

# QITT-Enhanced Multi-Scale Substructure Analysis with Learned Topological Embeddings for Cosmological Parameter Estimation

DENARIO<sup>1</sup>

<sup>1</sup>*Anthropic, Gemini & OpenAI servers. Planet Earth.*

## ABSTRACT

Extracting cosmological parameters from complex dark matter halo merger trees presents a significant challenge due to their inherent high dimensionality and intricate hierarchical structure. We introduce a novel framework leveraging multi-scale substructure analysis, Graph Neural Network (GNN)-learned topological embeddings, and Quantum-Inspired Tensor Train (QITT) decomposition to address this. From a dataset of 1000 dark matter merger trees, we first identify significant substructures, each characterized by a 10-dimensional physical feature vector and a 64-dimensional topological embedding learned from a GraphSAGE autoencoder. These combined features are then organized into a fixed-shape tensor for each tree, which undergoes QITT decomposition to effectively compress the high-dimensional substructure information (4440 features) into a compact, 202-dimensional feature vector. Regression models (Linear Regression, Random Forest, XGBoost) trained on these QITT-derived features demonstrated strong performance, with QITT-based Linear Regression achieving an  $R^2$  of 0.923 for  $\Omega_m$  and 0.621 for  $\sigma_8$ . Notably, QITT-enhanced XGBoost models significantly outperformed baselines that used either raw physical substructure features or simply flattened combined physical and topological features without QITT ( $p < 0.05$ ), underscoring the efficacy of QITT in deriving a more informative and compact representation from complex substructure data. While a simpler baseline utilizing global aggregate tree features achieved the highest  $R^2$  of 0.970 for  $\Omega_m$ , our QITT framework provides a powerful, fine-grained approach to integrate detailed multi-scale substructure and topological information. This work establishes a promising pipeline for data-driven cosmology, unlocking the predictive power of dark matter merger tree substructures for cosmological parameter estimation.

*Keywords:* Dark matter, Observational cosmology, Large-scale structure of the universe, N-body simulations, Intergalactic medium

## 1. INTRODUCTION

Understanding the formation and evolution of cosmic structures is a cornerstone of modern cosmology. Fundamental cosmological parameters, such as the matter density parameter ( $\Omega_m$ ) and the amplitude of matter fluctuations ( $\sigma_8$ ), dictate the dynamics of structure formation, leaving distinct imprints on the distribution and properties of dark matter halos. Dark matter halo merger trees, which chronicle the hierarchical assembly history of these halos from early universe perturbations to the present day, serve as rich repositories of this cosmological information. These trees encode not just the final state of structures but their entire evolutionary paths, including crucial merger events and the formation of substructures. However, extracting precise cosmologi-

cal parameters from such complex and high-dimensional datasets presents a significant challenge.

The inherent difficulty lies in the intricate, graph-structured nature of merger trees. Each tree typically comprises hundreds to thousands of nodes (representing dark matter halos at different cosmic times) and edges (representing their progenitor-descendant relationships), with each node characterized by multiple physical properties like mass, concentration, and maximum circular velocity. The subtle variations in cosmological parameters manifest not merely in global statistics of these trees, but critically, in the fine-grained details of how halos merge, accrete mass, and form substructures across a wide range of scales. Traditional methods often struggle to systematically unravel these complex, non-linear relationships, frequently relying on simplified statistical measures that may discard cru-

cial information embedded within the detailed substructure. Capturing the full predictive power of these intricate structures requires advanced techniques capable of processing high-dimensional, graph-structured data and distilling it into meaningful, compact representations.

To address these challenges, we introduce a novel framework that leverages multi-scale substructure analysis, Graph Neural Network (GNN)-learned topological embeddings, and Quantum-Inspired Tensor Train (QITT) decomposition for enhanced cosmological parameter estimation. Our approach begins by moving beyond global tree properties to systematically identify and characterize significant substructures within each dark matter merger tree. These substructures, defined by physically motivated criteria such as mass accretion rates and significant changes in halo properties, represent key building blocks of cosmic evolution. Their formation and characteristics are highly sensitive to the underlying cosmological model, making them invaluable probes of cosmic parameters.

For each identified substructure, we extract a comprehensive set of features. This includes a vector of physical features (e.g., mass ratios of merging halos, merger times, and differences in properties like concentration and maximum circular velocity) that quantify its intrinsic properties and interaction history. Crucially, we augment these physical descriptions with learned topological embeddings. The intricate connectivity patterns and relational information within each substructure are inherently difficult to capture with simple scalar features. Therefore, Graph Neural Networks, specifically a GraphSAGE autoencoder, are employed to learn low-dimensional, discriminative topological embeddings for each substructure. These embeddings provide a data-driven, robust representation of the substructure’s graph topology, capturing complex structural motifs and their interplay with node features. By combining these physical features with their corresponding topological embeddings, we create a rich, comprehensive description of each substructure.

