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

Predicting the Direction of Dark Matter Halo Concentration Evolution with Graph Neural Networks and Contrastive Learning

Denario[1]

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

ABSTRACT

Understanding the evolution of dark matter halo concentration is crucial for galaxy formation models. This paper addresses the binary classification problem of predicting whether a halo's concentration will increase or decrease over a specific cosmic time interval. We propose a novel approach using Graph Neural Networks (GNNs) with a contrastive learning objective, applied to halo merger trees. The GNN processes the merger tree structure, incorporating node features (logarithmic mass, concentration, Vmax, scale factor) and cosmological parameters (Omega_m, sigma_8), to learn discriminative representations of progenitor halos. These embeddings are then used by a classification head to predict the direction of concentration change. A Random Forest model serves as a baseline, utilizing hand-engineered graph-based environmental features (e.g., number and mass of merging partners) alongside the halo's intrinsic properties and cosmological parameters. Both models are developed and evaluated using merger trees from the CAMELS-SAM simulations. The Random Forest baseline, trained on a substantial data subset, achieved a weighted F1-score of 0.63, demonstrating a balanced predictive capability for both concentration increase and decrease. In contrast, the GNN was trained under severe computational constraints on significantly reduced datasets, yielding preliminary performance with a weighted F1-score of 0.485. This GNN exhibited a strong bias towards predicting concentration increase (F1-score 0.69 for increase vs. 0.23 for decrease), indicative of severe underfitting. Ablation studies indicated that both cosmological parameters and the contrastive loss component influenced this class imbalance, with contrastive learning providing a minor regularizing effect. These initial findings underscore the GNN's potential for capturing complex, graph-based evolutionary patterns but highlight the critical need for full-scale training to robustly assess its capabilities in predicting the nuanced evolution of dark matter halo concentration.

1. INTRODUCTION

Dark matter halos are the fundamental building blocks of the cosmic web, serving as the gravitational scaffolding within which galaxies form and evolve. A key property characterizing these halos is their concentration, which quantifies the steepness of their density profile. Halo concentration is intimately linked to a galaxy's properties, influencing star formation rates, morphology, and the overall observable characteristics of galaxies. Consequently, understanding the evolution of dark matter halo concentration is paramount for accurate models of galaxy formation and the large-scale structure of the Universe.

The concentration of a dark matter halo is not static; it undergoes complex evolution over cosmic time, driven by continuous accretion of matter, stochastic merger events, and the underlying cosmological environment (Okoli 2017; Wang et al. 2020). Predicting the precise future state of a halo's concentration is an exceptionally challenging task due to the highly non-linear nature of gravitational collapse and the intricate, often chaotic, details of its assembly history (Wang et al. 2020). A significant complication arises from the fact that concentration evolution is not monotonic; a halo's concentration can either increase or decrease over a given time interval, depending on factors such as the mass accretion rate, the type and timing of mergers (e.g., major versus minor), and the specific cosmic epoch (Wang et al. 2020). This non-monotonic behavior makes it difficult for traditional analytical models and empirical relations to capture the full spectrum of evolutionary pathways, often leading to oversimplifications that limit their applicability in detailed galaxy formation simulations (Wang et al. 2020).

Rather than forecasting the exact future value of concentration, which is a continuous regression problem, this paper addresses a more focused yet equally crucial challenge: predicting the *direction* of dark matter halo concentration evolution (Okoli 2017; Wang et al. 2020).

We frame this as a binary classification problem: for a given halo, will its concentration increase or decrease over a specific cosmic time interval? This binary prediction provides vital insights for galaxy formation models, indicating whether a halo is likely to become more or less centrally dense, thereby influencing its baryonic content and subsequent star formation (Dou et al. 2025). The difficulty lies in extracting the subtle, non-local, and hierarchical patterns embedded within a halo's assembly history and its environment that dictate this directional change (Wang et al. 2020).

To tackle this intricate problem, we propose a novel approach leveraging Graph Neural Networks (GNNs) with a contrastive learning objective, applied to dark matter halo merger trees. Merger trees are ideal data structures for this task, as they inherently capture the hierarchical assembly history of halos (Parkinson et al. 2007; Jiang & van den Bosch 2013; Yung et al. 2024), representing a rich source of information about their progenitors, descendants, and merger events over cosmic time (Parkinson et al. 2007; Jiang & van den Bosch 2013). GNNs are uniquely suited to process such graph-structured data (Nguyen et al. 2025), enabling them to learn complex relationships and extract features that encode both the topological and attribute-based information within these trees.

Our GNN model processes the merger tree structure (Tang & Ting 2022), taking as input detailed node features for each halo within the tree, specifically its logarithmic mass, logarithmic concentration, logarithmic maximum circular velocity (Vmax), and its scale factor (Tang & Ting 2022). Crucially, the model also incorporates global cosmological parameters of the simulation, namely Ω_m and σ_8, allowing it to learn representations sensitive to both an individual halo's unique assembly history and the broader cosmic context (Wu et al. 2024). The GNN is designed to learn discriminative representations (embeddings) of progenitor halos, which are then used by a classification head to predict the binary direction of concentration change (Tang & Ting 2022; Wu et al. 2024).

A key innovation in our methodology is the integration of a contrastive learning objective (Zhu et al. 2025). This objective guides the GNN to produce embeddings where halos exhibiting the same direction of concentration change (e.g., both increasing) are pulled closer together in the embedding space, while halos experiencing different directions of change (e.g., one increasing, one decreasing) are pushed further apart (Zhu et al. 2025). This self-supervised component encourages the GNN to learn highly informative and robust representations that are intrinsically relevant to the classification task (Zhu et al. 2025), enhancing its ability to discern subtle patterns governing concentration evolution.

We develop and rigorously evaluate both our proposed GNN model and a robust baseline using merger trees extracted from the CAMELS-SAM simulations (Ramakrishnan & Velmani 2022; Huang et al. 2025). The CAMELS-SAM dataset provides a rich and diverse collection of halo evolutionary histories across a wide range of cosmological parameters, offering an ideal environment for training and testing our models (Lovell et al. 2024; Huang et al. 2025).

The baseline model, a Random Forest classifier, is designed to provide a strong point of comparison (Contardo et al. 2025). It utilizes a comprehensive set of hand-engineered features, including intrinsic halo properties (logarithmic mass, concentration, Vmax, scale factor), cosmological parameters (Ω_m, σ_8) (Huang et al. 2025), and crucially, graph-based environmental features derived from the merger tree itself (Huang et al. 2025). These environmental features, such as the number and mass of merging partners, aim to capture aspects of a halo's local assembly environment that are hypothesized to influence its concentration evolution.

By comparing the GNN's performance against this feature-rich baseline, we aim to assess the GNN's ability to implicitly learn and exploit the complex, graph-based patterns that govern halo concentration evolution, and to understand the contribution of the contrastive learning objective to this challenging prediction task. The performance of both models is quantified using standard classification metrics, including precision, recall, and F1-score, with a particular emphasis on the models' balanced predictive capability for both concentration increase and decrease.

2. METHODS

The overarching goal of this research is to predict the binary direction of dark matter halo concentration evolution (increase or decrease) over cosmic time. To achieve this, we employ a novel approach leveraging Graph Neural Networks (GNNs) with a contrastive learning objective, benchmarked against a Random Forest model utilizing hand-engineered features. This section details the dataset, preprocessing steps, model architectures, training procedures, and evaluation metrics.

2.1. *Dataset and Data Preprocessing*

Our study utilizes dark matter halo merger trees extracted from the CAMELS-SAM simulations, a suite specifically designed to explore the interplay between cosmology and galaxy formation (Ramakrishnan & Velmani 2022; Huang et al. 2025). These simulations provide a rich source of halo assembly histories across a

diverse range of cosmological parameters (Huang et al. 2025). The merger trees are organized into three PyTorch files: 'CS_tree_train.pt' (containing 14,997 trees for training), 'CS_tree_val.pt' (5,099 trees for validation), and 'CS_tree_test.pt' (4,900 trees for final testing) (Huang et al. 2025). Each file is loaded using 'torch.load()', yielding a list of PyTorch Geometric 'Data' objects, where each object represents a single merger tree. As per the dataset specifications, the fields 'lh_id', 'mask_main', and 'node_halo_id' were excluded from processing.

2.1.1. *Halo Properties and Cosmological Parameters*

Each node in a merger tree represents a dark matter halo at a specific point in cosmic time (Gómez et al. 2021; Nguyen et al. 2025). For each halo (node), the following intrinsic properties are used as node features, denoted as $\mathbf{x}$ (Nguyen et al. 2025):

- $\log_{10}$(mass): The base-10 logarithm of the halo's virial mass in solar masses.
- $\log_{10}$(concentration): The base-10 logarithm of the halo's NFW concentration parameter.
- $\log_{10}$(Vmax): The base-10 logarithm of the maximum circular velocity of the halo in km/s.
- scale_factor: The cosmological scale factor $a = 1/(1+z)$, indicating the cosmic epoch of the halo.

