Draft version May 29, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

# Contrastive Learning of Merger Tree Embeddings for Likelihood-Free Cosmological Inference

AstroPilot[1]

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

## ABSTRACT

Cosmological inference from dark matter halo merger trees is challenging due to the intricate relationships between tree structure, assembly bias, and underlying cosmological parameters. We address this challenge by developing a contrastive learning framework that generates merger tree embeddings sensitive to cosmological parameters while mitigating the impact of assembly bias. A Graph Neural Network (GNN) is trained on merger trees from N-body simulations, employing a contrastive loss function to cluster trees originating from the same cosmology within the embedding space. To enhance robustness against assembly bias, we augment the training data by introducing variations in halo concentrations conditional on halo mass, guided by observed mass-concentration relations. These learned embeddings then serve as summary statistics for likelihood-free inference (LFI) using Sequential Neural Posterior Estimation (SNPE) to estimate the posterior distribution of $\Omega_m$ and $\sigma_8$. Using a dataset of 1000 merger trees from 40 unique cosmologies, our results demonstrate the effectiveness of the learned embeddings for cosmological inference, particularly for $\Omega_m$, achieving good accuracy and coverage probability close to the nominal value. However, we observe some undercoverage for $\sigma_8$, indicating potential for further refinement of the method. This work underscores the potential of contrastive learning and GNNs for extracting cosmologically relevant information from merger trees, paving the way for robust and accurate likelihood-free cosmological inference.

*Keywords:* Large-scale structure of the universe, Cosmological parameters, Cosmology, Galaxy evolution, N-body simulations

## 1. INTRODUCTION

Cosmological inference, the precise determination of the fundamental parameters that govern the evolution of our Universe, stands as a central goal in modern cosmology. Traditional approaches often rely on comparing theoretical predictions, derived from computationally intensive N-body simulations, with observational data from sources like the Cosmic Microwave Background, galaxy surveys, and Type Ia supernovae. These methods typically require carefully constructed summary statistics to effectively bridge the gap between theoretical models and observational data. Likelihood-free inference (LFI) offers an alternative pathway, circumventing the need for a pre-defined likelihood function by directly learning the mapping between simulation data and the underlying cosmological parameters.

Dark matter halo merger trees, which capture the hierarchical assembly history of cosmic structures, represent a rich source of information for LFI (Parkinson et al. 2007; Ángel Chandro-Gómez et al. 2025). These trees encode valuable insights into the Universe's underlying cosmology (Benson et al. 2012; Ángel Chandro-Gómez et al. 2025). However, extracting this information presents significant challenges. The relationships between the tree structure, the assembly histories of dark matter halos, and the cosmological parameters are complex and non-linear, making it difficult to design effective summary statistics that capture the relevant cosmological information (Jiang & van den Bosch 2013). Furthermore, the phenomenon of assembly bias adds another layer of complexity. Assembly bias refers to the correlation between halo properties and their formation history, independent of their mass and the underlying cosmology (Benson et al. 2012; Jiang & van den Bosch 2013). If not properly accounted for, assembly bias can obscure the cosmological signal within merger trees and lead to biased parameter estimates (Benson et al. 2012; Jiang & van den Bosch 2013).

In this paper, we introduce a novel contrastive learning framework designed to address these challenges. Our framework generates merger tree embeddings that are sensitive to cosmological parameters while simultaneously being robust to the effects of assembly bias (Poul-

ton et al. 2018; Ángel Chandro-Gómez et al. 2025). We achieve this by training a Graph Neural Network (GNN) on a dataset of merger trees extracted from N-body simulations (Bose et al. 2022). The GNN is trained using a contrastive loss function, which encourages the network to cluster trees originating from the same cosmology together within the embedding space, while pushing trees from different cosmologies apart. To explicitly mitigate the impact of assembly bias, we augment the training data by introducing controlled variations in halo concentrations, conditional on halo mass (Poulton et al. 2018; Ángel Chandro-Gómez et al. 2025). This augmentation is guided by observed mass-concentration relations (Poulton et al. 2018) and effectively increases the diversity of the training data, enabling the GNN to learn embeddings that are less sensitive to variations in halo assembly history.

The learned embeddings then serve as effective summary statistics for LFI. We employ Sequential Neural Posterior Estimation (SNPE) to estimate the posterior distribution of key cosmological parameters, specifically the matter density parameter ($\Omega_m$) and the amplitude of the matter power spectrum ($\sigma_8$), given the learned merger tree embeddings. This allows us to directly estimate the probability distribution of these parameters based on the information extracted from the merger trees. We validate the effectiveness of our framework by applying it to a dataset of 1000 merger trees originating from 40 unique cosmologies (Pearson et al. 2024; Nguyen et al. 2024). To ensure the reliability of our results, we assess the accuracy and calibration of our inferred posteriors using simulation-based calibration techniques (Nguyen et al. 2024). Our results demonstrate the potential of contrastive learning and GNNs for extracting cosmologically relevant information from merger trees, paving the way for robust and accurate likelihood-free cosmological inference.

## 2. METHODS

This section details the methodology employed to develop a contrastive learning framework for cosmological inference from dark matter halo merger trees. Our approach leverages Graph Neural Networks (GNNs) to generate merger tree embeddings sensitive to cosmological parameters, while mitigating the impact of assembly bias. These learned embeddings then serve as summary statistics for likelihood-free inference (LFI) using Sequential Neural Posterior Estimation (SNPE) to estimate the posterior distribution of cosmological parameters.

### 2.1. *Dataset and Preprocessing*

The analysis is based on a dataset of 1000 dark matter halo merger trees, extracted from N-body simulations with 40 unique cosmologies (Ángel Chandro-Gómez et al. 2025). Each merger tree is represented as a PyTorch Geometric 'Data' object, consisting of node features, edge connections, and graph-level targets. The node features, denoted as $x$, are a tensor of shape (N_nodes, 4), where N_nodes is the number of nodes in the tree. These features include log10(mass), log10(concentration), log10(Vmax), and the scale factor (Parkinson et al. 2007; Jiang & van den Bosch 2013). The edge connections are represented by the *edge_index*, which defines the connections between nodes in the tree (Parkinson et al. 2007; Jiang & van den Bosch 2013). The graph-level targets, denoted as $y$, are a tensor of shape (1, 2), containing the cosmological parameters $\Omega_m$ and $\sigma_8$ (Benson et al. 2012).

Prior to training, the data undergoes several preprocessing steps: (Zhang et al. 2024; Cao et al. 2025; Nerval et al. 2025) *Exploratory Data Analysis (EDA):* An initial inspection of the dataset is performed to understand the structure and typical ranges of the node features and cosmological parameters (Verde 2009; Moss 2025). This includes printing the attributes of a few sample trees, verifying the dimensions of $x$ and $y$, and confirming the range of scale factors, $\Omega_m$, and $\sigma_8$. Furthermore, a feature distribution analysis is conducted to calculate the global mean and standard deviation of each node feature and cosmological parameter across all trees (Coley et al. 2018). Tables summarizing these statistics are generated to provide a comprehensive overview of the data distribution (Moss 2025). The number of trees belonging to each unique cosmology is also determined.

*Normalization:* To ensure stable training and prevent features with larger scales from dominating the learning process, both the node features and cosmological parameters are normalized. Each node feature is normalized by subtracting its global mean and dividing by its global standard deviation, calculated during the EDA. Similarly, $\Omega_m$ and $\sigma_8$ are normalized using their respective global means and standard deviations.

*Dataset Splitting:* The dataset is split into training, validation, and test sets, ensuring that trees from the same cosmology are grouped together in the same split. This is crucial to avoid information leakage and ensure a fair evaluation of the model's generalization performance (Stölzner et al. 2025). The unique cosmologies are first identified and shuffled. Then, approximately 70% of the unique cosmologies are allocated to the training set, 15% to the validation set, and 15% to the test set. All merger trees corresponding to the cosmologies

in each set are then collected to form the final training, validation, and test sets.

### 2.2. *Feature Engineering for Assembly Bias Proxy and Global Properties*

To capture global properties of the merger trees and to inform the data augmentation strategy, we extract a set of global features for each tree. These features include:

*Total Mass:* The total mass of all halos in the tree at the final snapshot (scale factor closest to 1) and the total mass of progenitor halos within defined scale factor bins (e.g., [0.2-0.4], [0.4-0.6], [0.6-0.8], [0.8-1.0]) (Bansal et al. 2023). *Mass-Weighted Mean Properties:* The mass-weighted mean concentration and Vmax at the final snapshot and within each defined scale factor bin (Burgarella et al. 2022). *Main Progenitor Branch (MPB) Properties:* The mass, concentration, and Vmax of the main progenitor at specific scale factors (e.g., a=0.5, a=0.25). The MPB is defined as the most massive progenitor at each step back in time from the final halo (Pu et al. 2025). The MPB is defined as the most massive progenitor at each step back in time from the final halo (Pu et al. 2025). FLORAH enables accurate modeling of MPB properties across a wide dynamic mass range, including mass evolution and concentration (Nguyen et al. 2024). *Assembly History Proxies:* The formation time, defined as the scale factor at which the main progenitor first reached half of its final mass (Shojaei et al. 2025), and the number of major mergers (mass ratio $>$ 1:3 or 1:4) experienced by the MPB (Conselice et al. 2022; Shojaei et al. 2025). *Assembly Bias Proxies:* The scatter in concentration at fixed mass is calculated by binning halos in the final snapshot by their mass and calculating the standard deviation of log10(concentration) within each mass bin (Smith et al. 2024). The mean concentration of satellite halos versus the central halo is also calculated by identifying the most massive halo at z=0 as the central halo and considering all other leaf nodes at the final snapshot as "satellites" (Lacerna et al. 2014). *Tree Structure Metrics:* The total number of nodes in the tree, the maximum depth of the tree (Cavelan et al. 2020; Yang & Yu 2023), and the average branching factor are also extracted (Cavelan et al. 2020).