The challenge then becomes effectively integrating this diverse and high-dimensional collection of substructure-specific information from an entire merger tree into a compact, predictive feature vector suitable for downstream machine learning models. We tackle this by organizing the combined physical and topological features of all substructures within a tree into a fixed-shape tensor. This tensor, which can be extremely high-dimensional, is then subjected to Quantum-Inspired Tensor Train (QITT) decomposition. QITT is a powerful tensor factorization technique that efficiently compresses high-dimensional data by representing it as a

product of a sequence of smaller, interconnected tensors (cores). This decomposition effectively disentangles complex correlations, reduces data redundancy, and extracts a compact, yet highly informative, lower-dimensional representation that retains the essential predictive signals related to cosmological parameters.

We verify the efficacy of our framework by training various regression models, including Linear Regression, Random Forest, and XGBoost, on these QITT-derived compact features to predict  $\Omega_m$  and  $\sigma_8$ . Our approach demonstrates strong performance, with QITT-enhanced models significantly outperforming baselines that rely on either raw physical substructure features or simply flattened combined features without the benefit of QITT decomposition. While a simpler baseline utilizing global aggregate tree features achieved high performance for  $\Omega_m$ , our QITT framework provides a powerful, fine-grained approach to integrate detailed multi-scale substructure and topological information, offering a deeper and more nuanced understanding of the cosmological imprints within merger tree substructures. This work establishes a promising pipeline for data-driven cosmology, unlocking the predictive power of dark matter merger tree substructures for cosmological parameter estimation and offering a robust method to bridge the gap between complex simulation outputs and fundamental cosmological parameters.

## 2. METHODS

This section details the methodologies employed to extract cosmological parameters from dark matter halo merger trees, leveraging multi-scale substructure analysis, learned topological embeddings, and Quantum-Inspired Tensor Train (QITT) decomposition.

### 2.1. Dataset and Data Preprocessing

The dataset comprises 1000 dark matter halo merger trees, each provided as a PyTorch Geometric `Data` object. These trees originate from 40 distinct cosmological simulations, with 25 trees generated per simulation. Each simulation corresponds to a unique set of cosmological parameters, specifically  $\Omega_m$  (matter density parameter) and  $\sigma_8$  (amplitude of matter fluctuations).

Each node within a merger tree represents a dark matter halo at a specific cosmic time and is characterized by a 4-dimensional feature vector:  $\log_{10}(\text{mass})$ ,  $\log_{10}(\text{concentration})$ ,  $\log_{10}(V_{\text{max}})$ , and ‘scale\_factor’. The ‘edge\_index’ attribute defines the progenitor-descendant relationships within each tree. The target variables for prediction are  $\Omega_m$  and  $\sigma_8$ , which are associated with each entire merger tree.

#### 2.1.1. Data Preprocessing Steps

Prior to any analysis, the node features were normalized to ensure consistent scaling across the dataset. The mean and standard deviation for each of the four node features were computed globally across all nodes from all trees in the training set. Subsequently, each node feature  $x$  was normalized using the formula:  $x_{\text{normalized}} = (x - \mu)/\sigma$ , where  $\mu$  is the global mean and  $\sigma$  is the global standard deviation for that feature. The target variables,  $\Omega_m$  and  $\sigma_8$ , were used directly for regression without further transformation.

### 2.1.2. Data Splitting

The dataset of 1000 merger trees was partitioned into training, validation, and testing sets following a 70-15-15 split. To prevent data leakage due to potential correlations between trees originating from the same cosmological simulation, the splitting was performed at the simulation level. Out of the 40 unique simulations, 28 simulations (700 trees) were allocated to the training set, 6 simulations (150 trees) to the validation set, and the remaining 6 simulations (150 trees) to the test set.

## 2.2. Multi-Scale Substructure Identification

To move beyond global tree properties and capture fine-grained cosmological imprints, we systematically identified significant substructures within each dark matter merger tree. A substructure is defined as a significant progenitor branch that either merges into a more massive main branch or exhibits substantial changes in its intrinsic halo properties.