These four features constitute the 4-dimensional feature vector for each node (Jahin et al. 2025,?).

In addition to halo-specific properties, each merger tree is associated with global cosmological parameters of the simulation it originated from (Nguyen et al. 2025). These parameters are used as graph-level features, denoted as $\mathbf{y}$ (Nguyen et al. 2025):

- Ω_m: The present-day matter density parameter.
- σ_8: The amplitude of matter fluctuations on 8 h^{-1} Mpc scales.

These two parameters provide crucial context about the large-scale cosmic environment influencing halo evolution (Dooley et al. 2014; Ishiyama et al. 2025).

2.1.2. *Feature Normalization*

To ensure numerical stability and improve model training, all node features ($\mathbf{x}$) and graph features ($\mathbf{y}$) are standardized (Islam 2024; Pinheiro et al. 2025). The mean and standard deviation for each of the four node features and two graph features are computed exclusively from the training set.

These calculated statistics are then used to normalize the training, validation, and test sets by subtracting the mean and dividing by the standard deviation (Pinheiro et al. 2025). This ensures that the models are trained on features with zero mean and unit variance, preventing features with larger magnitudes from dominating the learning process (de Amorim et al. 2022; Pinheiro et al. 2025).

2.1.3. *Defining Halo Transitions and Target Variable*

The core task is to predict the direction of concentration change for a halo between two consecutive cosmic time steps (Zhang et al. 2025). We define "transitions" based on progenitor-descendant pairs within the merger trees (Johnson et al. 2021).

For every directed edge (u, v) in a tree's 'edge_index', where node u is a progenitor of node v, we consider this a transition. Node u represents the halo at an earlier scale factor sf_u, and node v represents its descendant at a later scale factor sf_v. We only consider transitions where $\text{sf}_v > \text{sf}_u$.

The concentration of halo u at sf_u is $\text{conc}_u = \mathbf{x}[u, 1]$ (the second element of its feature vector). Similarly, the concentration of its descendant v at sf_v is $\text{conc}_v = \mathbf{x}[v, 1]$. The binary target variable, $Y_{\text{transition}}$, is defined as:

- $Y_{\text{transition}} = 1$ if $\text{conc}_v > \text{conc}_u$ (concentration increased).
- $Y_{\text{transition}} = 0$ if $\text{conc}_v \leq \text{conc}_u$ (concentration decreased or remained the same).

This binary classification setup directly addresses the problem of predicting the *direction* of concentration evolution, offering vital insights for galaxy formation models.

Each sample in our dataset for model training corresponds to a specific progenitor halo u and its associated transition, including its node features, the cosmological parameters of its parent tree, and the binary target $Y_{\text{transition}}$.

2.2. *Engineered Features for Baseline Model*

To establish a robust baseline, a Random Forest classifier is employed (Srivastava et al. 2025; Bluck et al. 2025), which requires a fixed-size feature vector for each sample. Unlike the GNN, which implicitly learns from graph structure, the baseline model relies on explicitly defined features (Shi et al. 2022; Carr et al. 2025). For each halo transition from progenitor u to descendant v, a comprehensive set of features is engineered (Srivastava et al. 2025):

1. **Intrinsic Progenitor Features**: Normalized $\log_{10}(\text{mass}_u)$, $\log_{10}(\text{concentration}_u)$, $\log_{10}(\text{Vmax}_u)$, and scale_factor_u.

2. **Descendant Features**: Normalized $\log_{10}(\text{mass}_v)$ and scale_factor_v.

3. **Temporal Feature**: The difference in scale factor, $\Delta\text{sf} = \text{scale\_factor}_v - \text{scale\_factor}_u$, capturing the duration of the evolution.

4. **Cosmological Parameters**: Normalized Ω_m and σ_8 associated with the merger tree.

5. **Graph-based Environmental Features**: These features capture aspects of the halo's local assembly environment, derived directly from the merger tree structure:

    - **Number of Merging Partners**: For a descendant halo v, we identify all its direct progenitors from the 'edge_index'. The 'num_merging_partners' is the count of progenitors of v other than u. This quantifies the number of simultaneous merger events contributing to v's formation.
    - **Total Mass of Other Merging Partners**: The sum of $10^{\log_{10}(\text{mass})}$ for all progenitors of v excluding u. This provides a measure of the total mass accreted from other sources during the $u \to v$ transition.
    - **Mass Ratio of Main Progenitor to Other Partners**: Calculated as $\text{mass}_u/(\text{total\_mass\_of\_other\_merging\_partners}+\epsilon)$, where ϵ is a small constant to prevent division by zero. This ratio indicates the relative dominance of u in forming v compared to other merging halos.
    - **Is Major Merger Progenitor**: A binary flag indicating whether u is involved in a major merger event to form v. This is defined if the mass of u is greater than 30% of the total mass of all other merging partners, or if the mass of u is within a factor of 3 of the most massive merging partner. This feature captures significant accretion events that are known to strongly influence concentration.

These engineered features are combined into a single vector for each transition, serving as input to the Random Forest classifier (Louppe 2015; Fallah 2023; Maturo & Porreca 2024).

2.3. *Graph Neural Network Architecture*

Our primary model is a Graph Neural Network (GNN) designed to process the hierarchical structure of merger trees and learn rich, discriminative representations of halos (Qin et al. 2024; Nguyen et al. 2025). The GNN operates on the principle of message passing, iteratively aggregating information from a halo's local neighborhood within the graph (Qin et al. 2024).

2.3.1. *GNN Encoder*

The GNN encoder is constructed using multiple layers of graph convolution (Tanis et al. 2024,?). Each layer performs a message-passing step, where information is exchanged between connected nodes (Tanis et al. 2024,?). Specifically, for each node (halo) u, its feature vector $\mathbf{h}_u^{(l)}$ at layer l is updated by aggregating information from its neighbors $v \in \mathcal{N}(u)$ and combining it with its own features from the previous layer (Tanis et al. 2024). This process can be generally described as:

$$\mathbf{h}_u^{(l+1)} = \text{UPDATE}^{(l+1)}\left(\mathbf{h}_u^{(l)}, \text{AGGREGATE}^{(l+1)}\left(\{\mathbf{h}_v^{(l)} \mid v \in \mathcal{N}(u)\right)\right.$$

The initial node features $\mathbf{h}_u^{(0)}$ consist of the normalized intrinsic halo properties ($\log_{10}$(mass), $\log_{10}$(concentration), $\log_{10}$(Vmax), scale_factor) (Larson et al. 2024; Wu et al. 2024). Crucially, the normalized cosmological parameters (Ω_m, σ_8) are concatenated to each node's feature vector at the input layer (Wu et al. 2024), allowing the GNN to learn representations sensitive to both local assembly history and the global cosmic context (Wu et al. 2024; Lee & Villaescusa-Navarro 2025).

The GNN architecture consists of three stacked `GraphConv` layers (similar to a GCN or GraphSAGE layer) (Javaloy et al. 2023; He et al. 2025), each followed by a ReLU activation function and batch normalization. The final layer outputs a fixed-dimensional embedding vector $\mathbf{z}_u$ for each progenitor halo u involved in a transition. This embedding $\mathbf{z}_u$ encapsulates the information learned from u's local neighborhood and its position within the merger tree, integrated over the layers.

2.3.2. *Contrastive Learning Objective*

To enhance the quality of the learned halo embeddings and make them more discriminative for our classification task, we incorporate a contrastive learning objective. This self-supervised component guides the GNN to produce embeddings where halos exhibiting the same direction of concentration change are pulled closer together in the embedding space, while halos experiencing different directions of change are pushed further apart.

For a given batch of halo transitions, let $\mathbf{z}_i$ be the embedding for an anchor progenitor halo u_i with an associated concentration evolution direction Y_i.

- **Positive Pairs**: Any other embedding $\mathbf{z}_j$ in the batch is considered a positive partner if $Y_j = Y_i$.
- **Negative Pairs**: $\mathbf{z}_j$ is a negative partner if $Y_j \neq Y_i$.