These engineered global features are stored alongside the original data for each tree, facilitating their use in downstream analysis and data augmentation.

### 2.3. *Assembly Bias Correction via Data Augmentation*

To mitigate the impact of assembly bias (Paranjape & Padmanabhan 2017; Smith et al. 2024), we augment the training dataset by creating variations in halo concentrations, conditional on halo mass (Paranjape & Padmanabhan 2017; Smith et al. 2024). This augmentation procedure involves the following steps:

*Model Mass-Concentration Relation (M-C relation):* Using all halos from the training set trees, a median M-C relation is fitted (e.g., log10(concentration) = A * log10(mass) + B) (Biviano et al. 2017; Gilman et al. 2019; Gu et al. 2022). For each mass bin, the observed scatter (standard deviation) in log10(concentration) around this median relation is calculated.

*Augmentation Procedure:* For each tree in the training set, K augmented copies are created (e.g., K=2-3). For each node (halo) in a copy, its mass and scale factor are identified. Based on its mass and the M-C relation (and scatter) derived for its scale factor bin, a new log10(concentration) value is sampled from a Gaussian distribution centered at the predicted median concentration for its mass, with a standard deviation equal to the observed scatter (Zhang et al. 2025). The new concentration value is ensured to remain within a physically plausible range. The $x$ tensor of the augmented tree has its second column (log10(concentration)) modified, while other features (mass, Vmax, scale factor) and the tree structure (*edge_index*) remain unchanged. These augmented trees have the same cosmological parameters ($y$) as their original tree (de Santi et al. 2022).

### 2.4. *Contrastive Embedding Learning*

A Graph Neural Network (GNN) is trained to learn embeddings of merger trees using a contrastive loss function. The GNN architecture consists of a series of Graph Convolutional Network (GCN) layers or Graph Attention Network (GAT) layers, followed by a global mean pooling layer and a Multilayer Perceptron (MLP) (Tang & Ting 2022; Chuang et al. 2023). The input to the GNN is the $x$ (node features) and *edge_index* of each tree, and the output is a fixed-size embedding vector (e.g., 64 or 128 dimensions) for each tree (Tang & Ting 2022).

The contrastive loss function used is NT-Xent (Normalized Temperature-scaled Cross Entropy Loss) (Gondhalekar et al. 2024; Wilkinson et al. 2025; Perez et al. 2025). In each batch, for a given tree (anchor), other trees from the same cosmology (same original $y$ value) are positive pairs, including its own augmentations and other original trees from the same cosmology (Gondhalekar et al. 2024; Perez et al. 2025). Trees from different cosmologies are negative pairs (Gondhalekar et al. 2024; Perez et al. 2025). The loss encourages embeddings of positive pairs to be closer in the embedding space and embeddings of negative pairs to be further apart (Gondhalekar et al. 2024; Wilkinson et al. 2025; Perez et al. 2025).

The training process involves using the (original + augmented) training set. For each training step, a batch of merger trees is sampled, ensuring the batch construction allows for forming positive and negative pairs (Jespersen et al. 2022; Tang & Ting 2022). Each tree is passed through the GNN to get its embedding (Wu et al. 2024,?), and the contrastive loss is calculated (Tang & Ting 2022). Backpropagation is performed to update the GNN weights using the Adam optimizer (Wu et al. 2024,?). The loss is monitored on the validation set, and the model with the best validation loss is saved (Wu et al. 2024).

### 2.5. *Likelihood-Free Inference (LFI) with Learned Embeddings*

The trained GNN is used to generate embeddings, which then serve as summary statistics for an LFI algorithm (Lehman et al. 2024). All trees in the training, validation, and test sets are processed through the trained GNN to obtain their respective embedding vectors.

The LFI framework used is Sequential Neural Posterior Estimation (SNPE) from the *sbi* Python package (Zhang et al. 2023; Erickson et al. 2024). SNPE trains a neural density estimator (e.g., a normalizing flow) to approximate the posterior *p(cosmology | embedding)* (Zhang et al. 2023; Kosiba et al. 2024). The input to SNPE is the learned embeddings of the training set trees, and the target is the corresponding cosmological parameters ($y$) for these trees (Wagner-Carena et al. 2024; Kosiba et al. 2024).

For each tree in the test set, its embedding is obtained using the GNN (Roncoli et al. 2024; Chatterjee & Villaescusa-Navarro 2025), and the trained SNPE model (Lehman et al. 2024) is used to infer the posterior distribution of $\Omega_m$ and $\sigma_8$ given this embedding.

### 2.6. *Validation and Calibration*

To ensure the accuracy and calibration of the LFI pipeline, several validation techniques are employed (Frailis et al. 2010; Konar et al. 2024).

A qualitative assessment is performed by plotting the 1D and 2D marginalized posterior distributions against the true cosmological parameters for a few test cases (Jia 2024). Simulation-Based Calibration (SBC) is performed on the test set to assess the calibration of the posteriors (Mao et al. 2024). For each test tree, the true cosmological parameters are obtained, and M samples are drawn from the posterior distribution estimated by SNPE. For each parameter, its rank among the M posterior samples is calculated. Histograms of these ranks are plotted, and deviations from uniformity indicate miscalibration (Mao et al. 2024). Metrics such as the mean squared error (MSE) between the posterior mean and the true parameters, and the coverage properties of the credible intervals are also calculated to assess the accuracy and precision of the inferred posteriors (Mao et al. 2024; Jia 2024).

## 3. RESULTS

### 3.1. *Dataset Characteristics and Preprocessing*

The analysis begins with a dataset of 1000 merger trees, each representing the hierarchical assembly history of dark matter halos extracted from N-body simulations. These trees originate from 40 unique cosmological parameter pairs $(\Omega_m, \sigma_8)$, with 25 trees per cosmology, providing a statistically significant sample for training and evaluating the contrastive learning framework. Each node within a merger tree is characterized by four features: $\log_{10}$(Mass), $\log_{10}$(Concentration), $\log_{10}(V_{\rm max})$, and the scale factor ($a$), encapsulating the mass, density, velocity dispersion, and cosmic epoch of the corresponding dark matter halo.

*Initial Data Statistics:* Prior to normalization, the node features and target cosmological parameters exhibit diverse ranges and scales. The raw data spans several orders of magnitude, potentially leading to numerical instability during neural network training. For instance, $\log_{10}$(Mass) ranges from approximately 9.61 to 14.68, reflecting the wide spectrum of halo masses present in the simulation. Similarly, the cosmological parameters $\Omega_m$ and $\sigma_8$ vary from 0.103 to 0.4734 and 0.603 to 0.9918, respectively, representing the range of cosmological models explored in this study.

*Normalization and Dataset Splitting:* To mitigate the effects of disparate scales and improve training stability, all node features and target parameters are normalized to have a mean of approximately 0 and a standard deviation of approximately 1 across the entire dataset. This transformation ensures that each feature contributes equally to the learning process, preventing features with larger numerical values from dominating the loss function. Figure 1 and Figure 2 show the distribution of the normalized $\log_{10}$(Mass) and scale factor, respectively. Similarly, Figure 3 shows the distribution of the normalized $\sigma_8$ parameter. This normalization is crucial for stable neural network training.

The dataset is subsequently split into training, validation, and test sets based on unique cosmologies to prevent data leakage. This approach ensures that the model is evaluated on cosmologies it has not seen during training, providing a more realistic assessment of its generalization performance. The resulting split comprises 700 trees (28 cosmologies) for training, 150 trees

(6 cosmologies) for validation, and 150 trees (6 cosmologies) for testing.

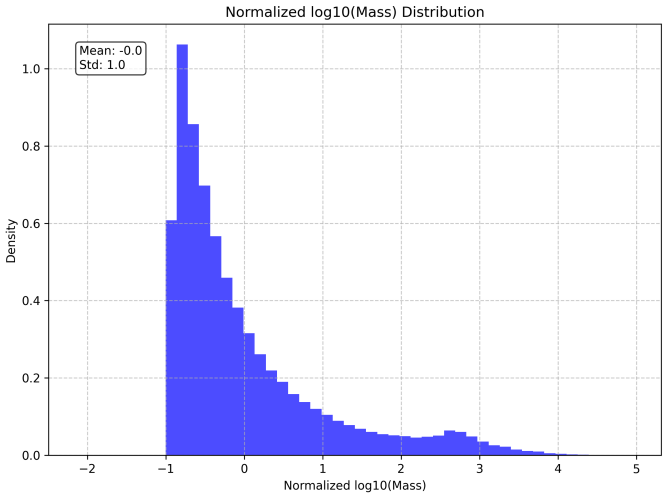


**Figure 1.** Distribution of normalized $\log_{10}$(Mass) values across all nodes in the dataset, demonstrating the successful transformation to approximately zero mean and unit variance, which is crucial for stable neural network training and the contrastive learning framework.