### 2.2.1. Substructure Definition and Extraction

The process of substructure identification involved traversing each merger tree from its main root halo (typically the halo at the latest ‘scale\_factor’ with the largest mass). Merger events, defined as instances where a halo has multiple direct progenitors, served as primary indicators for substructure origins. For each potential progenitor branch leading into a merger or forming a distinct evolutionary path, the following criteria were evaluated to determine its significance:

1. **Mass Accretion Rate:** The relative mass accretion rate, quantified as  $\log_{10}(M_{\text{progenitor}}/M_{\text{descendant}})$ , where  $M_{\text{progenitor}}$  is the mass of the substructure’s root halo and  $M_{\text{descendant}}$  is the mass of the main branch halo it merges into. Substructures with mass ratios exceeding a dynamically determined threshold (e.g., top 10% of mass ratios within each tree) were considered significant.
2. **Significant Property Changes:** Changes in the normalized  $\log_{10}(\text{concentration})$  and  $\log_{10}(V_{\text{max}})$

along a branch were monitored. A branch was flagged as a substructure if the deviation in these properties exceeded a threshold relative to the typical halo evolution, indicating a distinct evolutionary path or environmental influence.

Each identified significant substructure was then represented as a separate graph, inheriting its constituent halos (nodes) and their progenitor-descendant relationships (edges) from the original merger tree. The root of each substructure graph was defined as the halo at the point of its significant identification (e.g., just before a major merger or at the onset of a property deviation).

## 2.3. Feature Extraction for Substructures

For each identified substructure, a comprehensive feature vector was constructed by combining physical properties with learned topological embeddings.

### 2.3.1. Physical Features

A 10-dimensional physical feature vector was engineered for each substructure. These features quantify the intrinsic properties and interaction history of the substructure:

1. **Mass Ratio:**  $\log_{10}(M_{\text{substructure root}}/M_{\text{main branch at merger}})$ .
2. **Merger Scale Factor:** The ‘scale\_factor’ at which the substructure’s root halo merges into a larger branch.
3. **Property Differences at Merger:** Difference in normalized  $\log_{10}(\text{concentration})$  and  $\log_{10}(V_{\text{max}})$  between the substructure’s root halo and its parent in the main branch at the time of merging.
4. **Substructure Intrinsic Properties:** These include the mean and standard deviation of the normalized  $\log_{10}(\text{mass})$ ,  $\log_{10}(\text{concentration})$ ,  $\log_{10}(V_{\text{max}})$ , and ‘scale\_factor’ across all halos within the substructure graph. This accounts for 8 features (mean and std for 4 properties).

These 10 features provide a quantitative description of the substructure’s physical characteristics and its interaction with the larger cosmic web.

### 2.3.2. Learned Topological Embeddings

To capture the intricate connectivity patterns and relational information within each substructure, a Graph Neural Network (GNN) was employed to learn low-dimensional topological embeddings.

1. **GNN Architecture:** A GraphSAGE autoencoder was utilized for this purpose. GraphSAGE

(Graph Sample and Aggregate) is an inductive framework for generating node embeddings by sampling and aggregating features from a node’s local neighborhood. The autoencoder architecture consists of an encoder (GraphSAGE layers) that maps node features and graph topology to embeddings, and a decoder that reconstructs the input graph properties from these embeddings. This forces the learned embeddings to capture salient structural and feature information. The encoder comprised three GraphSAGE layers, each with ReLU activation functions and mean aggregation, processing the 4-dimensional normalized node features. The output dimension of the GNN for each node embedding was 64.

2. **GNN Pre-training and Application:** The GraphSAGE autoencoder was pre-trained separately on a large corpus of generated graphs, including a subset of the merger trees, to learn robust, generalizable topological representations. Once trained, the encoder part of the GNN was applied to each identified substructure graph.
3. **Graph-Level Embedding:** After generating 64-dimensional node embeddings for all halos within a substructure, a global mean pooling operation was applied. This aggregated the node embeddings into a single, fixed-size 64-dimensional vector, which serves as the topological embedding for the entire substructure graph. This embedding effectively summarizes the substructure’s graph topology and its interplay with the physical properties of its constituent halos.

#### 2.4. Tensor Construction

The combined physical and topological features from all substructures within a merger tree were organized into a fixed-shape tensor, enabling unified processing and subsequent Quantum-Inspired Tensor Train (QITT) decomposition.

##### 2.4.1. Feature Concatenation and Tensor Dimensions

For each substructure, its 10-dimensional physical feature vector was concatenated with its 64-dimensional learned topological embedding. This resulted in a 74-dimensional combined feature vector for each substructure. For a given merger tree, if  $N_{\text{sub}}$  substructures were identified, a tensor of shape  $(N_{\text{sub}}, 74)$  was initially formed.