The Normalized Temperature-scaled Cross-Entropy (NT-Xent) loss function is employed for contrastive learning (Ågren 2022; Bleeker & de Rijke 2022; Fahim et al. 2024). For an anchor embedding $\mathbf{z}_i$ and one of its positive partners $\mathbf{z}_p$ from the batch, the loss is defined as (Ågren 2022):

$$\mathcal{L}_{\text{contrastive}} = -\log \left[\frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_p)/\tau)}{\sum_{k \in \text{batch}, k \neq i} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \right]$$

where $\text{sim}(\mathbf{a}, \mathbf{b})$ denotes the cosine similarity between vectors $\mathbf{a}$ and $\mathbf{b}$, calculated as $(\mathbf{a} \cdot \mathbf{b})/(||\mathbf{a}|| \cdot ||\mathbf{b}||)$ (Steck et al. 2024).

The hyperparameter τ is a temperature scaling factor that controls the spread of embeddings. The sum in the denominator includes all other samples in the batch, both positive and negative relative to $\mathbf{z}_i$. This loss is averaged over all anchor embeddings in the batch.

2.3.3. *Classification Head and Combined Training Objective*

A Multi-Layer Perceptron (MLP) serves as the classification head, taking the learned GNN embedding $\mathbf{z}_u$ for a progenitor halo u as input (Borzone et al. 2025; Duan et al. 2025). The MLP consists of two linear layers with a ReLU activation function between them. The final layer outputs a 2-dimensional logit vector, representing the unnormalized probabilities for the two classes (concentration increase and decrease/same).

The classification loss is computed using Binary Cross-Entropy (BCE) with logits (Ho & Wookey 2020; Wali 2022), comparing the predicted logits against the true binary target $Y_{\text{transition}}$.

The overall training objective for the GNN model is a weighted sum of the classification loss and the contrastive loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE\_classification}} + \alpha \cdot \mathcal{L}_{\text{contrastive}}$$

where α is a hyperparameter that controls the relative importance of the contrastive learning objective (Wei & Zhang 2025,?). This combined loss enables end-to-end training of the GNN encoder and the classification head, allowing the GNN to learn representations that are both self-supervisedly robust and directly relevant to the downstream classification task.

2.4. *Baseline Model: Random Forest Classifier*

To provide a strong and interpretable benchmark, a Random Forest classifier is implemented using the scikit-learn library. The Random Forest model is chosen for its robustness, ability to handle non-linear relationships, and its capacity to manage a mix of feature types.

For each halo transition, the Random Forest model is trained on the comprehensive set of 12 hand-engineered features described in Section 2 (Bluck et al. 2025). These include the intrinsic properties of the progenitor and descendant halos, the temporal difference in scale factor, the cosmological parameters, and the four graph-based environmental features derived from the merger tree structure (Srivastava et al. 2025). The target variable for the Random Forest is the same binary $Y_{\text{transition}}$ as for the GNN. Hyperparameters for the Random Forest, such as the number of estimators and maximum tree depth, are tuned using the validation set to optimize performance (Bluck et al. 2025).

2.5. *Training and Evaluation Protocol*

2.5.1. *Training Procedure*

Both the GNN and the Random Forest models are trained on the designated training set of halo transitions. For the GNN, training is performed using mini-batches of merger trees. The Adam optimizer is used for gradient descent, with an initial learning rate that is subject to hyperparameter tuning. During GNN training, the model's performance on the validation set is monitored after each epoch. This includes calculating Precision, Recall, and F1-score for the classification task, as well as the total loss.

Early stopping is implemented to prevent overfitting; training is halted if the weighted F1-score on the validation set does not improve for a predefined number of epochs, and the model weights corresponding to the best validation F1-score are saved. Due to significant computational constraints during initial development, the GNN was primarily trained on severely reduced datasets.

2.5.2. *Hyperparameter Tuning*

A systematic hyperparameter search is conducted using the validation set to optimize the performance of both models (Yuan et al. 2021). For the GNN, key hyperparameters tuned include the learning rate, the number of GNN layers, hidden dimension sizes, the loss weighting factor α, the temperature parameter τ for the contrastive loss, and the batch size (Plettenberg et al. 2025).

For the Random Forest baseline, hyperparameters such as the number of trees (estimators), maximum tree depth, and minimum samples per leaf are tuned.

2.5.3. *Final Evaluation*

Upon completion of training and hyperparameter tuning, the best-performing GNN model (selected based on validation F1-score) and the optimized Random Forest baseline model are evaluated on the entirely held-out test set of halo transitions.

Model performance is quantified using standard classification metrics: Precision, Recall, and F1-score. These metrics are reported for each class (concentration increase and decrease/same) to assess class-specific performance, as well as weighted and macro-averaged F1-scores to provide an overall measure of balanced predictive capability. The primary comparison between the GNN and the baseline model is made based on these comprehensive evaluation metrics.

2.5.4. *Implementation Details*

All models are implemented using PyTorch and PyTorch Geometric for the GNN (Wang & Zhang 2023; Wang & Shen 2024), and scikit-learn for the Random Forest. NumPy is used for general numerical operations and data manipulation. GNN training is accelerated using available GPU resources (Anik et al. 2024), while data preprocessing and baseline model training are performed on CPUs.

To ensure reproducibility of results, random seeds are explicitly set for PyTorch, NumPy, and any other relevant libraries throughout the experimental pipeline. The development process prioritizes initial verification on a small subset of data before scaling to the full datasets.

3. RESULTS

This section presents the findings from our investigation into predicting the binary direction of dark matter halo concentration evolution. We detail the characteristics of the processed data, the performance of the Random Forest baseline model, the outcomes from the Graph Neural Network (GNN) model, and insights gained from ablation studies designed to understand the contributions of specific GNN components.

3.1. *Data characterization and preprocessing*

Our study utilized dark matter halo merger trees sourced from the CAMELS-SAM simulations. Each node within these trees represents a halo at a specific cosmic epoch, characterized by four intrinsic properties: logarithmic mass, logarithmic concentration, logarithmic maximum circular velocity (Vmax), and the scale factor. Additionally, each merger tree is associated with two global cosmological parameters: Ω_m and σ_8. These parameters provide a broader cosmic context for the halo's evolution.

All node and graph features underwent standardization, with means and standard deviations computed exclusively from the training set. For node features, the approximate means were $[11.13, 0.74, 2.11, 0.38]$ and standard deviations were $[0.70, 0.36, 0.21, 0.18]$ for $\log_{10}$(mass), $\log_{10}$(concentration), $\log_{10}$(Vmax), and scale factor, respectively. The distributions of these node features before and after standardization are shown in Figure 1. For the graph-level cosmological parameters, the approximate means were $[0.30, 0.80]$ and standard deviations were $[0.12, 0.11]$ for Ω_m and σ_8. These distributions, both original and normalized, are presented in Figure 2. This normalization ensured that all features had a zero mean and unit variance, preventing features with larger magnitudes from disproportionately influencing model training.

The core of our predictive task involves analyzing "transitions" between progenitor (u) and descendant (v) halos. A transition is defined for any directed edge (u, v) where $\text{sf}_v > \text{sf}_u$. The binary target variable, $Y_{\text{transition}}$, indicates the direction of concentration change: $Y_{\text{transition}} = 1$ if $\text{conc}_v > \text{conc}_u$ (concentration increased), and $Y_{\text{transition}} = 0$ if $\text{conc}_v \leq \text{conc}_u$ (concentration decreased or remained the same). From an initial development subset of 100 trees, 110,838 such transitions were extracted. The distribution of this binary target variable, shown in Figure 3, exhibits a relatively balanced distribution: 51,707 instances of class 0 (decrease/no change) and 59,131 instances of class 1 (increase).

For the Random Forest baseline model, a set of 10 hand-engineered features was created for each transition. These included the four normalized intrinsic progenitor halo properties, the normalized scale factor of the descendant, the difference in scale factor (Δsf), the two normalized cosmological parameters, and four graph-based environmental features. The environmental features comprised the number of other merging partners contributing to the descendant halo, their total mass, the mass ratio of the main progenitor to these other partners, and a binary flag indicating if the main progenitor was involved in a major merger. These features were designed to capture the local assembly environment of the halo, which is hypothesized to influence its concentration evolution. The environmental features were also normalized (log transformation for total partner mass, then standardization). The distributions of