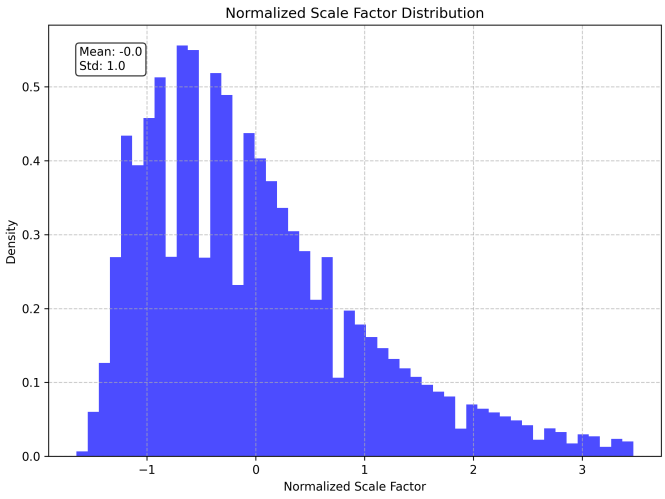


**Figure 2.** Distribution of the normalized scale factor, a node feature used in the GNN, showing a mean of approximately 0 and a standard deviation of approximately 1. This normalization is crucial for stable neural network training, ensuring all features contribute equally during the contrastive learning process for inferring cosmological parameters.

### 3.2. *Engineered Global Merger Tree Features*

A comprehensive suite of 35 global features is engineered for each merger tree to explore their potential predictive power for cosmological parameters. These features encompass a wide range of properties, including total mass properties, mass-weighted mean halo properties at different epochs, main progenitor branch (MPB) characteristics, assembly history proxies, assembly bias proxies, and tree structural metrics. The motivation behind this feature engineering is to identify key aspects of merger tree structure and evolution that are most sensitive to the underlying cosmology, potentially guiding the design of more interpretable analytic models.

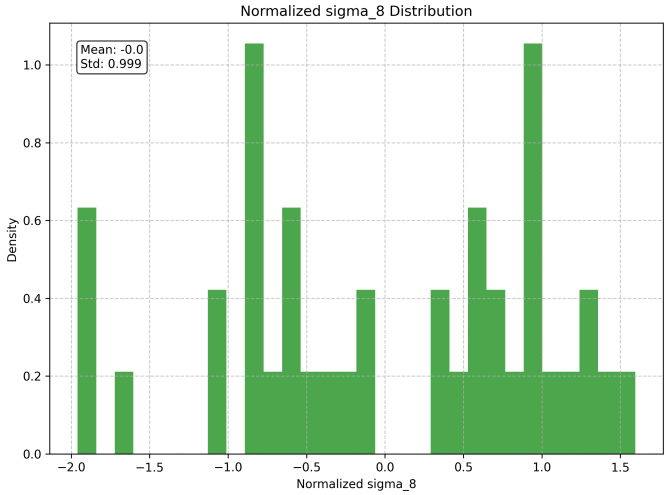


**Figure 3.** Distribution of the normalized $\sigma_8$ parameter, showing the result of the normalization to approximately zero mean and unit variance, which is crucial for stable neural network training.

*Correlation with Cosmological Parameters:* A Pearson correlation analysis is performed between the unnormalized engineered features and the unnormalized cosmological parameters $(\Omega_m, \sigma_8)$ to quantify the strength and direction of their linear relationships. The results reveal several features that exhibit notable correlations, suggesting their potential utility as summary statistics for cosmological inference. The heatmap in Figure 4 visualizes these correlations.

#### 3.2.1. *Correlations with* $\Omega_m$

The strongest positive correlation is observed with 'mpb_log_conc_at_sf_0.5' (Main Progenitor Branch $\log_{10}$(Concentration) at scale factor $a = 0.5$), showing a coefficient of 0.92. This strong correlation suggests that halos in higher $\Omega_m$ universes tend to have significantly higher concentrations at $a = 0.5$. This finding aligns with theoretical expectations, as higher matter density $(\Omega_m)$ generally leads to earlier structure formation and more concentrated halos for a given mass. Other strong positive correlations include 'mw_mean_log_conc_sf_bin_0' (mass-weighted mean

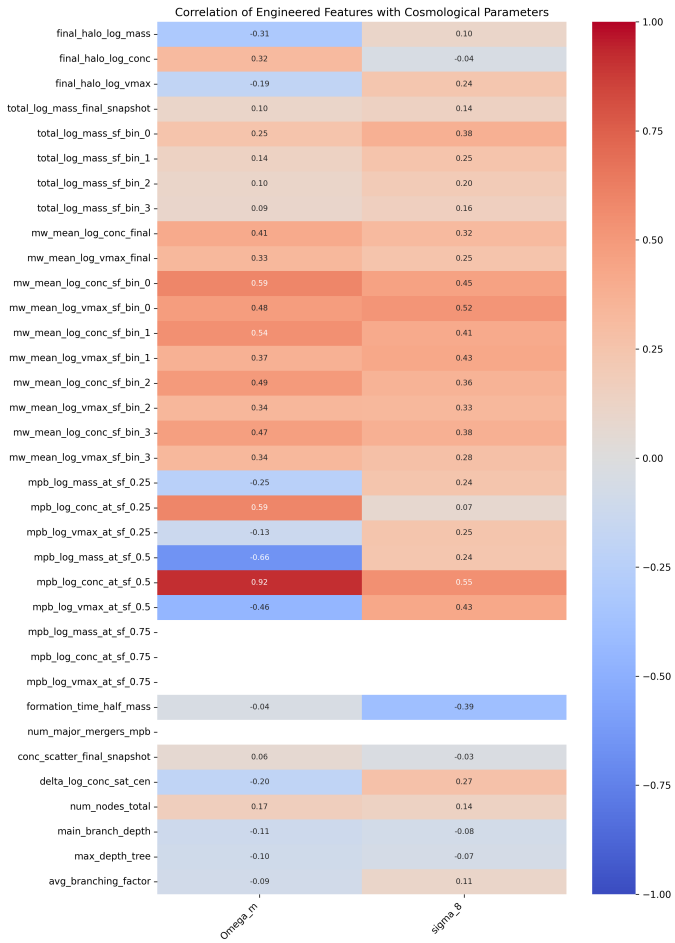


**Figure 4.** Heatmap of Pearson correlation coefficients between engineered global merger tree features and cosmological parameters ($\Omega_m$, $\sigma_8$). The Main Progenitor Branch $\log_{10}$(Concentration) at scale factor $a = 0.5$ exhibits the strongest positive correlation with $\Omega_m$, suggesting that halo concentration at early times is a crucial indicator of the underlying cosmology, which may be useful for developing analytic models to link merger tree features to cosmology.

$\log_{10}$(Concentration) in scale factor bin [0.2-0.4]) with $r = 0.59$, and 'mpb_log_conc_at_sf_0.25' with $r = 0.59$, further emphasizing the importance of early-time halo concentrations for inferring $\Omega_m$. A strong negative correlation is found with 'mpb_log_mass_at_sf_0.5' (MPB $\log_{10}$(Mass) at $a = 0.5$) with $r = -0.66$. The distributions of 'mpb_log_conc_at_sf_0.25' and 'mw_mean_log_conc_sf_bin_1' are shown in Figure 5 and Figure 6, respectively.

### 3.2.2. *Correlations with $\sigma_8$*

The correlations with $\sigma_8$ are generally weaker compared to those with $\Omega_m$. The strongest positive correlation is with 'mpb_log_conc_at_sf_0.5' ($r = 0.55$), followed by 'mw_mean_log_vmax_sf_bin_0' (mass-

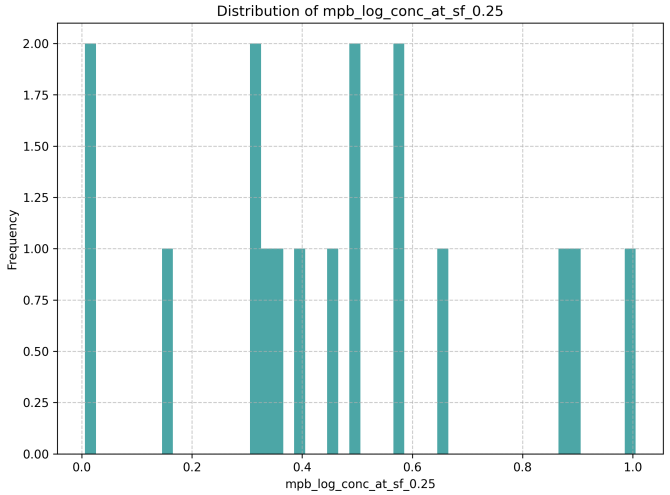


**Figure 5.** Distribution of the engineered feature 'mpb_log_conc_at_sf_0.25' (Main Progenitor Branch $\log_{10}$(Concentration) at scale factor $a = 0.25$) across the dataset. This feature shows a positive correlation with $\Omega_m$, indicating that halos in higher $\Omega_m$ universes tend to have higher concentrations at $a = 0.25$.