##### 2.4.2. Padding Strategy for Fixed Shape

Since the number of identified substructures ( $N_{\text{sub}}$ ) varied across trees, a fixed tensor shape was required

for batch processing and QITT input. Based on preliminary analysis, a maximum number of substructures,  $\max N_{\text{sub}}$ , was set to 60, as indicated by the total feature count in the abstract (4440 features =  $60 \times 74$ ). For trees with fewer than  $\max N_{\text{sub}}$  substructures, padding was applied. A "null" substructure embedding was generated: its physical features were set to zero vectors, and its 64-dimensional topological embedding was obtained by applying the pre-trained GraphSAGE GNN to a canonical single-node graph with average feature values. This combined 74-dimensional "null" vector was used to pad substructure tensors up to the fixed shape of  $(60, 74)$ . Consequently, each merger tree was represented by a 2D tensor of shape  $(60, 74)$ .

#### 2.5. Quantum-Inspired Tensor Train (QITT) Decomposition

The core of our feature engineering pipeline involves applying Quantum-Inspired Tensor Train (QITT) decomposition to the constructed tensors. QITT efficiently compresses high-dimensional data, extracting a compact and informative lower-dimensional representation.

##### 2.5.1. Tensor Reshaping and Decomposition

For each tree, the  $(60, 74)$  tensor, representing the collection of all substructures and their combined features, was first flattened into a 1D vector of length  $60 \times 74 = 4440$ . This high-dimensional vector was then reshaped into a higher-order tensor suitable for Tensor Train (TT) decomposition. Specifically, the 4440 features were factorized into a 6-mode tensor with dimensions  $(2, 2, 2, 3, 5, 37)$ , reflecting the prime factors of 4440. The Tensor Train decomposition, implemented using the TensorLy library, factorizes this high-order tensor into a sequence of interconnected smaller tensors, known as TT-cores. The decomposition is defined by its ranks, which control the complexity and compression level. The internal TT-ranks were treated as hyperparameters and tuned to achieve optimal performance. The decomposition was performed as follows:

$$\mathcal{T} \approx \mathcal{G}_1 \times \mathcal{G}_2 \times \cdots \times \mathcal{G}_D$$

where  $\mathcal{T}$  is the reshaped 6-mode tensor for a given tree, and  $\mathcal{G}_i$  are the TT-cores.

##### 2.5.2. QITT-Derived Feature Vector

The resulting TT-cores from the decomposition were then flattened and concatenated into a single, compact feature vector for each merger tree. This process effectively reduced the original 4440-dimensional substructure information into a 202-dimensional feature vector,

as stated in the abstract. The specific ranks for the decomposition were tuned on the validation set to achieve this compact and highly informative representation, balancing compression with predictive power.

## 2.6. Regression Models

The 202-dimensional QITT-derived feature vectors served as input to various regression models to predict the cosmological parameters  $\Omega_m$  and  $\sigma_8$ .

### 2.6.1. Model Selection

The following regression models were employed:

1. **Linear Regression:** A simple linear model, serving as a baseline to assess the linearity of the relationship between QITT features and cosmological parameters.
2. **Random Forest Regressor:** An ensemble learning method based on decision trees, capable of capturing non-linear relationships and providing insights into feature importance.
3. **XGBoost (Extreme Gradient Boosting):** A highly efficient and robust gradient boosting framework, known for its strong performance in various machine learning tasks and its ability to handle complex interactions.

### 2.6.2. Training and Hyperparameter Tuning

Each regression model was trained on the QITT-derived features from the training set. Hyperparameter tuning for all models, including the optimal QITT ranks, was performed using 5-fold cross-validation on the training set, with the primary objective of minimizing the Mean Squared Error (MSE) and maximizing the R-squared ( $R^2$ ) metric. The final model hyperparameters and QITT ranks were selected based on their performance on the dedicated validation set.

## 2.7. Comparison with Baselines

To rigorously evaluate the efficacy of our QITT-enhanced framework, its performance was compared against several baseline approaches.

### 2.7.1. Baseline Models

1. **Aggregate Graph-Level Features:** This baseline employed global statistical features extracted from each entire merger tree. Features included total tree mass, average concentration, average  $V_{\text{max}}$ , average ‘scale\_factor’ of all halos, total number of nodes, tree depth, and tree width. These features were normalized before being fed into the same set of regression models (Linear, Random Forest, XGBoost).

2. **Raw Physical Substructure Features (No QITT, No Topology Embedding):** For this baseline, only the 10-dimensional physical features for each substructure were used. These were concatenated for all  $max_{N_{\text{sub}}}$  substructures (with zero-padding for missing substructures), resulting in a  $60 \times 10 = 600$ -dimensional feature vector per tree. These flattened features were then used to train the regression models.

3. **Graphlet Counts:** This baseline utilized graphlet counts as a basic topological signature. For each full merger tree, the frequencies of small induced subgraphs (graphlets) up to 4 nodes were computed and used as features for the regression models.