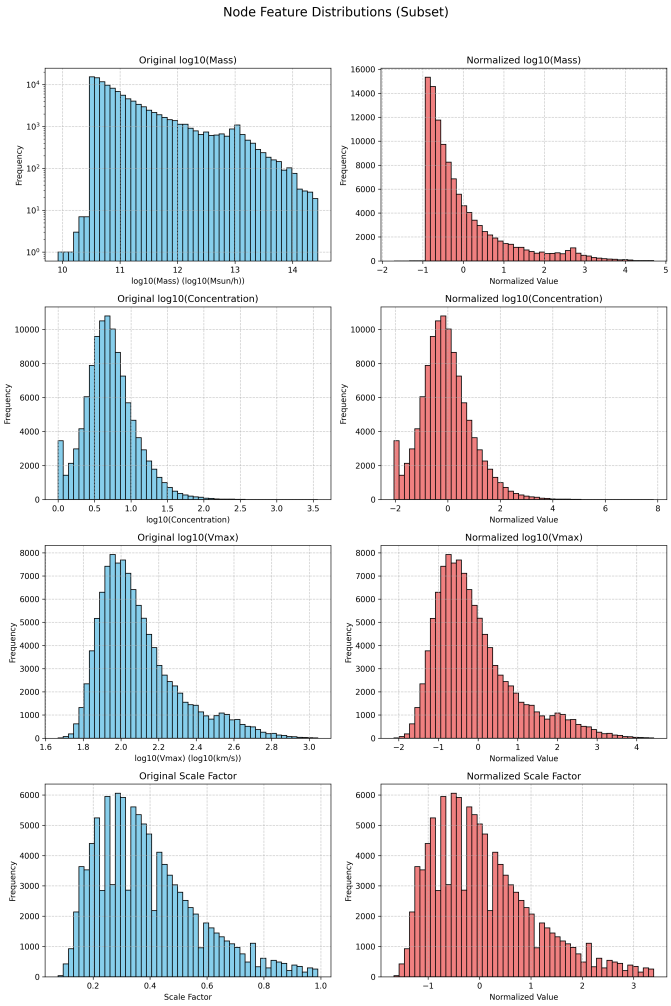

Figure 1. Distributions of the four halo node features ($\log_{10}$(Mass), $\log_{10}$(Concentration), $\log_{10}$(Vmax), and Scale Factor) before (blue) and after (red) standardization. The plots confirm that normalization effectively centers the feature distributions around zero mean and scales them to unit standard deviation, an essential preprocessing step for model training.

these engineered environmental features are presented in Figure 4.

It is crucial to note that, due to significant computational constraints (specifically, execution timeouts), the GNN model training, evaluation, and subsequent ablation studies were performed on severely reduced subsets of the full training, validation, and test datasets. For instance, GNN training was typically limited to 10-50 training trees instead of the approximately 15,000 available. While the baseline model was trained on a larger subset (200 training trees), it also did not utilize the full dataset. Consequently, the reported GNN performance metrics should be considered preliminary and indicative of potential capabilities rather than definitive assessments on the complete dataset.

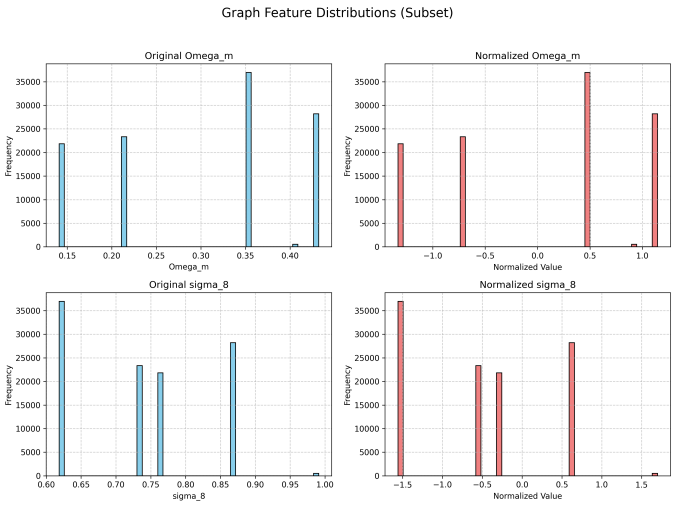

Figure 2. Histograms displaying the distributions of the cosmological parameters, Ω_m and σ_8, for a subset of the dataset. Original distributions are shown in blue, and their normalized counterparts are in red. This visualization confirms that the normalization process effectively centers and scales these graph-level features, which is essential for consistent model training.

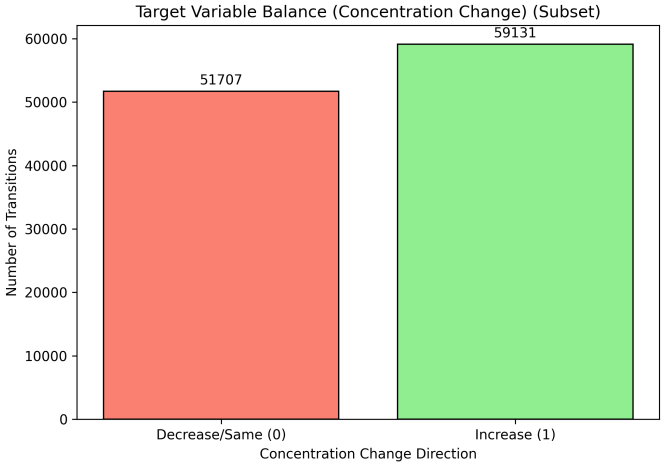

Figure 3. Distribution of the binary target variable for halo concentration change, derived from 110,838 transitions extracted from a development subset of 100 merger trees. The plot shows a relatively balanced distribution, with 51,707 instances of concentration decrease or no change (Class 0) and 59,131 instances of concentration increase (Class 1).

3.2. *Baseline model performance*

A Random Forest classifier was implemented as a robust baseline. This model utilized a total of 10 features per transition: the four normalized intrinsic properties

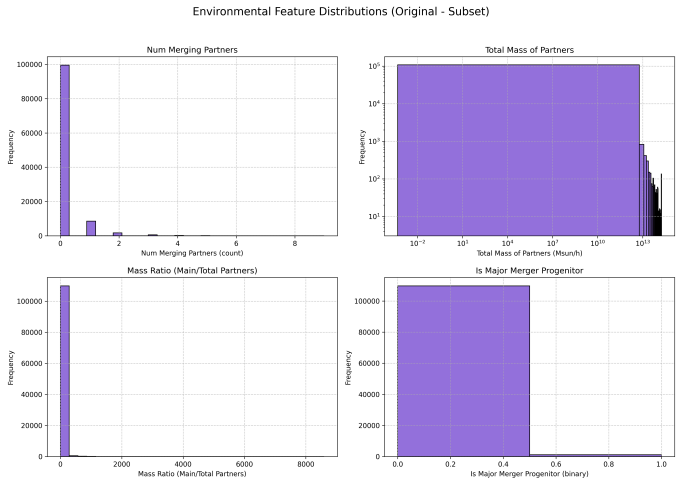

Figure 4. Distributions of four engineered environmental features for progenitor halos. The plots show that most halos have few merging partners, with their total mass spanning several orders of magnitude, and that major mergers are relatively infrequent. These characteristics describe the environmental context provided to the baseline model.

of the progenitor halo, the normalized descendant scale factor, the normalized scale factor difference, the two normalized cosmological parameters, and the four normalized/binary engineered environmental features detailed previously.

Hyperparameter tuning for the Random Forest was conducted using GridSearchCV on a subset of 200 training trees (221,770 transitions) and validated on a subset of 100 validation trees (109,866 transitions). The optimal hyperparameters identified were ‘’class_weight’: ’balanced_subsample’, ’max_depth’: 10, ’min_samples_leaf’: 20, ’min_samples_split’: 50, ’n_estimators’: 50‘.

The performance of the tuned baseline model on the validation subset is summarized in Table 1. The Random Forest baseline achieved an overall accuracy of 63% and a weighted average F1-score of 0.63. The model demonstrated a relatively balanced predictive capability, performing slightly better at identifying halos that would increase in concentration (F1-score of 0.66) compared to those that would decrease or maintain concentration (F1-score of 0.59). The confusion matrix in Figure 5 further illustrates these classification outcomes, showing a slight bias towards predicting concentration increase. This indicates that the engineered features, particularly the intrinsic properties of the progenitor halo, provide a reasonable basis for predicting the direction of concentration evolution.

An analysis of feature importances, depicted in Figure 6, revealed that the most influential features were ‘norm_log10_Concentration_prog‘, ‘norm_ScaleFactor_prog‘, ‘norm_log10_Mass_prog‘, and ‘norm_log10_Vmax_prog‘. These intrinsic properties of the progenitor halo appear to be the primary drivers of the classification decision. The cosmological parameters (Ω_m, σ_8) and the engineered graph-based environmental features (e.g., total mass of merging partners, number of merging partners) had lower but still measurable importance, suggesting they contribute to the prediction but are secondary to the halo’s immediate state.