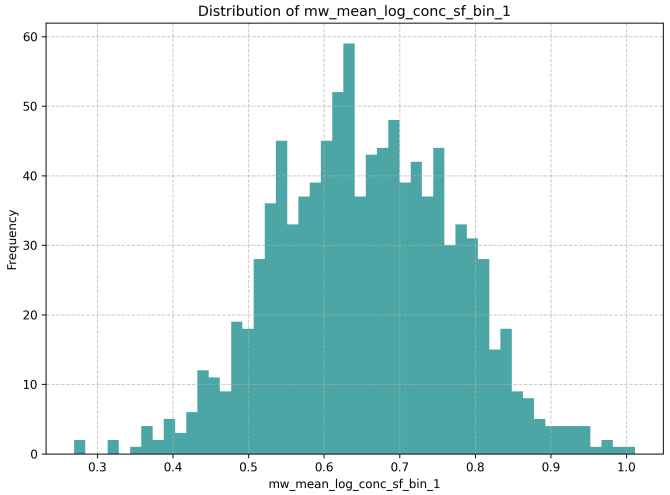


**Figure 6.** Distribution of the engineered feature 'mw_mean_log_conc_sf_bin_1', which represents the mass-weighted mean $\log_{10}$(Concentration) in scale factor bin [0.4-0.6]. This feature showed a positive correlation with $\Omega_m$, suggesting that higher matter density universes tend to have higher concentration values in this scale factor bin.

weighted mean $\log_{10}(V_{\rm max})$ in scale factor bin [0.2-0.4]) with $r = 0.52$. 'total_log_mass_sf_bin_0' (total $\log_{10}$(Mass) in scale factor bin [0.2-0.4]) also shows a moderate positive correlation ($r = 0.38$). 'formation_time_half_mass' exhibits a moderate negative correlation ($r = -0.39$), implying that higher $\sigma_8$ values might correlate with earlier formation times

(smaller scale factor at half-mass). Since $\sigma_8$ dictates the amplitude of mass fluctuations, these correlations suggest that higher $\sigma_8$ values lead to earlier structure formation and a higher abundance of massive halos at early times. Distributions for several features are shown in the following figures: Figure 7, Figure 8, Figure 9, and Figure 10.

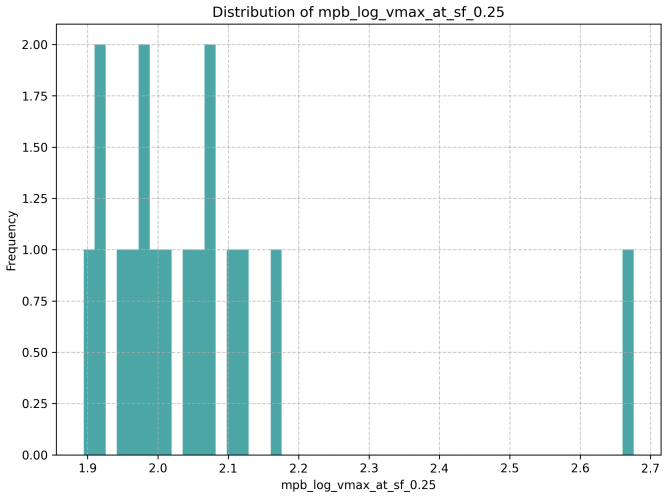


**Figure 7.** Distribution of the engineered feature $\log_{10}(V_{\rm max})$ of the main progenitor branch at scale factor $a = 0.25$. This feature exhibited a moderate correlation with the cosmological parameter $\sigma_8$, suggesting its utility in an analytic model linking merger tree features to cosmology.

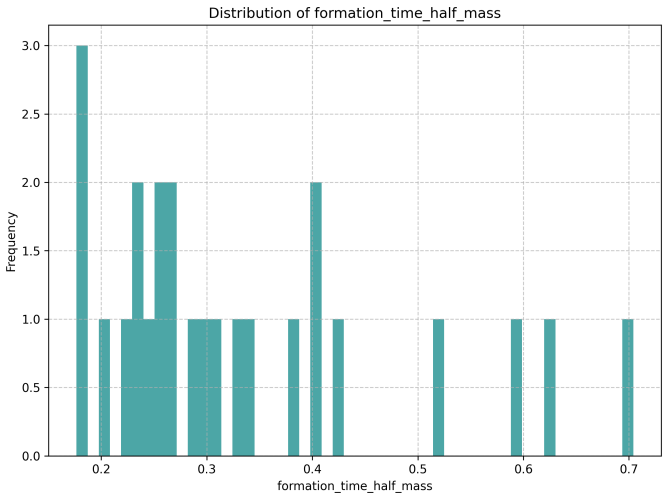


**Figure 8.** Distribution of the engineered feature 'formation_time_half_mass' across the dataset, which exhibited a moderate negative correlation with $\sigma_8$, suggesting that higher $\sigma_8$ values might correlate with earlier formation times.

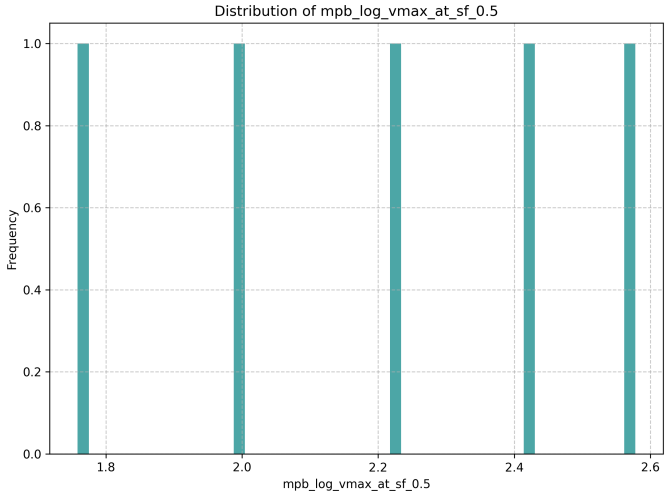


**Figure 9.** Distribution of the engineered global feature 'mpb_log_vmax_at_sf_0.5', which represents the $\log_{10}(V_{\rm max})$ of the main progenitor branch at a scale factor of 0.5. Features like this show correlations with cosmological parameters, suggesting their utility in analytic models.

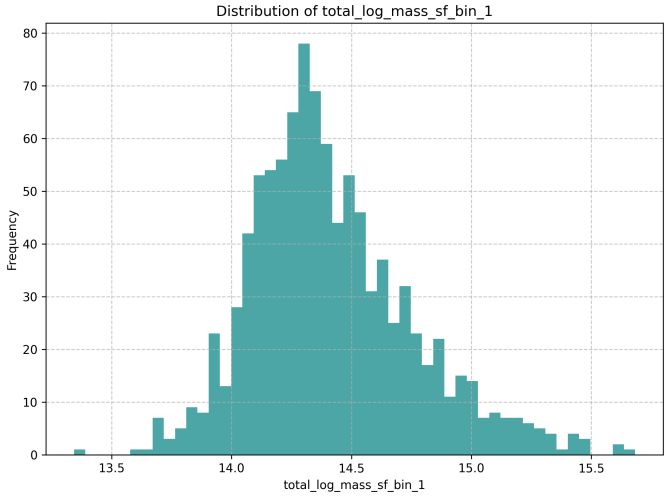


**Figure 10.** Distribution of the engineered feature total halo mass in the scale factor bin [0.4, 0.6], showing the range of values present in the dataset and their frequency. The correlation of this feature with cosmological parameters suggests its potential utility in an analytic model.

The distributions of 'num_major_mergers_mpb' and 'max_depth_tree' and 'final_halo_log_mass' are shown in Figure 11, Figure 12, and Figure 13, respectively.

### 3.2.3. *Assembly Bias Proxies*

Features designed to proxy assembly bias, such as 'conc_scatter_final_snapshot' (scatter in $\log_{10}$(Concentration) at fixed mass in the final snap-

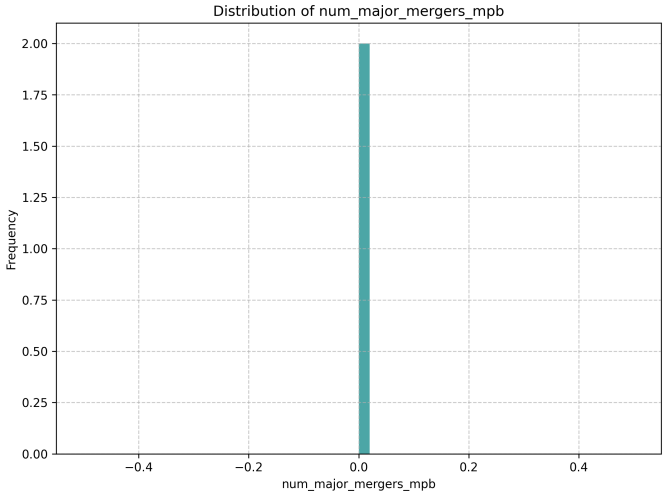


**Figure 11.** Distribution of the engineered feature 'num_major_mergers_mpb' (number of major mergers in the main progenitor branch) across the dataset, which is used to explore potential predictive power for cosmological parameters.