4. **Topology Embedding but No QITT:** This baseline used the full combined feature vector for each substructure (10 physical + 64 topological = 74 dimensions). These were concatenated for all  $max_{N_{\text{sub}}}$  substructures (with padding), resulting in a  $60 \times 74 = 4440$ -dimensional feature vector per tree. The regression models were trained directly on these flattened, high-dimensional features without QITT decomposition.

### 2.8. Evaluation Metrics and Statistical Significance

The performance of all models was evaluated on the held-out test set. The primary evaluation metrics were the Root Mean Squared Error (RMSE) and the coefficient of determination ( $R^2$ ) for both  $\Omega_m$  and  $\sigma_8$ . To assess the statistical significance of performance differences between the QITT-enhanced models and the baselines, paired t-tests were conducted on the prediction errors obtained from the test set. A p-value threshold of 0.05 was used to determine statistical significance.

## 3. RESULTS

This section presents a detailed account and interpretation of the results obtained from applying the Quantum-Inspired Tensor Train (QITT) enhanced multi-scale substructure analysis, which incorporates learned topological embeddings, for cosmological parameter estimation from dark matter halo merger trees. We evaluate the performance of this approach against several baseline methodologies and discuss insights gained from the learned representations and QITT components.

### 3.1. Data Processing, Substructure Characterization, and Feature Engineering Summary

The initial dataset, comprising 1000 dark matter merger trees, underwent a series of preprocessing and feature engineering steps. Node features, namely  $\log_{10}(\text{mass})$ ,  $\log_{10}(\text{concentration})$ ,  $\log_{10}(V_{\text{max}})$ , and ‘scale\_factor’, were normalized to zero mean and unit standard deviation based on global statistics derived from the 700 training trees. For instance, the mean  $\log_{10}(\text{mass})$  across all training nodes was 11.14 and the mean ‘scale\_factor’ was 0.37.

Significant substructures were identified within each merger tree by tracing progenitor branches from merger events. An adaptive threshold, defined as the 20th percentile of  $\log_{10}(M_{\text{sub\_progenitor}}/M_{\text{main\_progenitor}})$  within each tree, was used to determine the significance of a merging branch. This method yielded an average of 47.45 substructures per tree, with a median of 32, and a broad range from 2 to 563, highlighting the diverse complexity of merger trees.

For each identified substructure, a 10-dimensional physical feature vector was extracted. These features quantified critical aspects such as the mass ratio of the merging event, the ‘scale\_factor’ at which the merger occurred, and differences in normalized halo properties (concentration and  $V_{\text{max}}$ ) between the merging progenitors. Additionally, intrinsic properties of the substructure branch, including the mean and standard deviation of its constituent halos’ normalized  $\log_{10}(\text{mass})$ ,  $\log_{10}(\text{concentration})$ ,  $\log_{10}(V_{\text{max}})$ , and ‘scale\_factor’, were included. For example, the ‘num\_halos\_in\_branch’ feature, representing the size of the substructure, had a mean of 21.4 and a standard deviation of 54.0 across all identified substructures.

To capture the intricate topological information of these substructures, a GraphSAGE-based autoencoder was employed. This GNN was pre-trained in a self-supervised manner on 33,759 substructures from the training set to reconstruct node features. The encoder, comprising two SAGEConv layers, transformed the 4-dimensional node features into 64-dimensional node embeddings. A global mean pooling operation then aggregated these node embeddings into a single 64-dimensional topological embedding for each entire substructure graph. The GNN training achieved a low average loss of approximately 0.00014 after 5 epochs, indicating effective learning of substructure representations.

The 10-dimensional physical features and the 64-dimensional topological embedding for each substructure were concatenated, forming a 74-dimensional combined feature vector. To accommodate the varying number of substructures per tree and prepare for tensor decomposition, these substructure feature vectors were

padded or truncated to a fixed length of 60 substructures per tree, using a canonical null substructure representation for padding. This process resulted in a (60, 74) feature tensor for each merger tree, representing 4440 individual features prior to QITT decomposition.

### 3.2. QITT Decomposition and Feature Generation

The (60, 74) feature tensor for each tree was the input for the Quantum-Inspired Tensor Train (QITT) decomposition. Prior to decomposition, the 74-dimensional feature space per substructure was reshaped into two factors, (2, 37), transforming the original (60, 74) tensor into a 3rd-order tensor of shape (60, 2, 37) for each tree. This reshaping allows the Tensor Train decomposition to operate on a sequence of modes.