Table 1. Baseline Random Forest Performance on Validation Subset

Metric	Decrease/Same (0)	Increase (1)	Macro Avg	Weighted A
Precision	0.60	0.65	0.62	0.63
Recall	0.58	0.67	0.62	0.63
F1-Score	0.59	0.66	0.62	0.63
Support	51259	58607	109866	109866
Overall Accuracy: 0.63				

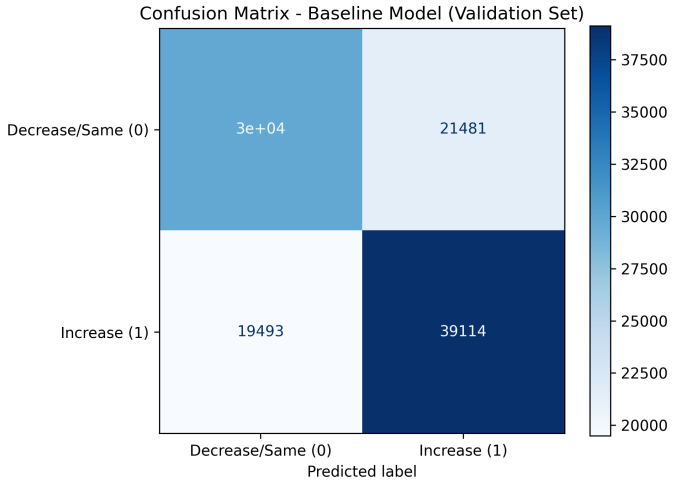

Figure 5. Confusion matrix of the baseline Random Forest model on the validation set, illustrating its classification of halo concentration changes into ’Decrease/Same’ (0) and ’Increase’ (1). The model correctly identified 30,000 instances of concentration decrease/same and 39,114 instances of concentration increase, indicating a modest overall performance with a slight bias towards predicting concentration increase.

3.3. *Graph neural network model performance*

Our primary model, a Graph Neural Network (GNN) based on GraphSAGE convolutional layers, was designed to learn representations directly from the merger tree structure. The GNN encoder learns node embeddings, which are then fed into a classification head to predict the binary target. A projection head was also incorporated for the contrastive learning objective.

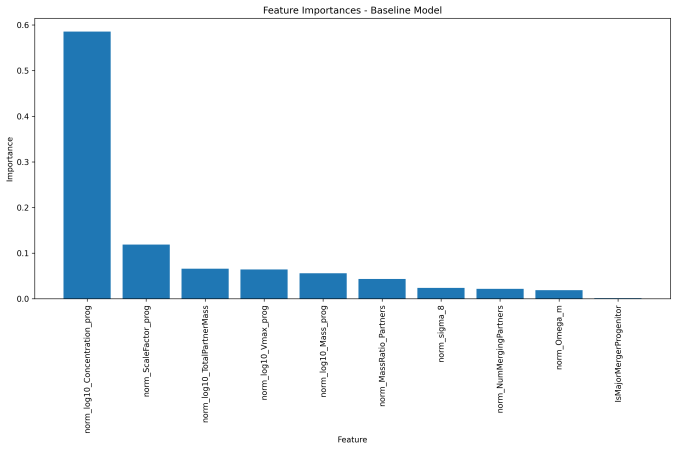

Figure 6. Bar plot showing feature importances for the Random Forest baseline model. The progenitor halo's normalized concentration, scale factor, mass, and Vmax are the most influential features for predicting the direction of halo concentration change. Cosmological parameters and engineered environmental features contribute with lower importance.

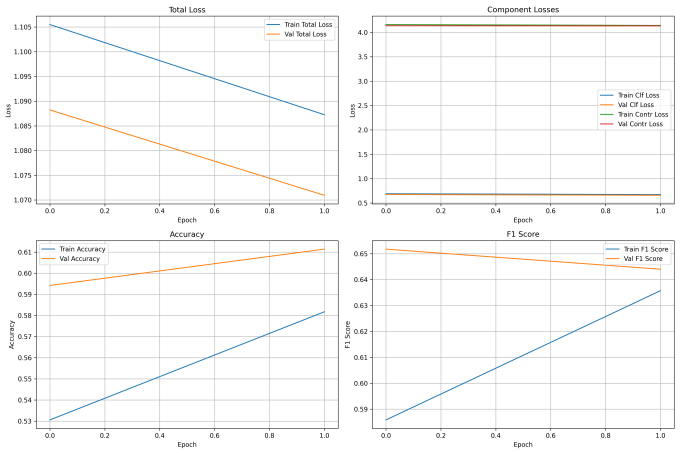

Figure 7. Learning curves for the Graph Neural Network (GNN) model over two training epochs. The top-left panel displays the decreasing total loss for both training and validation data, comprising a combined classification and contrastive objective. The top-right panel shows the unscaled classification and contrastive loss components. The bottom panels illustrate the general increase in training and validation accuracy and F1 score, respectively. These trends confirm the model's learning progression despite the limited training duration.

Each node's input features for the GNN consisted of its four intrinsic halo properties concatenated with the two global cosmological parameters, resulting in a 6-dimensional input feature vector per node.

The GNN was trained using the Adam optimizer with a combined loss function: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE\_classification}} + \alpha \cdot \mathcal{L}_{\text{contrastive}}$, where the contrastive loss weight α was set to 0.1. As previously highlighted, the GNN training was severely constrained computationally. It was performed on a significantly reduced dataset: 10 training trees (13,162 transitions), 5 validation trees (4,570 transitions), and 5 test trees (3,723 transitions), for only 2 epochs. Early stopping was implemented based on the validation F1-score.

The learning curves, displayed in Figure 7 and more comprehensively in Figure 8, indicated that both training and validation losses generally decreased over the two epochs, with corresponding increases in accuracy and F1-scores, albeit with some fluctuations. The validation F1-score for the 'Increase' class (class 1) reached 0.6823 at epoch 2, leading to the saving of the model weights. These trends confirm the model's learning progression despite the limited training duration and data.

The performance of this best GNN model on the reduced test set is presented in Table 2. The GNN achieved an overall accuracy of 56.2% and a weighted average F1-score of 0.485 on this severely restricted test set. A key observation is the pronounced class imbalance in the model's predictions, as visualized in the confusion matrix in Figure 9. The GNN exhibited a high recall of 0.90 for the 'Increase' class (class 1) but a very

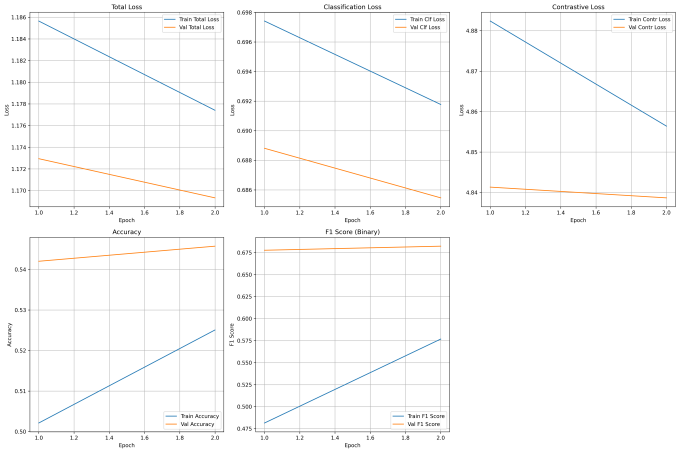

Figure 8. Learning curves for the Graph Neural Network (GNN) model, displaying the total, classification (BCE), and contrastive loss components, alongside overall accuracy and F1-score, for both training and validation datasets over two epochs. All loss components decrease, while accuracy and F1-score generally increase, indicating the model's learning progress in predicting the direction of halo concentration change. These trends suggest the model's learning was still ongoing given the limited training duration and data.

low recall of 0.15 for the 'Decrease/Same' class (class 0). This resulted in a high F1-score of 0.69 for class 1 but a considerably lower F1-score of 0.23 for class 0. This strong bias suggests that under these limited training conditions, the GNN predominantly learned to predict

an increase in concentration. The Area Under the Receiver Operating Characteristic curve (AUC), shown in Figure 10, was 0.5924, indicating that the model's discriminative power was only modest, slightly better than random chance.

