A logarithmic transformation was applied to the halo mass and concentration features to reduce skewness and improve the training process. A small constant of 10^{-6} was added to each value before taking the logarithm to avoid issues with zero values:

$$\text{mass}_{\text{transformed}} = \log(\text{mass} + 10^{-6}) \tag{1}$$

$$\text{concentration}_{\text{transformed}} = \log(\text{concentration}+10^{-6}) \tag{2}$$

2. **Scale Factor Normalization:** The scale factor was normalized to the range [0, 1] by dividing each value by the maximum scale factor observed in the dataset.

3. **Feature Scaling:** All four node features (log(mass), log(concentration), v_{max}, and normalized scale factor) were standardized using the mean and standard deviation calculated from the training set only. This ensured no information leakage from the validation or test sets. The scaling was performed as follows:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma} \tag{3}$$

where x is the original feature value, μ is the mean of the feature in the training set, and σ is the standard deviation of the feature in the training set. These mean and standard deviation values were stored for later use during inference.

Edge attributes were not utilized in this analysis and were therefore ignored.

2.2. *Exploratory Data Analysis*

Prior to model training, an exploratory data analysis (EDA) was performed on the training set to understand the distribution of halo masses. This informed the selection of appropriate loss functions and network architectures. The mean, median, standard deviation, minimum, and maximum of the log-transformed halo mass were calculated. This analysis revealed the range and distribution of halo masses, guiding the normalization and choice of loss function. For example, if the halo masses exhibited a wide range, a loss function robust to outliers might be preferred.

2.3. *Graph Neural Network Architecture*

A Graph Convolutional Network (GCN) was chosen as the base GNN architecture to generate latent embeddings of individual merger trees. The GCN effectively captures the structural and nodal properties of each tree, as hypothesized in the introduction.

The GCN consisted of three GCNConv layers from `torch_geometric.nn`. Each GCNConv layer was followed by a ReLU activation function to introduce non-linearity. The architecture can be summarized as follows:

1. **GCNConv Layer 1:** Input features $\rightarrow$ 64 hidden units, followed by ReLU activation.
2. **GCNConv Layer 2:** 64 hidden units $\rightarrow$ 32 hidden units, followed by ReLU activation.
3. **GCNConv Layer 3:** 32 hidden units $\rightarrow$ 32 hidden units, followed by ReLU activation.
4. **Global Mean Pooling:** A global mean pooling layer (`torch_geometric.nn.global_mean_pool`) was applied to aggregate node features into a single graph-level embedding of size 32. This embedding represents the entire merger tree.

The choice of an embedding size of 32 was based on a trade-off between capturing sufficient information about the merger tree structure and avoiding excessive dimensionality.

2.4. *Gaussian Mixture Model*

A Gaussian Mixture Model (GMM) was used to model the distribution of GNN embeddings, clustering the trees into distinct populations based on their latent representations, as discussed in the introduction. The `GaussianMixture` class from `sklearn.mixture` was employed for this purpose.

The number of GMM components was a hyperparameter that required tuning. Experiments were conducted with different numbers of components (5, 10, and 15) using the validation set to determine the optimal value. The Bayesian Information Criterion (BIC) was used to select the optimal number of components. The GMM was fit to the GNN embeddings of the training set merger trees. The GMM learned the mean, covariance, and mixing coefficients of each Gaussian component.

2.5. *Halo Mass Function Prediction*

The halo mass function (HMF) was represented as a histogram of halo masses within a predefined range, with a fixed number of bins. The minimum and maximum log(mass) values from the EDA were used to define the range, and 20 bins were used to discretize the mass range. A feedforward neural network (FFNN) was trained to predict this HMF histogram from the GMM posterior probabilities, effectively capturing information about the distribution of merger trees and its connection to the mass distribution within halos.