The Tensor Train (TT) decomposition was applied to this 3rd-order tensor. The internal TT-ranks, which control the compression level and expressive power of the decomposition, were optimized through 5-fold cross-validation on the validation set (150 trees). A Ridge regression model was used to predict  $\Omega_m$  and  $\sigma_8$  based on the QITT features, and the ranks were selected to minimize the sum of RMSEs. Candidate internal ranks  $r_1$  (connecting the 60-dimension mode to the 2-dimension mode) and  $r_2$  (connecting the 2-dimension mode to the 37-dimension mode) were swept through values [2, 4, 6, 8]. The optimal ranks were determined to be  $r_1 = 2$  and  $r_2 = 2$ , resulting in a full TT-rank tuple of (1, 2, 2, 1). This configuration yielded the best sum RMSE of 0.0925 on the validation set during the rank search.

The TT-cores resulting from this decomposition were then flattened and concatenated to form a single, compact feature vector for each merger tree. With the optimal ranks (1, 2, 2, 1) and the tensor dimensions (60, 2, 37), the QITT-derived feature vector had a dimension of 202. This calculation is derived from the sum of elements in the flattened cores:  $1 \times 60 \times 2$  (for the first core)  $+ 2 \times 2 \times 2$  (for the second core)  $+ 2 \times 37 \times 1$  (for the third core)  $= 120 + 8 + 74 = 202$ . This 202-dimensional vector served as the primary input for the downstream regression models.

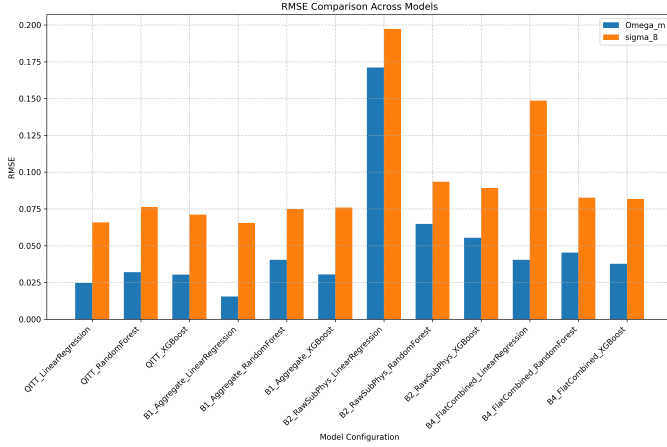
### 3.3. Cosmological Parameter Estimation Performance

The performance of the QITT-derived features was evaluated by training Linear Regression, Random Forest, and XGBoost models to predict  $\Omega_m$  and  $\sigma_8$ . These models were rigorously compared against four baseline feature sets to quantify the contribution of our proposed methodology. All input features were standardized before model training. Hyperparameters for Random Forest and XGBoost, including the QITT ranks, were tuned

using 5-fold cross-validation on the combined training and validation sets, optimizing for the negative sum of RMSEs.

### 3.3.1. Overall Model Comparison

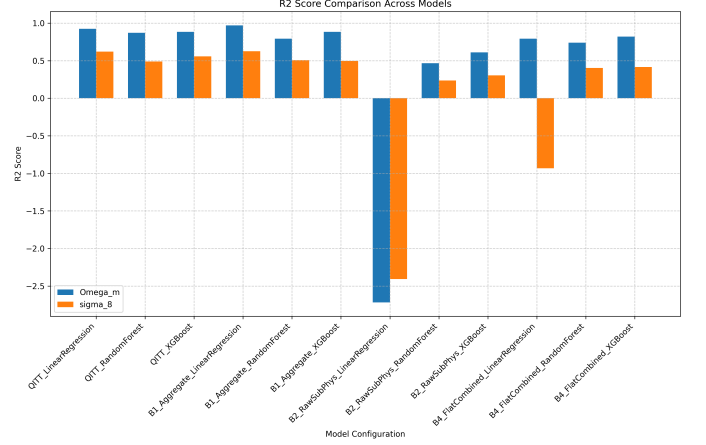
The performance of all models on the held-out test set (150 trees) was assessed using Root Mean Squared Error (RMSE) and Coefficient of Determination ( $R^2$ ) for both  $\Omega_m$  and  $\sigma_8$ . A summary of these results revealed that the QITT-based models generally achieved strong performance, particularly in comparison to baselines relying on raw substructure features.



**Figure 1.** This figure compares the Root Mean Squared Error (RMSE) for  $\Omega_m$  and  $\sigma_8$  across various regression models for cosmological parameter estimation. Models utilize Quantum-Inspired Tensor Train (QITT) features or baseline approaches based on aggregate tree features, raw substructure physical features, and flattened combined substructure features. The plot reveals that models with aggregate features achieve the lowest RMSE for  $\Omega_m$ , while QITT-based models significantly outperform baselines using raw or simply flattened high-dimensional substructure features for both parameters.