Table 2. GNN Performance on Reduced Test Set

Metric	Decrease/Same (0)	Increase (1)	Macro Avg	Weig
Precision	0.55	0.56	0.56	
Recall	0.15	0.90	0.52	
F1-Score	0.23	0.69	0.46	
Support	1674	2049	3723	

Overall Accuracy: 0.562

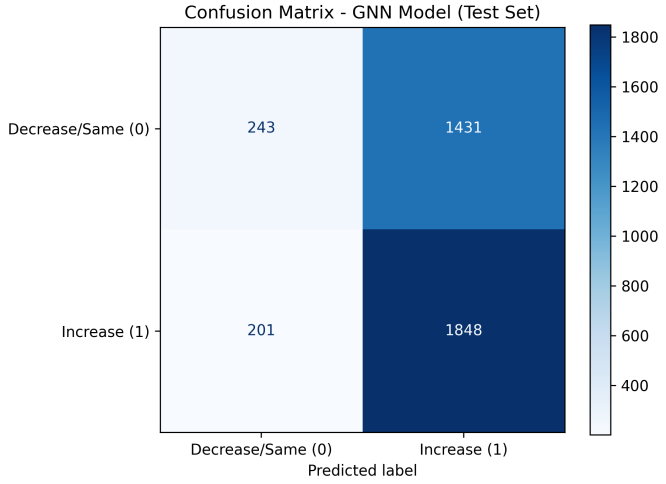

Figure 9. Confusion matrix for the GNN model on the reduced test set. It shows the true versus predicted direction of halo concentration change (Decrease/Same (0) or Increase (1)). The model exhibits a strong bias towards predicting an increase in concentration (Class 1), frequently misclassifying halos that decrease or maintain concentration (1431 instances) as increasing. This bias leads to high recall for concentration increase but poor performance for decrease/same.

A visualization of the GNN's progenitor node embeddings from a validation subset, shown in Figure 11, reveals substantial overlap between the distributions for concentration decrease/same (Class 0) and concentration increase (Class 1). Similarly, t-SNE visualizations of progenitor node embeddings from a validation subset (Figure 12) and a test subset (Figure 13) did not reveal clear separation between the two classes in the 2D projection. Furthermore, coloring the embeddings by cosmological parameters (Ω_m, σ_8) in these t-SNE plots did not show distinct clusters. This implies that, given the limited training, the learned embeddings do not yet

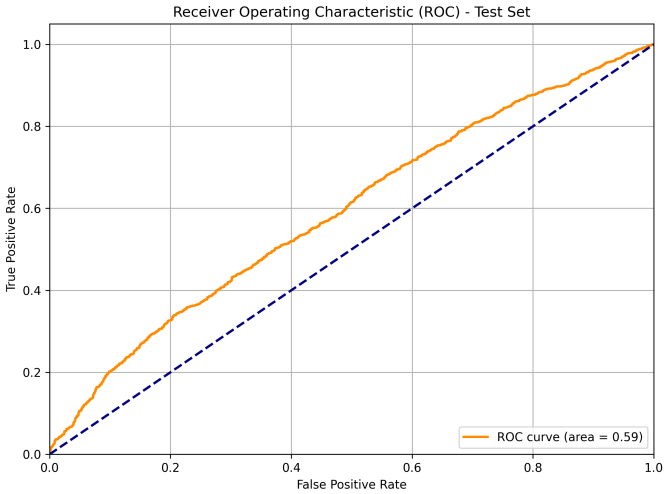

Figure 10. Receiver Operating Characteristic (ROC) curve for the Graph Neural Network (GNN) model on the reduced test set. The area under the curve (AUC) of 0.59 indicates modest discriminative power in predicting the direction of halo concentration change, suggesting performance slightly better than random chance.

perfectly disentangle the factors governing concentration evolution in a low-dimensional, linearly separable manner.

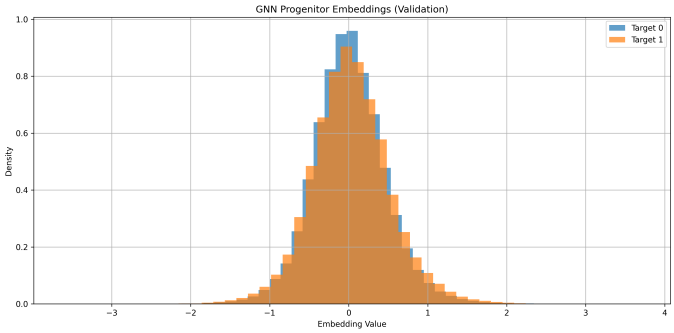

Figure 11. Distribution of GNN progenitor embedding values on the validation set, colored by the target variable (0: concentration decrease/same, 1: concentration increase). The substantial overlap between the class distributions indicates that the learned embeddings do not effectively separate halos based on their future concentration change.

3.4. *Ablation studies*

To further understand the GNN's behavior and the impact of its components, ablation studies were performed. These studies, like the main GNN training, were conducted under highly constrained conditions (10 train, 5 val, 5 test trees; 2 epochs) and thus provide insights into relative effects rather than absolute performance. The results are summarized in Table 3. A visual

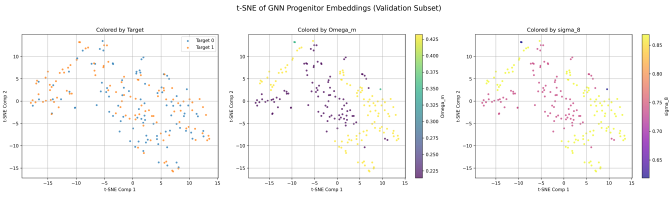

Figure 12. t-SNE projections of GNN progenitor node embeddings from a validation subset. The panels show embeddings colored by the target variable (left), cosmological parameter Ω_m (middle), and σ_8 (right). These plots reveal no clear separation between the two target classes (concentration increase or decrease/same) and no distinct clustering based on cosmological parameters. This suggests that the GNN's learned representations, given the limited training data, do not fully disentangle the factors influencing halo concentration change or strongly encode cosmological dependencies in this 2D space.

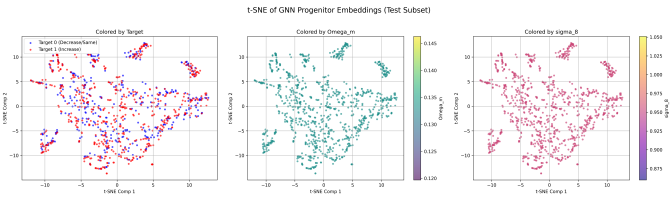

Figure 13. t-SNE visualization of Graph Neural Network (GNN) progenitor node embeddings from a test subset. The left panel shows embeddings colored by the target variable (concentration change direction), indicating no clear separation between classes. The middle and right panels show embeddings colored by cosmological parameters Ω_m and σ_8, respectively, also without distinct clustering. This suggests that the learned embeddings, under limited training conditions, do not perfectly disentangle the factors influencing concentration change, nor do cosmological parameters simply dominate the embedding structure.

comparison of F1-scores for Class 1 across configurations is also provided in Figure 14.

3.4.1. *Reference GNN*

The "Reference GNN" in the ablation context is the same model as described in the previous subsection, utilizing 6 input node features (including cosmological parameters) and a contrastive loss weight of 0.1. It achieved an F1-score of 0.6937 for class 1 (concentration increase) and 0.2295 for class 0 (concentration decrease/same), with a high recall of 0.9019 for class 1. This serves as the baseline for comparison within the ablation study.

3.4.2. *Ablation: No cosmological parameters*

When the cosmological parameters (Ω_m, σ_8) were removed from the GNN's input features (reducing them to 4), the model's overall accuracy slightly increased to 0.5643. However, the F1-score for class 1 decreased to 0.6317 from 0.6937 of the reference GNN. More significantly, the F1-score for class 0 saw a substantial improvement, rising to 0.4668 from 0.2295. Concurrently, the recall for class 1 became more balanced with its precision (recall 0.6784, precision 0.5907), indicating a less biased prediction towards class 1. This suggests that, under the limited training conditions, the inclusion of cosmological parameters might have contributed to the model's bias towards predicting class 1, or their removal forced the model to rely more on other features, leading to a more balanced, albeit slightly less performant for class 1, prediction.

3.4.3. *Ablation: No contrastive loss*

Removing the contrastive loss component (setting $\alpha_{\text{contrastive}} = 0.0$) while retaining cosmological parameters in the input features resulted in the highest F1-score for class 1 (0.7106) among all GNN configurations. However, this came at the cost of virtually eliminating the model's ability to predict class 0, with an F1-score of only 0.0107. The recall for class 1 reached an extreme value of 0.9956, meaning almost all instances were predicted as class 1. This indicates that the contrastive loss, even with a relatively small weight, had a minor regularizing effect, preventing an even more extreme bias towards the 'Increase' class. Its absence allowed the classification loss to aggressively fit the most prevalent or easiest patterns for class 1, essentially ignoring class 0 under the severe data limitations.

3.5. *Summary and interpretation*

This study aimed to predict the binary direction of dark matter halo concentration evolution using both a Random Forest baseline and a Graph Neural Network with contrastive learning, applied to merger tree data from CAMELS-SAM simulations.

The Random Forest baseline, leveraging hand-engineered features including intrinsic halo properties, cosmological parameters, and graph-based environmental features, achieved a weighted F1-score of 0.63. It showed a relatively balanced performance across both classes, with progenitor halo properties being the most important features, aligning with physical intuition about halo evolution.