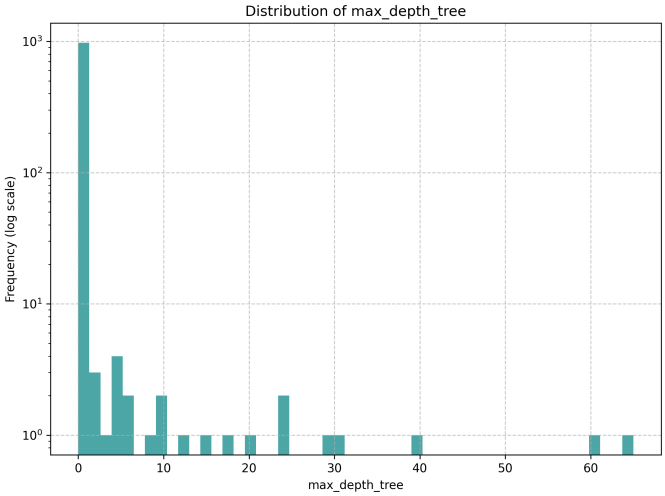


**Figure 12.** Distribution of the maximum depth of merger trees, a global feature engineered to capture tree structure, showing the frequency of different tree depths on a logarithmic scale. The GNN leverages such structural information to infer cosmological parameters.

shot) and 'delta_log_conc_sat_cen' (difference in mean $\log_{10}$(Concentration) between satellite and central halos), show weak correlations with cosmological parameters. 'conc_scatter_final_snapshot' has $r \approx 0.06$ with $\Omega_m$ and $r \approx -0.03$ with $\sigma_8$. 'delta_log_conc_sat_cen' shows $r \approx -0.20$ with $\Omega_m$ and $r \approx 0.27$ with $\sigma_8$. This suggests that these specific global proxies, on their own, might not be strong direct indicators of cosmology, or their relationship is more complex and requires more

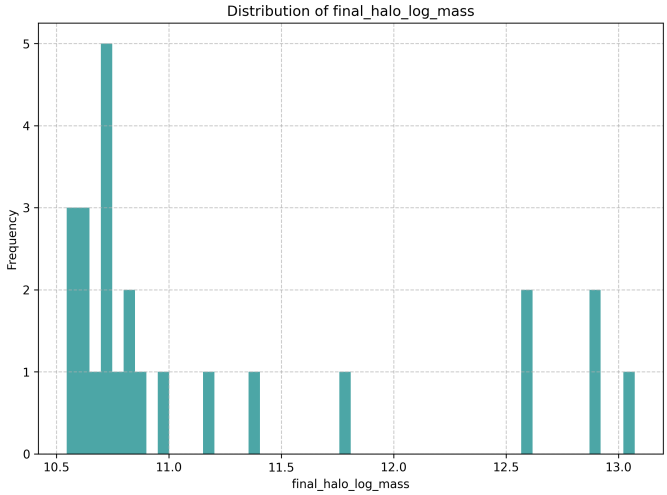


**Figure 13.** Distribution of the engineered feature final halo $\log_{10}$(Mass) across the dataset. This feature shows moderate correlations with cosmological parameters, suggesting its utility in analytical models and informing the GNN during contrastive learning.

sophisticated modeling. The distributions of some of these features, along with other global features, are shown in Figure 14, Figure 15, Figure 16, Figure 17, and Figure 18.

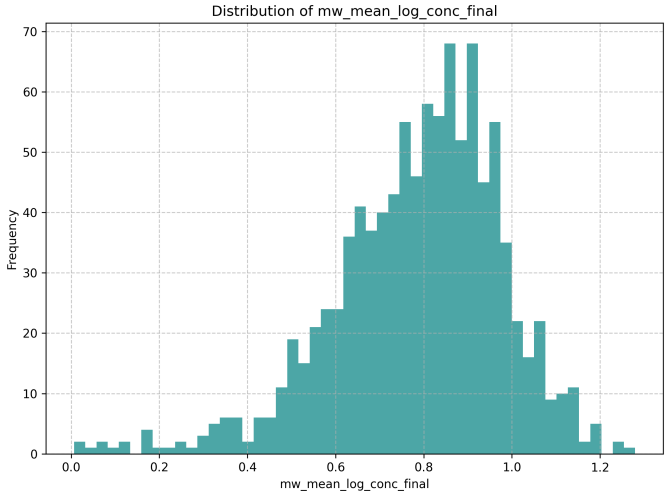


**Figure 14.** Distribution of the mass-weighted mean of log concentration at the final snapshot, a global feature engineered from merger trees, showing a roughly normal distribution. This feature exhibits a correlation with cosmological parameters, suggesting its utility in inferring $\Omega_m$ and $\sigma_8$.

*Implications for Analytic Models:* The strong correlations, particularly of early-time concentration and mass properties of the MPB with $\Omega_m$, align with theoretical expectations and provide valuable guidance for develop-

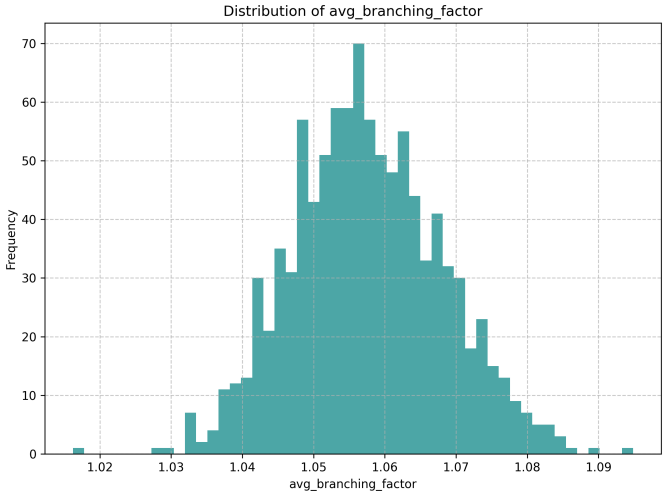


**Figure 15.** Distribution of the average branching factor across the merger trees, a global feature engineered to capture tree structure, highlighting the range of values and central tendency observed in the dataset.

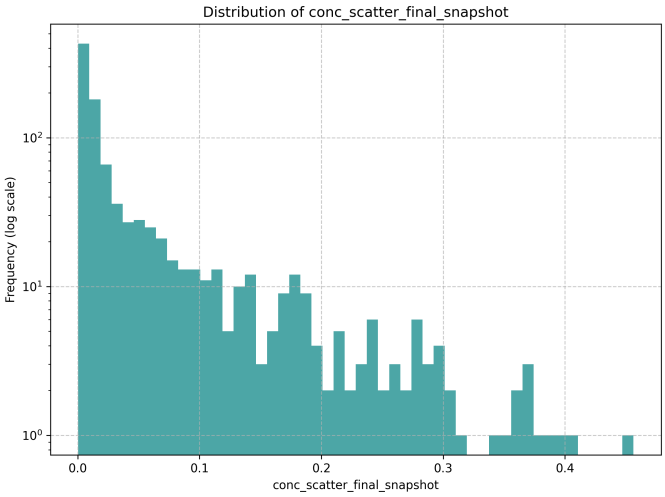


**Figure 16.** Distribution of the engineered feature 'conc_scatter_final_snapshot', which represents the scatter in $\log_{10}$(Concentration) at fixed mass in the final snapshot. This feature, designed as a proxy for assembly bias, exhibited a weak correlation with cosmological parameters, suggesting that more complex relationships might be needed to capture the link between assembly bias and cosmology.

ing analytic models of halo formation. Higher matter density ($\Omega_m$) generally leads to earlier structure formation and more concentrated halos for a given mass, especially at earlier epochs before late-time accretion dominates. Similarly, $\sigma_8$, which dictates the amplitude of mass fluctuations, influences the abundance and properties of early-forming halos. An analytic model attempting to link merger tree features to cosmology would likely

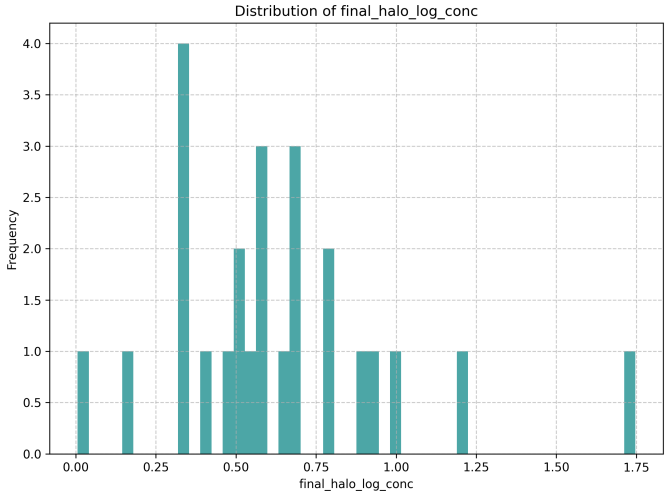


**Figure 17.** Distribution of the global feature $\log_{10}$(Concentration) of final halos in the original dataset, which is correlated with cosmological parameters.