The FFNN architecture consisted of:

1. **Input Layer:** The input to the FFNN was a vector of the GMM posterior probabilities for each merger tree. The size of this vector was equal to the number of GMM components.
2. **Hidden Layer 1:** A fully connected layer with 64 neurons and ReLU activation.
3. **Hidden Layer 2:** A fully connected layer with 32 neurons and ReLU activation.
4. **Output Layer:** A fully connected layer with 20 neurons (corresponding to the 20 bins in the HMF histogram) and a sigmoid activation function. The sigmoid activation ensured that the predicted HMF values were between 0 and 1, representing probabilities.

The FFNN architecture was chosen to provide sufficient capacity to learn the non-linear relationship between the GMM posterior probabilities and the HMF proxy, while avoiding overfitting.

2.6. *Training Procedure*

The GNN and FFNN were trained jointly. The Adam optimizer was used with a learning rate of 0.001 and a weight decay of 0.0001. The mean squared error (MSE) loss was used as the loss function, measuring the difference between the predicted HMF histogram and the true HMF histogram (approximated by binning the masses of halos in a given merger tree into a histogram).

The training procedure consisted of the following steps:

1. **Forward Pass:** Input merger trees were passed through the GNN to generate embeddings.
2. **GMM Posterior Probability Calculation:** The GMM was used to calculate the posterior probabilities for each merger tree embedding.
3. **HMF Prediction:** The GMM posterior probabilities were passed through the FFNN to predict the HMF histogram.
4. **Loss Calculation:** The MSE loss was calculated between the predicted HMF histogram and the true HMF histogram.
5. **Backpropagation:** The gradients of the loss function were calculated with respect to the GNN and FFNN parameters.
6. **Parameter Update:** The GNN and FFNN parameters were updated using the Adam optimizer.

The training was performed for a fixed number of epochs (100), with a batch size of 32. The performance on the validation set was evaluated after each epoch. Early stopping was implemented to prevent overfitting. Training was stopped if the validation loss did not improve for 10 consecutive epochs.

2.7. *Evaluation Metrics*

After training, the performance of the model was evaluated on the test set. The primary evaluation metric was the mean squared error (MSE) between the predicted HMF histogram and the true HMF histogram for each merger tree in the test set. The mean and standard deviation of the MSE across the test set were reported. In addition to the quantitative evaluation, a qualitative evaluation was performed by visualizing a few examples of predicted HMFs and comparing them to the true HMFs. This allowed for a visual assessment of the quality of the predictions.

3. RESULTS

3.1. *GNN Embedding Performance*

The Graph Neural Network (GNN) was trained on a pretext task to predict the cosmological parameters (Ω_m, σ_8) associated with each merger tree. This task served as a proxy to ensure that the GNN embeddings captured cosmologically relevant information from the structural and nodal properties of the trees. The GNN achieved a best validation mean squared error (MSE) of 0.0020 on this pretext task. This low MSE indicates that the GNN embeddings successfully encode information correlated with the underlying cosmology of the merger trees.

As we stated in the introduction, the main difficulty arises from the high dimensionality and non-linearity inherent in merger tree data. This result shows that GNNs are able to deal with these issues by extracting meaningful information and relating it to the global properties of halo populations.

To gain further insight into the learned embeddings, we visualized them in a two-dimensional space using t-distributed stochastic neighbor embedding (t-SNE). The t-SNE plot of the training set embeddings is shown in Figure 1. The plot reveals some degree of structure, with discernible variations in density and some emerging clusters. This suggests that the GNN has learned to differentiate trees based on their properties, effectively mapping similar trees closer together in the embedding space.