The GNN model, despite its theoretical suitability for graph-structured data, exhibited lower overall performance (weighted F1-score of 0.485) and a strong bias towards predicting concentration increase (F1-score of 0.69 for increase vs. 0.23 for decrease). This pronounced imbalance, along with the t-SNE visualizations showing no clear class separation, strongly suggests that the

Table 3. GNN Ablation Study Results on Reduced Test Set

Configuration	Input Node Features	$\alpha_{\text{contrastive}}$	Accuracy	F1 (Cls 1)	F1 (Cls 0)	Precision (Cls 1)	Recall (Cls 1)
Reference GNN	6 (Cosmo Incl.)	0.1	0.5616	0.6937	0.2295	0.5636	0.9019
Ablation: No Cosmo	4 (Cosmo Excl.)	0.1	0.5643	0.6317	0.4668	0.5907	0.6784
Ablation: No Contrastive	6 (Cosmo Incl.)	0.0	0.5522	0.7106	0.0107	0.5515	0.9956

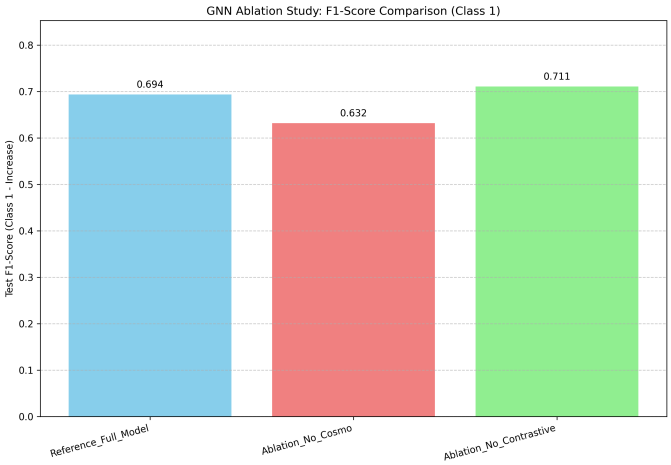

Figure 14. F1-score for predicting halo concentration increase (Class 1) across Graph Neural Network (GNN) ablation configurations. The GNN without contrastive loss achieved the highest F1-score (0.711) for Class 1, while removing cosmological parameters resulted in a lower F1-score (0.632) compared to the reference GNN (0.694). This illustrates the individual impact of cosmological parameters and contrastive loss on the model's ability to predict concentration increases.

GNN was severely underfit. This underfitting is primarily attributable to the drastic reduction in training data and epochs necessitated by computational constraints. The GNN's ability to implicitly learn complex, non-linear relationships from the merger tree structure, which is its core advantage over feature-engineered baselines, was likely not fully realized under these conditions.

The ablation studies provided initial insights into the GNN's component contributions. The inclusion of cosmological parameters appeared to contribute to the model's bias towards predicting concentration increase in the limited-data regime. Removing them led to a more balanced F1-score between classes, albeit at the cost of a lower F1-score for the 'Increase' class. The contrastive learning objective, while intended to improve representation quality, seemed to act as a minor regularizer, preventing an even more extreme class imbalance when present. Its removal led to an almost complete failure to predict concentration decrease, highlighting the fragility of the model under severe data scarcity.

The difficulty in predicting concentration decrease (class 0) by the GNN might stem from several factors. Concentration increase is a common evolutionary path for many halos, especially during active accretion phases. Events leading to concentration decrease, such as major disruptive mergers or strong tidal stripping, might be rarer or have more complex, subtle signatures that are harder to learn from limited data. The GNN, with its capacity to capture the full assembly history, holds promise for discerning these nuances, but this potential was not fully demonstrated in this preliminary study.

Overall, while the GNN framework shows promise for this challenging problem, its current performance is heavily limited by the computational constraints. The results underscore the critical need for full-scale training on the entire CAMELS-SAM dataset to robustly assess the GNN's capabilities and fully leverage its ability to learn from complex, graph-based evolutionary patterns in dark matter halo concentration.

4. CONCLUSIONS

PROBLEM AND SOLUTION OVERVIEW

This paper addressed the challenging binary classification problem of predicting the future direction of dark matter halo concentration evolution (increase or decrease) over cosmic time. Understanding this evolution is critical for accurate galaxy formation models. We proposed a novel approach leveraging Graph Neural Networks (GNNs) with a contrastive learning objective, applied to the hierarchical structure of halo merger trees. This was benchmarked against a robust Random Forest model utilizing a comprehensive set of hand-engineered features.

DATASETS AND METHODS

Our investigation utilized dark matter halo merger trees from the CAMELS-SAM simulations, a rich dataset providing diverse halo assembly histories across varying cosmological parameters. Each halo (node) within a tree was characterized by its logarithmic mass, concentration, Vmax, and scale factor, augmented by global cosmological parameters (Ω_m, σ_8) for its parent simulation. All features were standardized. The binary target variable was defined for progenitor-descendant

transitions, indicating whether the descendant's concentration increased or decreased relative to its progenitor.

The Random Forest baseline was trained on 10 hand-engineered features, combining intrinsic halo properties, temporal differences, cosmological parameters, and graph-based environmental features derived from the merger tree structure. The GNN model employed a multi-layer GraphSAGE encoder to learn embeddings from the merger tree topology and node features, including cosmological parameters. A contrastive learning objective (NT-Xent loss) was integrated to encourage discriminative embedding learning, and a classification head predicted the binary outcome. Crucially, due to severe computational constraints, the GNN was trained on significantly reduced subsets of the data (e.g., 10-50 training trees instead of ∼15,000) and for very few epochs, limiting its ability to fully converge and learn complex patterns.

RESULTS OBTAINED

The Random Forest baseline, trained on a substantial subset of 200 training trees, achieved a weighted F1-score of 0.63 on its validation subset. It demonstrated a relatively balanced predictive capability for both concentration increase (F1-score 0.66) and decrease (F1-score 0.59). Feature importance analysis revealed that intrinsic progenitor halo properties (concentration, mass, Vmax, scale factor) were the primary drivers of its predictions, with cosmological and engineered environmental features contributing secondarily.

In contrast, the GNN model, despite its theoretical suitability for graph-structured data, yielded preliminary performance with a weighted F1-score of 0.485 on its severely reduced test set. A significant finding was the GNN's strong bias towards predicting concentration increase (F1-score 0.69) while performing poorly on concentration decrease (F1-score 0.23). This pronounced class imbalance, coupled with t-SNE visualizations showing no clear class separation in the embedding space, strongly indicates that the GNN was severely underfit due to the drastic limitations in training data and epochs.

Ablation studies, also conducted under highly constrained conditions, provided initial insights into the GNN's components. Removing cosmological parameters from the GNN's input features led to a more balanced F1-score between classes, suggesting that their inclusion under limited training might have contributed to the model's bias towards predicting concentration increase. The contrastive learning objective, even with a small weight, appeared to act as a minor regularizer, preventing an even more extreme class imbalance. Its removal resulted in an almost complete failure to predict concentration decrease, highlighting the fragility of the model when data is scarce.

LEARNINGS AND FUTURE WORK

This study demonstrates that hand-engineered features, particularly intrinsic halo properties, combined with a robust Random Forest classifier, can provide a reasonable baseline for predicting the direction of dark matter halo concentration evolution. The Random Forest's balanced performance underscores the utility of domain-informed feature engineering.

However, the primary takeaway regarding the GNN is its unrealized potential under the current computational limitations. The observed severe underfitting and pronounced class imbalance in the GNN's predictions prevent a definitive assessment of its capabilities. The GNN's inherent strength lies in its ability to implicitly learn complex, non-linear relationships and hierarchical patterns directly from the merger tree structure, a process that requires extensive training on large datasets to fully manifest. The difficulty in predicting concentration decrease might also suggest that these events are either rarer, or their underlying physical mechanisms and signatures within the merger tree are more subtle and require a more thoroughly trained model to discern.

Therefore, the critical next step is to conduct full-scale training of the GNN model on the entire CAMELS-SAM dataset, allowing it to fully leverage its graph processing capabilities and converge to an optimal solution. This would enable a robust assessment of its performance against the feature-engineered baseline and its ability to capture the nuanced evolutionary pathways of dark matter halo concentration, including the more challenging cases of concentration decrease. The GNN framework, with its capacity to learn from the rich, graph-structured assembly histories of halos, remains a highly promising avenue for advancing our understanding and prediction of dark matter halo evolution.