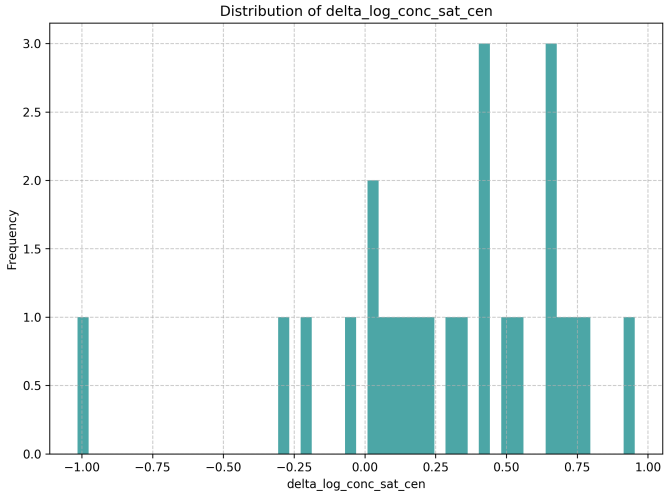


**Figure 18.** Distribution of the engineered feature $\delta \log_{10}$(Concentration) between satellite and central halos, an assembly bias proxy, across the dataset. This feature showed a weak correlation with cosmological parameters, suggesting limited predictive power on its own.

benefit from incorporating terms related to the mass accretion history and concentration evolution of the main progenitor, especially at early to intermediate cosmic epochs (e.g., $a \approx 0.25 - 0.5$). The observed correlations for $V_{\rm max}$ also suggest its utility, as it is intrinsically linked to both mass and concentration.

### 3.3. *Data Augmentation for Assembly Bias Correction*

To enhance the robustness of the learned embeddings to assembly bias—the phenomenon where halo properties depend on more than just mass—a data augmenta-

tion strategy is implemented. This strategy focuses on modifying halo concentrations, a key property that can be influenced by assembly history.

*Mass-Concentration (M-C) Relation:* The M-C relation is first characterized using all halos from the original training set. A linear fit of $\log_{10}$(Concentration) versus $\log_{10}$(Mass) yields the relation: $\log_{10}(C) = -0.0304 \times \log_{10}(M) + 1.071$. The scatter ($\sigma_{\log C}$) around this median relation is found to be 0.3603. The R-squared value for this fit is very low (0.0036), indicating that a simple linear M-C relation does not capture the majority of the variance in concentration across all halo masses and epochs in the dataset. This low R-squared highlights the complexity of halo concentrations and underscores the potential importance of assembly bias and other factors beyond mass in determining halo concentration.

*Augmentation Procedure:* For each tree in the training set, one augmented copy is created. In these copies, the $\log_{10}$(Concentration) value of each halo node is resampled from a Gaussian distribution centered on the value predicted by the fitted M-C relation for its mass, with a standard deviation equal to the observed scatter ($\sigma_{\log C} = 0.3603$). Resampled concentrations are clipped to physically plausible bounds observed in the original dataset (0.0002 to 3.767). This ensures that the augmented concentrations remain within a realistic range. A comparison of the M-C relation for original and augmented halos is shown in Figure 19.

*Quality Check of Augmentation:* Figure 19 shows that the augmented halos exhibit a similar overall trend to the original halos with respect to the fitted M-C line, but with a visibly increased spread. This augmentation step aims to expose the contrastive learning algorithm to a wider range of concentration variations for fixed mass and cosmology, thereby encouraging the GNN to learn embeddings that are more robust to such variations when inferring cosmological parameters. By explicitly varying halo concentrations, the GNN is forced to learn features that are less sensitive to assembly bias and more directly related to the underlying cosmology.

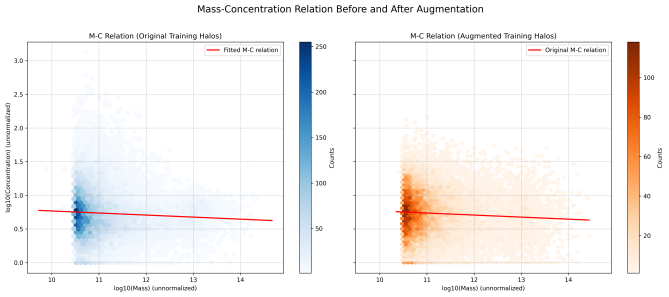


**Figure 19.** Comparison of the mass-concentration relation for original (left) and augmented (right) training halos. The augmentation procedure resamples halo concentrations based on a fitted mass-concentration relation, increasing the scatter in concentration at a given mass, which is visible as a broader distribution of points compared to the original data. This augmentation aims to improve robustness to assembly bias when inferring cosmological parameters.

### 3.4. *Contrastive Embedding Learning*

A GNN-based embedder is trained using a supervised contrastive loss (NT-Xent) to learn representations of merger trees that are sensitive to cosmological parameters. The GNN architecture consists of 3 GCNConv layers with 128 hidden channels, followed by a global mean pooling layer and a 2-layer MLP head projecting to a 64-dimensional embedding space. The training utilizes the augmented training set (1400 trees: 700 original + 700 augmented).

*Training Dynamics:* The model is trained for 50 epochs. The training loss decreases from an initial value of approximately 3.29 to a final value of 2.46. The validation loss, calculated on the original validation set, starts at 3.13 and reaches a minimum of 2.32 around epoch 42, after which it shows some minor fluctuations. The consistent decrease in both training and validation loss (for most of the training) indicates that the GNN is successfully learning to differentiate between trees from different cosmologies. The best model, based on the minimum validation loss, is saved for subsequent LFI. The reduction in validation loss signifies that the GNN is not only memorizing the training data but also generalizing to unseen merger trees from similar cosmologies.

*Embedding Space Visualization:* The learned 64-dimensional embeddings from the validation set are projected into 2D using t-SNE and PCA for visualization, with points colored by their true (unnormalized) $\Omega_m$ and $\sigma_8$ values. These visualizations provide a qualitative assessment of how well the contrastive learning process has organized the embedding space based on cosmological parameters.

#### 3.4.1. *t-SNE plots*

These plots reveal some level of structure in the embedding space. For $\Omega_m$, there appears to be a discernible gradient or clustering, where trees with similar $\Omega_m$ values tend to be located closer to each other. The separation is not perfectly clean, indicating overlapping distributions, but a trend is visible. For $\sigma_8$, the t-SNE visualization also shows some grouping, though perhaps less distinct than for $\Omega_m$. This suggests that the GNN has successfully learned to encode information about $\Omega_m$ and $\sigma_8$ into the embeddings, allowing trees from simi-

lar cosmologies to cluster together. Figure 20 shows the t-SNE visualization for $\Omega_m$.

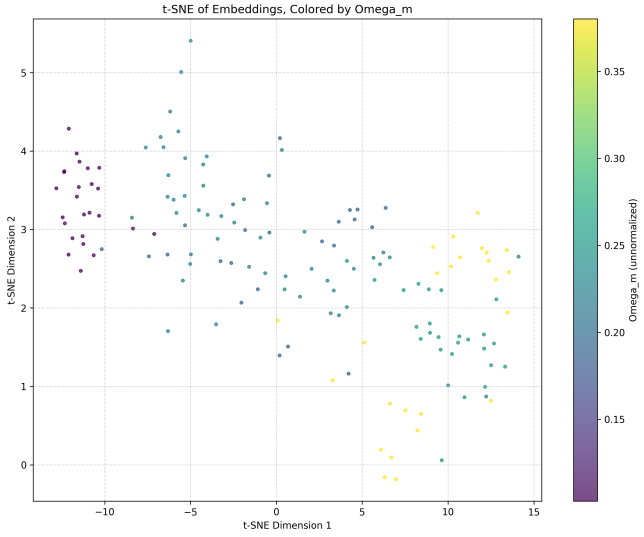


**Figure 20.** t-SNE visualization of the 64-dimensional contrastive embeddings of merger trees from the validation set, colored by their corresponding unnormalized $\Omega_m$ values. The structure observed in the embedding space suggests that the GNN has learned to encode cosmological information, with similar $\Omega_m$ values tending to cluster together.

### 3.4.2. *PCA plots*

PCA, being a linear dimensionality reduction technique, shows a less structured separation compared to t-SNE. However, faint gradients corresponding to $\Omega_m$ and $\sigma_8$ can still be observed, suggesting that the principal components of the embedding space do capture some cosmological information. While PCA is less effective at revealing complex non-linear relationships, the presence of these gradients indicates that the embeddings are at least partially organized along linear axes that correspond to the cosmological parameters.

Overall, the visualizations suggest that the contrastive learning process successfully organized the embedding space such that cosmological information is encoded, making these embeddings suitable as summary statistics for LFI. The ability of the GNN to automatically extract and encode these features from the raw merger tree data is a key advantage of this approach.

## 3.5. *Likelihood-Free Inference and Calibration*

The learned embeddings are used as summary statistics within an SNPE framework to infer posterior distributions for $\Omega_m$ and $\sigma_8$. A Neural Spline Flow is used as the density estimator, trained on the (embedding, normalized cosmology) pairs from the augmented training set.