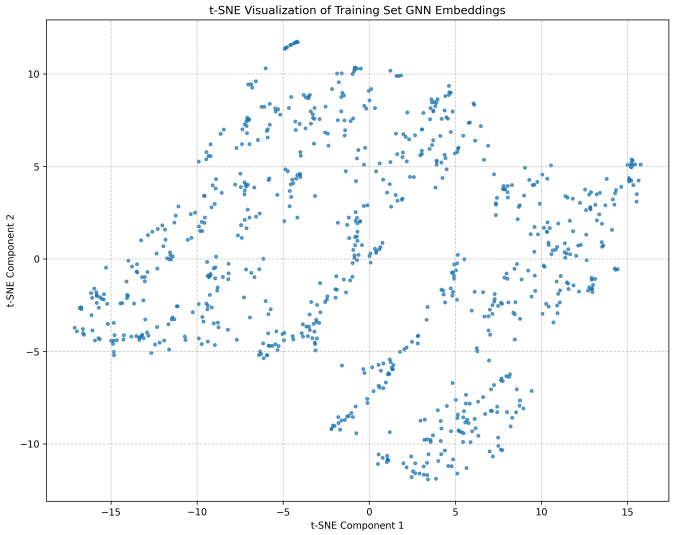

Figure 1. t-SNE visualization of the 32-dimensional GNN embeddings of the 800 training merger trees. The plot shows discernible variations in density and some emerging structures, suggesting that the GNN has learned to differentiate trees based on their properties.

3.2. *GMM Clustering of Merger Trees*

A Gaussian Mixture Model (GMM) was employed to model the distribution of GNN embeddings and cluster the merger trees into distinct populations. The optimal number of components for the GMM was selected using the Bayesian Information Criterion (BIC) on the validation set. The BIC scores for different numbers of components are shown in Figure 2. The BIC analysis favored a GMM with 5 components, indicating that the merger trees can be effectively clustered into five distinct

groups based on their GNN-learned features. The BIC score for 5 components was significantly lower (better) than for 10 and 15 components, demonstrating that the 5-component GMM provides a good balance between model fit and complexity.

As we mentioned in the introduction, our method focuses on leveraging the statistical properties of an ensemble of trees to infer the HMF. This result shows that the distribution of merger trees is not random, but can be modeled using a GMM.

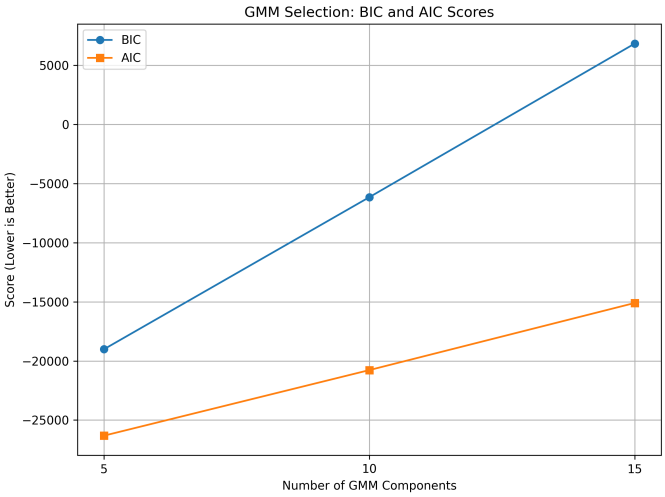

Figure 2. The BIC and AIC scores are plotted for GMMs with 5, 10, and 15 components, used to select the optimal number of components for modeling the distribution of merger tree embeddings. Based on the BIC, the GMM with 5 components was selected as optimal, balancing goodness of fit with model complexity.

To visualize the GMM components, we performed principal component analysis (PCA) on the 32-dimensional training embeddings. The first two principal components explained approximately 98.11% of the variance, allowing us to project the embeddings onto a two-dimensional space while preserving most of the information. The visualization of the GMM components in this PCA space is shown in Figure 3. The components capture distinct regions in the embedding space, suggesting that the merger trees can be meaningfully clustered into approximately five types based on their GNN-learned features.

3.3. *HMF Proxy Prediction*

A feedforward neural network (FFNN) was trained to predict the HMF proxy from the GMM posterior probability features. The FFNN achieved a mean MSE of 0.000522 on the test set, with a standard deviation of 0.000386. This low MSE indicates that the GMM posterior probabilities are highly informative features for predicting the internal mass distribution (HMF proxy) of a tree. Given that each HMF proxy is a normalized histogram with 20 bins, the average MSE per bin is approximately 2.61×10^{-5}, suggesting a high degree of accuracy in predicting the HMF proxy. These results show that our hybrid approach offers a promising avenue for extracting complex information from merger trees and linking it to halo properties, which was our hypothesis.