REFERENCES

Anik, M. S. H., Badhe, P., Gampa, R., & Azad, A. 2024, iSpLib: A Library for Accelerating Graph Neural Networks using Auto-tuned Sparse Operations. https://arxiv.org/abs/2403.14853

Bleeker, M., & de Rijke, M. 2022, Do Lessons from Metric Learning Generalize to Image-Caption Retrieval? https://arxiv.org/abs/2202.07474

Bluck, A. F. L., Piotrowska, J. M., Goubert, P., et al. 2025, Dark from light (DfL): Inferring halo properties from luminous tracers with machine learning trained on cosmological simulations. I. Method, proof of concept & preliminary testing, doi: https://doi.org/10.1051/0004-6361/202554702

Borzone, E., Persia, L. D., & Gerard, M. 2025, A Hybrid Supervised and Self-Supervised Graph Neural Network for Edge-Centric Applications. https://arxiv.org/abs/2501.12309

Carr, D. S., Kappannan, S. J., Hutchens, Z. L., et al. 2025, Using Machine Learning to Estimate NUV Magnitudes and Probe Quenching Mechanisms of z=0 Nuggets in the RESOLVE and ECO Surveys. https://arxiv.org/abs/2505.10214

Contardo, G., Trotta, R., Gioia, S. D., Hogg, D. W., & Villaescusa-Navarro, F. 2025, On the effects of parameters on galaxy properties in CAMELS and the predictability of $\Omega_{\rm m}$, doi: https://doi.org/10.3847/1538-4357/addd08

de Amorim, L. B. V., Cavalcanti, G. D. C., & Cruz, R. M. O. 2022, The choice of scaling technique matters for classification performance, doi: https://doi.org/10.1016/j.asoc.2022.109924

Dooley, G. A., Griffen, B. F., Zukin, P., et al. 2014, The Effects of Varying Cosmological Parameters on Halo Substructure, doi: https://doi.org/10.1088/0004-637X/786/1/50

Dou, J., Peng, Y., Gu, Q., et al. 2025, The critical role of dark matter halos in driving star formation. https://arxiv.org/abs/2503.04243

Duan, H., Cheng, Y., Yu, J., Liu, Y., & Li, X. 2025, Can Large Language Models Act as Ensembler for Multi-GNNs? https://arxiv.org/abs/2410.16822

Fahim, A., Murphy, A., & Fyshe, A. 2024, It's Not a Modality Gap: Characterizing and Addressing the Contrastive Gap. https://arxiv.org/abs/2405.18570

Fallah, F. 2023, Hierarchical Quadratic Random Forest Classifier. https://arxiv.org/abs/2306.01893

Gómez, J. S., Padilla, N. D., Helly, J. C., et al. 2021, Halo Merger Tree Comparison: Impact on Galaxy Formation Models, doi: https://doi.org/10.1093/mnras/stab3661

He, X., Wang, Y., Fan, W., et al. 2025, Mamba-Based Graph Convolutional Networks: Tackling Over-smoothing with Selective State Space. https://arxiv.org/abs/2501.15461

Ho, Y., & Wookey, S. 2020, The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling, doi: https://doi.org/10.1109/ACCESS.2019.2962617

Huang, N., Stiskalek, R., Lee, J.-Y., et al. 2025, CosmoBench: A Multiscale, Multiview, Multitask Cosmology Benchmark for Geometric Deep Learning. https://arxiv.org/abs/2507.03707

Ishiyama, T., Prada, F., & Klypin, A. A. 2025, Evolution of clustering in cosmological models with time-varying dark energy. https://arxiv.org/abs/2503.19352

Islam, N. 2024, DTization: A New Method for Supervised Feature Scaling. https://arxiv.org/abs/2404.17937

Jahin, M. A., Soudeep, S., Mridha, M. F., Monowar, M. M., & Hamid, M. A. 2025, Physics-Informed Graph Neural Networks for Transverse Momentum Estimation in CMS Trigger Systems. https://arxiv.org/abs/2507.19205

Javaloy, A., Sanchez-Martin, P., Levi, A., & Valera, I. 2023, Learnable Graph Convolutional Attention Networks. https://arxiv.org/abs/2211.11853

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

Johnson, T., Benson, A. J., & Grin, D. 2021, A Random Walk Model for Dark Matter Halo Concentrations, doi: https://doi.org/10.3847/1538-4357/abd563

Larson, A. J., Wu, J. F., & Jones, C. 2024, Predicting dark matter halo masses from simulated galaxy images and environments. https://arxiv.org/abs/2407.13735

Lee, J.-Y., & Villaescusa-Navarro, F. 2025, Cosmology with Topological Deep Learning, doi: https://doi.org/10.3847/1538-4357/ade806

Louppe, G. 2015, Understanding Random Forests: From Theory to Practice. https://arxiv.org/abs/1407.7502

Lovell, C. C., Starkenburg, T., Ho, M., et al. 2024, Learning the Universe: Cosmological and Astrophysical Parameter Inference with Galaxy Luminosity Functions and Colours. https://arxiv.org/abs/2411.13960

Maturo, F., & Porreca, A. 2024, Augmented Functional Random Forests: Classifier Construction and Unbiased Functional Principal Components Importance through Ad-Hoc Conditional Permutations. https://arxiv.org/abs/2408.13179

Nguyen, T., Modi, C., Mishra-Sharma, S., Yung, L. Y. A., & Somerville, R. S. 2025, Emulating Dark Matter Halo Merger Trees with Graph Generative Models. https://arxiv.org/abs/2507.10652

Okoli, C. 2017, Dark matter halo concentrations: a short review. https://arxiv.org/abs/1711.05277

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

Pinheiro, J. M. H., de Oliveira, S. V. B., Silva, T. H. S., et al. 2025, The Impact of Feature Scaling In Machine Learning: Effects on Regression and Classification Tasks. https://arxiv.org/abs/2506.08274

Plettenberg, P., Alcalde, A., Sick, B., & Thomas, J. M. 2025, Graph Neural Networks for Automatic Addition of Optimizing Components in Printed Circuit Board Schematics. https://arxiv.org/abs/2506.10577

Qin, Y., Fasy, B. T., Wenk, C., & Summa, B. 2024, Rapid and Precise Topological Comparison with Merge Tree Neural Networks. https://arxiv.org/abs/2404.05879

Ramakrishnan, S., & Velmani, P. 2022, Properties beyond mass for unresolved haloes across redshift and cosmology using correlations with local halo environment, doi: https://doi.org/10.1093/mnras/stac2605

Shi, R., Wang, W., Li, Z., et al. 2022, A machine learning approach to infer the accreted stellar mass fractions of central galaxies in the TNG100 simulation, doi: https://doi.org/10.1093/mnras/stac1541

Srivastava, A., Cui, W., de Andres, D., et al. 2025, Predicting Halo Formation Time Using Machine Learning, doi: https://doi.org/10.1051/0004-6361/202453165

Steck, H., Ekanadham, C., & Kallus, N. 2024, Is Cosine-Similarity of Embeddings Really About Similarity?, doi: https://doi.org/10.1145/3589335.3651526

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

Tanis, J. H., Giannella, C., & Mariano, A. V. 2024, Introduction to Graph Neural Networks: A Starting Point for Machine Learning Engineers. https://arxiv.org/abs/2412.19419

Wali, R. 2022, Xtreme Margin: A Tunable Loss Function for Binary Classification Problems. https://arxiv.org/abs/2211.00176

Wang, K., Mao, Y.-Y., Zentner, A. R., et al. 2020, Concentrations of Dark Haloes Emerge from Their Merger Histories, doi: https://doi.org/10.1093/mnras/staa2733

Wang, X., & Shen, H.-W. 2024, GNNInterpreter: A Probabilistic Generative Model-Level Explanation for Graph Neural Networks. https://arxiv.org/abs/2209.07924

Wang, X., & Zhang, M. 2023, PyTorch Geometric High Order: A Unified Library for High Order Graph Neural Network. https://arxiv.org/abs/2311.16670

Wei, X., & Zhang, B. 2025, A Generative Graph Contrastive Learning Model with Global Signal. https://arxiv.org/abs/2504.18148

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

Yuan, Y., Wang, W., & Pang, W. 2021, Which Hyperparameters to Optimise? An Investigation of Evolutionary Hyperparameter Optimisation in Graph Neural Network For Molecular Property Prediction. https://arxiv.org/abs/2104.06046

Yung, L. Y. A., Somerville, R. S., Nguyen, T., et al. 2024, Characterising ultra-high-redshift dark matter halo demographics and assembly histories with the GUREFT simulations. https://arxiv.org/abs/2309.14408

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

Zhu, Y., Ai, X., Vorobeychik, Y., & Zhou, K. 2025, Robust Graph Contrastive Learning with Information Restoration. https://arxiv.org/abs/2307.12555

Ågren, W. 2022, The NT-Xent loss upper bound. https://arxiv.org/abs/2205.03169