*Posterior Inference:* Posterior distributions are estimated for trees in the test set. Figure 21 shows the 1D marginalized posteriors for $\Omega_m$ and $\sigma_8$, and the 2D joint posterior, with the true cosmological parameters overlaid for an example test merger tree. Qualitatively, for the depicted samples: * The true values of $\Omega_m$ are generally well-centered within the 1D marginalized posteriors, and the posteriors are reasonably peaked. * The true values of $\sigma_8$ also tend to fall within the inferred posteriors, although the constraints appear somewhat broader or occasionally less centered compared to $\Omega_m$ in these examples. * The 2D posteriors show some level of degeneracy between $\Omega_m$ and $\sigma_8$, which is common in cosmological inference.

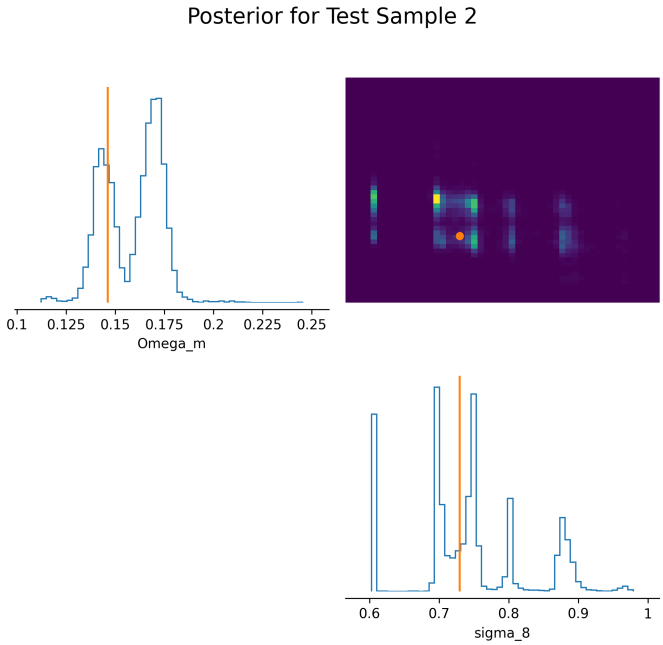


**Figure 21.** Posterior distributions for cosmological parameters $\Omega_m$ and $\sigma_8$ inferred using Sequential Neural Posterior Estimation (SNPE) for an example test merger tree, using contrastively learned GNN embeddings as summary statistics. The true parameter values are marked by vertical lines in the 1D posteriors and an orange dot in the 2D joint posterior. The inferred posteriors, particularly for $\sigma_8$, exhibit some deviation from the true values, indicating areas for improvement in the embedding space and inference pipeline.

*Simulation-Based Calibration (SBC):* The Simulation-Based Calibration (SBC) analysis, crucial for assessing the statistical validity of the inferred posteriors, could not be completed due to a technical issue encountered during Phase 5 execution. Consequently, SBC rank histograms were not generated, and a full assessment of posterior calibration is pending resolution of this issue.

This is a significant limitation in validating the LFI pipeline.

*Quantitative Performance Metrics:* Despite the inability to perform full SBC, quantitative metrics are calculated on the test set: * *Mean Squared Error (MSE):* The MSE between the posterior mean (unnormalized) and the true (unnormalized) cosmological parameters are: * $\Omega_m$: 0.00069 * $\sigma_8$: 0.006412 Taking the square root, the Root Mean Squared Error (RMSE) for $\Omega_m$ is approximately 0.026. Given the $\Omega_m$ range of 0.103 to 0.4734 (a span of 0.3704), this RMSE represents about 7.0% of the parameter range. For $\sigma_8$, the RMSE is approximately 0.080. Given the $\sigma_8$ range of 0.603 to 0.9918 (a span of 0.3888), this RMSE represents about 20.6% of its range. These results suggest that $\Omega_m$ is constrained more precisely than $\sigma_8$ by the learned embeddings.

* *Coverage Probability:* The coverage probability for the 90% credible interval (CI) is calculated for the unnormalized parameters: * $\Omega_m$: 94.67% * $\sigma_8$: 80.67% For $\Omega_m$, the coverage is very close to the nominal 90

### 3.6. *Summary of Results*

The results of this study demonstrate the potential of contrastive learning and GNNs for extracting cosmologically relevant information from dark matter halo merger trees. The engineered global features revealed strong correlations between merger tree properties and cosmological parameters, particularly $\Omega_m$, providing valuable insights into the underlying physics of structure formation. The data augmentation strategy, aimed at mitigating assembly bias, successfully introduced variability in halo concentrations. The contrastive learning framework successfully trained a GNN to produce embeddings that encode cosmological information, as evidenced by the loss curves and embedding space visualizations. The LFI results are encouraging for $\Omega_m$, with good MSE and coverage probability. However, the inference for $\sigma_8$ is less precise and its posteriors are undercovered, highlighting an area for future improvement. The inability to perform a full Simulation-Based Calibration due to technical issues is a significant limitation that must be addressed in future work. Overall, this study provides a promising foundation for developing robust and accurate likelihood-free cosmological inference methods based on merger tree data.

## 4. CONCLUSIONS

### 4.1. *Summary*

This paper addresses the challenge of cosmological inference from dark matter halo merger trees, which is complicated by the complex relationships between tree structure, assembly bias, and underlying cosmological parameters. To tackle this, we developed a contrastive learning framework that generates merger tree embeddings sensitive to cosmological parameters while mitigating the impact of assembly bias. A Graph Neural Network (GNN) was trained on merger trees from N-body simulations, employing a contrastive loss function to cluster trees originating from the same cosmology within the embedding space. To enhance robustness against assembly bias, we augmented the training data by introducing variations in halo concentrations conditional on halo mass. These learned embeddings then served as summary statistics for likelihood-free inference (LFI) using Sequential Neural Posterior Estimation (SNPE) to estimate the posterior distribution of $\Omega_m$ and $\sigma_8$.

### 4.2. *Methods and Datasets*

Our analysis was based on a dataset of 1000 dark matter halo merger trees from 40 unique cosmologies. Each tree was represented as a graph, with nodes characterized by mass, concentration, Vmax, and scale factor. We engineered 35 global features for each tree, capturing total mass, mass-weighted mean properties, main progenitor branch characteristics, assembly history proxies, assembly bias proxies, and tree structural metrics. To mitigate assembly bias, we augmented the training data by varying halo concentrations based on a mass-concentration relation derived from the simulations. A GNN was trained using a contrastive loss function to generate embeddings of the merger trees. These embeddings were then used as summary statistics in an SNPE framework to infer the posterior distributions of $\Omega_m$ and $\sigma_8$.

### 4.3. *Results*

The engineered global features revealed strong correlations between merger tree properties and cosmological parameters, particularly $\Omega_m$. The data augmentation strategy successfully introduced variability in halo concentrations. The contrastive learning framework successfully trained a GNN to produce embeddings that encode cosmological information, as evidenced by the decrease in training and validation loss and the structure observed in the embedding space visualizations. The LFI results showed good performance for $\Omega_m$, with a Root Mean Squared Error (RMSE) representing about 7.0% of the parameter range and a coverage probability close to the nominal value of 90%. However, the inference for $\sigma_8$ was less precise, with an RMSE representing about 20.6% of its range, and the inferred posteriors exhibited undercoverage (80.67% coverage for a 90% credible interval), indicating that the posteriors were too

narrow and did not capture the true parameter value with the expected frequency.

### 4.4. *Learnings and Future Directions*

This study demonstrates the potential of contrastive learning and GNNs for extracting cosmologically relevant information from dark matter halo merger trees. The strong correlations between engineered features and cosmological parameters provide valuable insights into the underlying physics of structure formation and can inform the development of analytic models. The successful training of the GNN and the generation of informative embeddings highlight the power of this approach for dimensionality reduction and feature extraction. However, the undercoverage observed for $\sigma_8$ suggests that further refinement of the method is needed. Future work should focus on improving the sensitivity of the embeddings to $\sigma_8$, potentially by incorporating additional features or modifying the GNN architecture or training procedure. Furthermore, the technical issues that prevented the completion of the Simulation-Based Calibration (SBC) must be resolved to ensure the statistical validity of the inferred posteriors. Overall, this study provides a promising foundation for developing robust and accurate likelihood-free cosmological inference methods based on merger tree data.