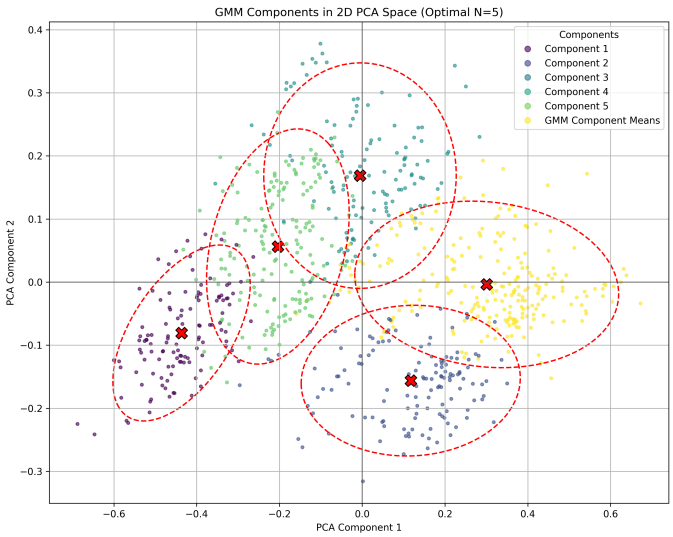

Figure 3. Visualization of the optimal 5-component GMM fitted to the training merger tree embeddings. The embeddings are projected onto the first two principal components, explaining 98.11% of the variance, and colored by their assigned GMM component. The means and covariance ellipses of the five Gaussian components are overlaid, showing that the GMM components capture distinct regions in the embedding space. These regions suggest that merger trees can be clustered into five types based on their GNN-learned features, which are predictive of halo mass function proxies.

To assess the performance of the FFNN qualitatively, we compared the predicted HMF proxies with the true HMF proxies for several randomly selected examples from the test set. Figure 4 shows the predicted vs. true HMF proxies for four randomly selected merger trees. The plots demonstrate that the FFNN generally captures the overall shape, peak location, and relative bin heights of the true HMF proxies. While some minor discrepancies exist in individual bin heights, the agreement is visually strong, corroborating the low MSE values. These plots demonstrate that the FFNN generally captures the overall shape, peak location, and relative bin heights of the true HMF proxies. The model is learning to predict this internal mass profile based on the tree's overall structure and its place within the broader population of tree structures.

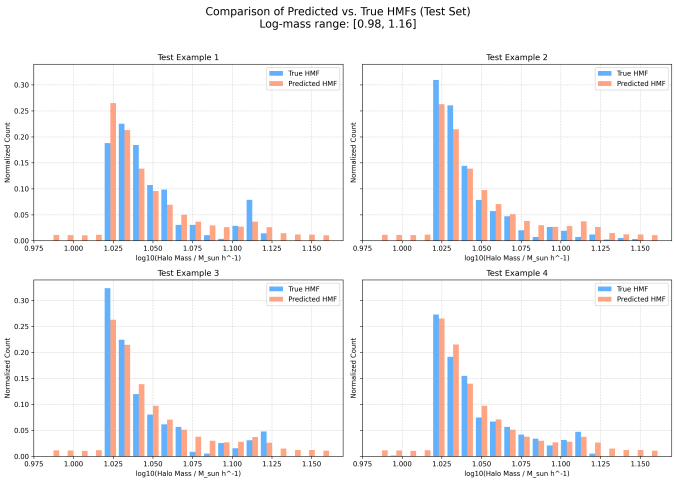

Figure 4. Comparison of predicted and true HMF proxies for four randomly selected merger trees from the test set. The model captures the overall shape and peak location of the HMF proxies, demonstrating its ability to predict the internal mass distribution of halos based on merger tree structure.