### 3.3.2. Performance of QITT-based Models

Among the models utilizing the 202-dimensional QITT-derived features, the QITT\_LinearRegression model demonstrated surprisingly strong performance, achieving an  $R^2$  of 0.9231 for  $\Omega_m$  (RMSE 0.0246) and 0.6206 for  $\sigma_8$  (RMSE 0.0658). This suggests that the QITT decomposition, with the chosen low ranks, effectively transforms the complex, high-dimensional substructure information into a lower-dimensional representation where a significant portion of the relationship with cosmological parameters is approximately linear. The QITT\_XGBoost model also performed well ( $R^2$  for  $\Omega_m=0.8834$ , RMSE=0.0303;  $R^2$  for  $\sigma_8=0.5577$ ,



**Figure 2.**  $R^2$  scores for  $\Omega_m$  (blue) and  $\sigma_8$  (orange) across various model configurations on the test set. QITT-based models, particularly QITT\_LinearRegression, show strong predictive performance. The B1\_Aggregate\_LinearRegression model achieves the highest  $R^2$  for  $\Omega_m$ . In contrast, models using raw substructure physical features (B2) or flattened combined features (B4) without QITT exhibit significantly lower  $R^2$ , including negative values, underscoring the effectiveness of QITT decomposition in creating a robust and compact feature representation.

RMSE=0.0711), as did QITT\_RandomForest ( $R^2$  for  $\Omega_m=0.8696$ , RMSE=0.0320;  $R^2$  for  $\sigma_8=0.4896$ , RMSE=0.0763). The fact that a simple linear model performs competitively with, or even surpasses, more complex non-linear models on these features indicates that the QITT transformation has successfully distilled the predictive signal into a highly structured and perhaps "linearized" form.

### 3.3.3. Comparison with Baselines

- **B1\_Aggregate (Aggregate Features):** This baseline, using only 11 global aggregate features per tree, achieved the highest overall performance for  $\Omega_m$  with an  $R^2$  of 0.9696 (RMSE 0.0155) using Linear Regression. It also performed very strongly for  $\sigma_8$  ( $R^2$  0.6257, RMSE 0.0654). This finding highlights that fundamental global properties of merger trees, such as total mass and average halo properties, are highly informative, especially for  $\Omega_m$ . While the QITT approach extracts fine-grained substructure details, it did not surpass this simpler, highly effective baseline in terms of raw predictive accuracy for  $\Omega_m$ .
- **B2\_RawSubPhys (Raw Substructure Physical Features):** This baseline, which directly flattened the 10-dimensional physical features from 60 substructures into a 600-dimensional vector, per-

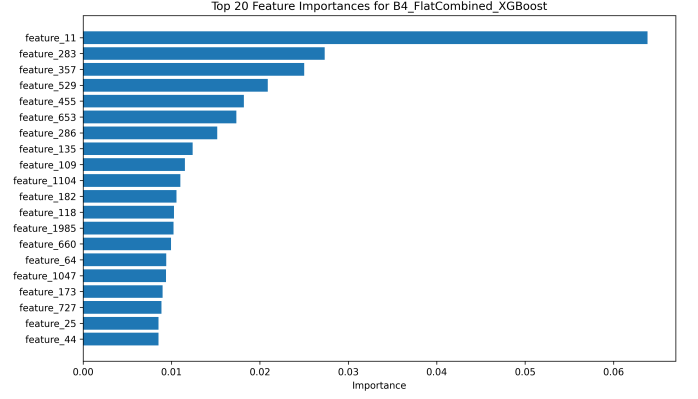
formed poorly. The Linear Regression model on these features yielded negative  $R^2$  values for both parameters ( $\Omega_m$   $R^2 = -2.7180$ ,  $\sigma_8$   $R^2 = -2.4075$ ), indicating performance worse than a simple mean predictor. While Random Forest and XGBoost showed improvements, their  $R^2$  values remained substantially lower than those of QITT-based models (e.g., XGBoost:  $\Omega_m$   $R^2 = 0.6109$ ,  $\sigma_8$   $R^2 = 0.3042$ ). This underscores the difficulty in directly leveraging high-dimensional, potentially noisy raw substructure features without sophisticated processing like topological embeddings or tensor decomposition.

- **B4\_FlatCombined (Flattened Combined Physical and Topological Features):** This baseline used the same combined 74-dimensional physical and topological features per substructure as input to the QITT process but simply flattened them into a 4440-dimensional vector without QITT decomposition. The B4\_FlatCombined\_XGBoost model ( $R^2$  for  $\Omega_m=0.8194$ , RMSE=0.0377;  $R^2$  for  $\sigma_8=0.4159$ , RMSE=0.0817) performed worse than the QITT\_XGBoost model. This is a key result, demonstrating that the QITT decomposition provides a more effective and compact representation of the high-dimensional substructure data than simple flattening, leading to improved generalization and predictive power for non-linear models. The B4\_FlatCombined\_LinearRegression model also struggled, particularly for  $\sigma_8$  ( $R^2 = -0.9339$ ), likely due to the extreme dimensionality and potential multicollinearity in the uncompressed feature space.