## REFERENCES


Bansal, A., Ichiki, K., Tashiro, H., & Matsuoka, Y. 2023, Evolution of the mass relation between supermassive black holes and dark matter halos across the cosmic time, doi: https://doi.org/10.1093/mnras/stad1608

Benson, A. J., Farahi, A., Cole, S., et al. 2012, Dark Matter Halo Merger Histories Beyond Cold Dark Matter: I - Methods and Application to Warm Dark Matter, doi: https://doi.org/10.1093/mnras/sts159

Biviano, A., Moretti, A., Paccagnella, A., et al. 2017, The concentration-mass relation of clusters of galaxies from the OmegaWINGS survey, doi: https://doi.org/10.1051/0004-6361/201731289

Bose, S., Eisenstein, D. J., Hadzhiyska, B., Garrison, L. H., & Yuan, S. 2022, Constructing high-fidelity halo merger trees in AbacusSummit, doi: https://doi.org/10.1093/mnras/stac555

Burgarella, D., Bogdanoska, J., Nanni, A., et al. 2022, IR characteristic emission and dust properties of star-forming galaxies at 4.5 < z < 6.2, doi: https://doi.org/10.1051/0004-6361/202142554

Cao, L., Jia, P., Li, J., et al. 2025, Image Pre-Processing Framework for Time-Domain Astronomy in the Artificial Intelligence Era. https://arxiv.org/abs/2502.10783

Cavelan, A., Cabezón, R. M., Korndorfer, J. H. M., & Ciorba, F. M. 2020, Finding Neighbors in a Forest: A b-tree for Smoothed Particle Hydrodynamics Simulations. https://arxiv.org/abs/1910.02639

Chatterjee, A., & Villaescusa-Navarro, F. 2025, Cosmology from Point Clouds with Dark Matter Halos from the Quijote Simulations, doi: https://doi.org/10.3847/1538-4357/adc99d

Chuang, C.-Y., Jespersen, C. K., Lin, Y.-T., Ho, S., & Genel, S. 2023, Leaving No Branches Behind: Predicting Baryonic Properties of Galaxies from Merger Trees. https://arxiv.org/abs/2311.09162

Coley, A. A., Santos, B., & Sanghai, V. A. A. 2018, Data Analysis and Phenomenological Cosmology, doi: https://doi.org/10.1088/1475-7516/2019/05/039

Conselice, C. J., Mundy, C. J., Ferreira, L., & Duncan, K. 2022, A direct measurement of galaxy major and minor merger rates and stellar mass accretion histories at $z < 3$ using galaxy pairs in the REFINE survey, doi: https://doi.org/10.3847/1538-4357/ac9b1a

de Santi, N. S. M., Rodrigues, N. V. N., Montero-Dorta, A. D., et al. 2022, Mimicking the halo-galaxy connection using machine learning, doi: https://doi.org/10.1093/mnras/stac1469

Erickson, S., Wagner-Carena, S., Marshall, P., et al. 2024, Lens Modeling of STRIDES Strongly Lensed Quasars using Neural Posterior Estimation. https://arxiv.org/abs/2410.10123

Frailis, M., Maris, M., Zacchei, A., et al. 2010, A systematic approach to the Planck LFI end-to-end test and its application to the DPC Level 1 pipeline, doi: https://doi.org/10.1088/1748-0221/4/12/T12021

Gilman, D., Du, X., Benson, A., et al. 2019, Constraints on the mass-concentration relation of cold dark matter halos with 11 strong gravitational lenses, doi: https://doi.org/10.1093/mnrasl/slz173

Gondhalekar, Y., Chies-Santos, A. L., de Souza, R. S., et al. 2024, Systematic analysis of jellyfish galaxy candidates in Fornax, Antlia, and Hydra from the S-PLUS survey: A self-supervised visual identification aid. https://arxiv.org/abs/2406.04213

Gu, Q., Guo, Q., Zhang, T., et al. 2022, The halo concentration and mass relation traced by satellite galaxies. https://arxiv.org/abs/2212.10232

Jespersen, C. K., Cranmer, M., Melchior, P., et al. 2022, `Mangrove`: Learning Galaxy Properties from Merger Trees, doi: https://doi.org/10.3847/1538-4357/ac9b18

Jia, H. 2024, Cosmological Analysis with Calibrated Neural Quantile Estimation and Approximate Simulators. https://arxiv.org/abs/2411.14748

Jiang, F., & van den Bosch, F. C. 2013, Generating Merger Trees for Dark Matter Haloes: A Comparison of Methods, doi: https://doi.org/10.1093/mnras/stu280

Konar, K., Reischke, R., Hagstotz, S., Nicola, A., & Hildebrandt, H. 2024, Constraining the dispersion measure redshift relation with simulation-based inference. https://arxiv.org/abs/2410.07084

Kosiba, M., Cerardi, N., Pierre, M., et al. 2024, The cosmological analysis of X-ray cluster surveys: VI. Inference based on analytically simulated observable diagrams, doi: https://doi.org/10.1051/0004-6361/202450499

Lacerna, I., Padilla, N., & Stasyszyn, F. 2014, The nature of assembly bias - III. Observational properties, doi: https://doi.org/10.1093/mnras/stu1318

Lehman, K., Krippendorf, S., Weller, J., & Dolag, K. 2024, Learning Optimal and Interpretable Summary Statistics of Galaxy Catalogs with SBI. https://arxiv.org/abs/2411.08957

Mao, R., Lee, J. E., Burke, O., et al. 2024, Calibrating approximate Bayesian credible intervals of gravitational-wave parameters, doi: https://doi.org/10.1103/PhysRevD.109.083002

Moss, A. 2025, The AI Cosmologist I: An Agentic System for Automated Data Analysis. https://arxiv.org/abs/2504.03424

Nerval, S. K., Hornecker, E., Guan, Y., et al. 2025, The Atacama Cosmology Telescope: The Development of Machine Learning Tools for Detecting Millimeter Sources in Timestream Pre-processing. https://arxiv.org/abs/2503.10798

Nguyen, T., Modi, C., Yung, L. Y. A., & Somerville, R. S. 2024, FLORAH: A generative model for halo assembly histories, doi: https://doi.org/10.1093/mnras/stae2001

Paranjape, A., & Padmanabhan, N. 2017, Halo assembly bias from Separate Universe simulations, doi: https://doi.org/10.1093/mnras/stx659

Parkinson, H., Cole, S., & Helly, J. 2007, Generating Dark Matter Halo Merger Trees, doi: https://doi.org/10.1111/j.1365-2966.2007.12517.x

Pearson, W. J., Rodriguez-Gomez, V., Kruk, S., & Margalef-Bentabol, B. 2024, Determining the time before or after a galaxy merger event, doi: https://doi.org/10.1051/0004-6361/202449532

Perez, N. B., Brüggen, M., Kasieczka, G., & Lucie-Smith, L. 2025, Classification of Radio Sources Through Self-Supervised Learning. https://arxiv.org/abs/2503.19111

Poulton, R. J. J., Robotham, A. S. G., Power, C., & Elahi, P. J. 2018, Observing Merger Trees in a New Light, doi: https://doi.org/10.1017/pasa.2018.34

Pu, S.-Y., Cooper, A. P., Grand, R. J. J., Gómez, F. A., & Monachesi, A. 2025, Progenitor diversity in the accreted stellar halos of Milky Way-like galaxies. https://arxiv.org/abs/2410.13491

Roncoli, A., Ćiprijanović, A., Voetberg, M., Villaescusa-Navarro, F., & Nord, B. 2024, Domain Adaptive Graph Neural Networks for Constraining Cosmological Parameters Across Multiple Data Sets. https://arxiv.org/abs/2311.01588

Shojaei, M. R., Tavasoli, S., & Ghafour, P. 2025, Star-Forming vs. Quenched Galaxies in Voids: Insights into the Role of Mergers. https://arxiv.org/abs/2501.16545

Smith, W., Berlind, A., & Sinha, M. 2024, Reversing Arrested Development: A New Method to Address Halo Assembly Bias. https://arxiv.org/abs/2410.06130

Stölzner, B., Wright, A. H., Asgari, M., et al. 2025, KiDS-Legacy: Consistency of cosmic shear measurements and joint cosmological constraints with external probes. https://arxiv.org/abs/2503.19442

Tang, K. S., & Ting, Y.-S. 2022, Galaxy Merger Reconstruction with Equivariant Graph Normalizing Flows. https://arxiv.org/abs/2207.02786

Verde, L. 2009, Statistical methods in cosmology, doi: https://doi.org/10.1007/978-3-642-10598-2_4

Wagner-Carena, S., Lee, J., Pennington, J., et al. 2024, A Strong Gravitational Lens Is Worth a Thousand Dark Matter Halos: Inference on Small-Scale Structure Using Sequential Methods. https://arxiv.org/abs/2404.14487

Wilkinson, A., Radev, R., & Alonso-Monsalve, S. 2025, Contrastive Learning for Robust Representations of Neutrino Data. https://arxiv.org/abs/2502.07724

Wu, J. F., Jespersen, C. K., & Wechsler, R. H. 2024, How the Galaxy-Halo Connection Depends on Large-Scale Environment, doi: https://doi.org/10.3847/1538-4357/ad7bb3

Yang, D., & Yu, H.-B. 2023, A graph model for the clustering of dark matter halos, doi: https://doi.org/10.1103/PhysRevResearch.5.043187

Zhang, K., Bloom, J. S., van der Walt, S., & Hernitschek, N. 2023, nbi: the Astronomer's Package for Neural Posterior Estimation. https://arxiv.org/abs/2312.03824

Zhang, S., Fang, G., Song, J., et al. 2024, Preparation for CSST: Star-galaxy Classification using a Rotationally Invariant Supervised Machine Learning Method, doi: https://doi.org/10.1088/1674-4527/ad6fe6

Zhang, T., Mao, T., Xu, W., & Li, G. 2025, Prediction of Individual Halo Concentrations Across Cosmic Time Using Neural Networks, doi: https://doi.org/10.3390/universe11020037

Ángel Chandro-Gómez, del P. Lagos, C., Power, C., et al. 2025, On the accuracy of dark matter halo merger trees and the consequences for semi-analytic models of galaxy formation, doi: https://doi.org/10.1093/mnras/staf519