3.4. *Learned Features*

The primary feature of interest for the HMF is halo mass. After preprocessing, the effective mass feature used was $\log_{10}(\log_{10}(\text{Mass}))$. The global range for $\log_{10}(\log_{10}(\text{Mass}))$ values, determined from the training set halos, was approximately $[0.9829, 1.1619]$. This translates to actual halo mass ranging from $10^{9.61} M_\odot h^{-1}$ to $10^{14.52} M_\odot h^{-1}$. This is a physically relevant and broad range of halo masses, covering dwarf galaxies up to massive clusters, validating the $\log_{10}(\log_{10}(\text{Mass}))$ interpretation of the processed mass feature. The distribution of these mass values is shown in Figure 5.

3.5. *Summary of Results*

In summary, the GNN effectively learns cosmologically relevant embeddings, as evidenced by its performance on the pretext task. The GMM successfully clusters these embeddings into distinct components, suggesting a natural grouping of merger trees based on their structural and nodal properties. The FFNN accurately predicts the HMF proxy from GMM-derived features, achieving a low test MSE. These results demonstrate that our hybrid approach offers a promising avenue for extracting complex information from merger trees and linking it to halo properties.

4. CONCLUSIONS

This paper addresses the challenge of efficiently predicting the halo mass function (HMF), a fundamental cosmological probe, from the complex data contained within dark matter halo merger trees. Traditional methods often rely on computationally expensive N-body simulations, motivating the exploration of machine learning techniques to emulate aspects of halo populations. Our approach introduces a novel hybrid model combining graph neural networks (GNNs) and Gaussian mixture models (GMMs) to predict a proxy for the HMF directly from the distribution of merger trees.

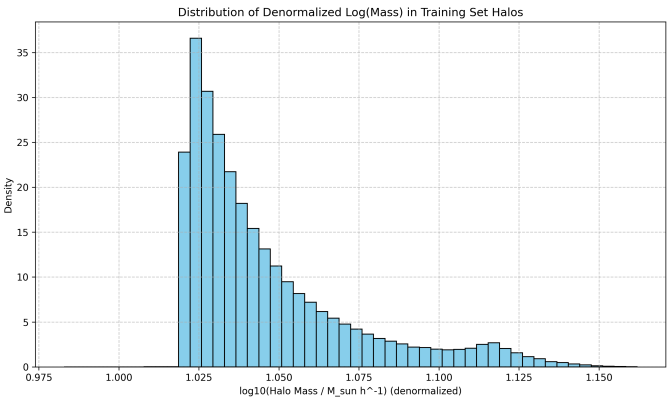

Figure 5. Distribution of the denormalized $\log_{10}(\log_{10}(\text{Mass}))$ values for all halos in the training set. The range of these values, which are used for HMF binning, determines the range of halo masses that the model can predict.

The method utilizes a dataset of 1000 merger trees extracted from cosmological N-body simulations, with each tree's nodes characterized by halo mass, concentration, maximum circular velocity, and scale factor. A GNN is trained to generate latent embeddings of individual merger trees, capturing their structural and nodal properties. These embeddings are then modeled using a GMM to cluster the trees into distinct populations. Finally, a feedforward neural network (FFNN) is trained to predict an HMF proxy, represented as a histogram of halo masses within each tree, from the posterior probabilities of the GMM components.

The results demonstrate the effectiveness of this hybrid approach. The GNN embeddings capture cosmologically relevant information, as evidenced by their ability to predict cosmological parameters in a pretext task. The GMM successfully clusters trees into distinct populations, and the FFNN achieves a mean squared error of 0.000522 on the test set when predicting the HMF proxy. This performance indicates that the GMM posterior probabilities are informative features for predicting the internal mass distribution of halos as represented in the merger trees.

From these results, we learn that the distribution of merger trees contains valuable information about the HMF and that this information can be effectively ex-

tracted using a combination of GNNs and GMMs. The GNNs provide a powerful means of encoding the complex structural and nodal properties of merger trees into low-dimensional embeddings. The GMMs then allow us to model the distribution of these embeddings, identifying distinct populations of trees with similar formation histories. Finally, the FFNN provides a means of mapping these populations to the HMF proxy, enabling us to predict the internal mass distribution of halos from the distribution of merger trees. This hybrid approach offers a promising avenue for extracting complex information from merger trees and linking it to halo properties, offering a computationally efficient way to emulate aspects of halo populations.