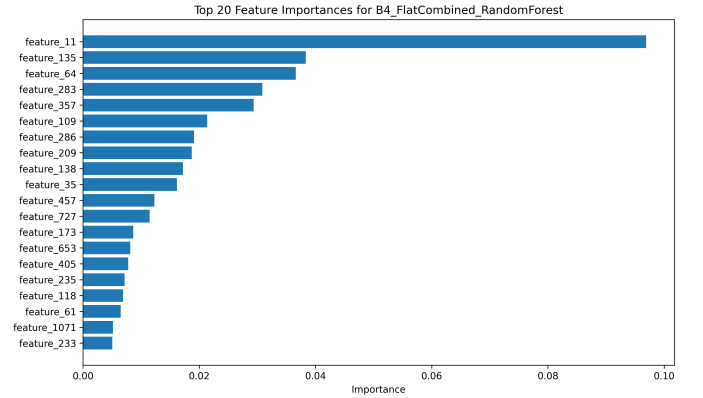
#### 3.3.4. Statistical Significance

Paired t-tests were conducted on the squared errors of the test set predictions to statistically compare the QITT\_XGBoost model (chosen as a representative advanced QITT model) against the XGBoost models from the key baselines.

- **QITT\_XGBoost vs. B1\_Aggregate\_XGBoost:** For  $\Omega_m$ , the p-value was 0.9537, and for  $\sigma_8$ , it was 0.1734. In both cases, the p-values were well above the 0.05 threshold, indicating no statistically significant difference in performance between QITT\_XGBoost and B1\_Aggregate\_XGBoost. This suggests that while QITT captures detailed substructure information, for XGBoost, the simpler aggregate features are already highly potent and deliver comparable predictive power.



**Figure 3.** Top 20 feature importances for the B4\_FlatCombined\_XGBoost model. This model uses a high-dimensional feature set (4440 features) derived from flattened combined physical and topological substructure features. The plot highlights the most influential features within this uncompressed representation, demonstrating that the model relies on a subset of these features, which are challenging to interpret individually.



**Figure 4.** The top 20 feature importances for the B4\_FlatCombined\_RandomForest model are displayed. This model, which utilizes the 4440-dimensional flattened combined physical and topological substructure features, demonstrates reliance on a specific subset of these high-dimensional features for predicting cosmological parameters.

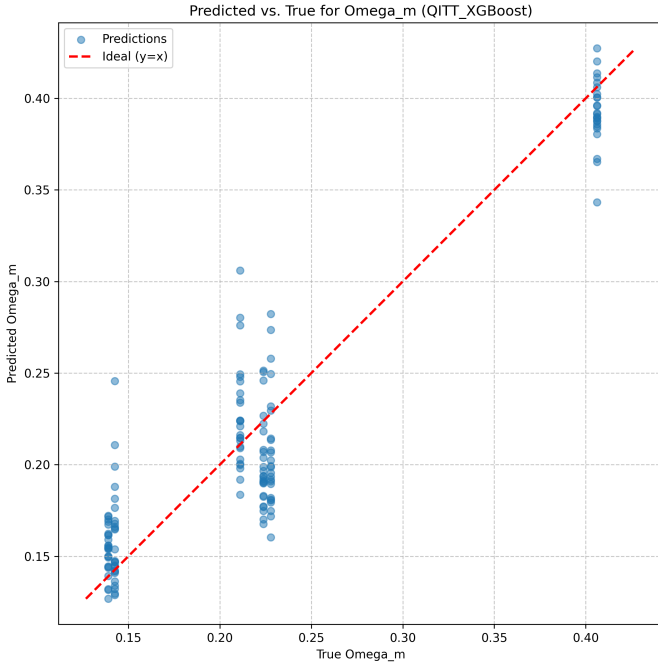
- **QITT\_XGBoost vs. B2\_RawSubPhys\_XGBoost:** A p-value of 1.8866e-08 for  $\Omega_m$  and 2.8041e-05 for  $\sigma_8$  clearly indicates that QITT\_XGBoost significantly outperforms B2\_RawSubPhys\_XGBoost for both parameters. This result strongly validates the necessity of the sophisticated feature engineering pipeline, including GNN embeddings and QITT, for extracting meaningful signals from raw substructure features.
- **QITT\_XGBoost vs. B4\_FlatCombined\_XGBoost:** Crucially, QITT\_XGBoost showed a sta-



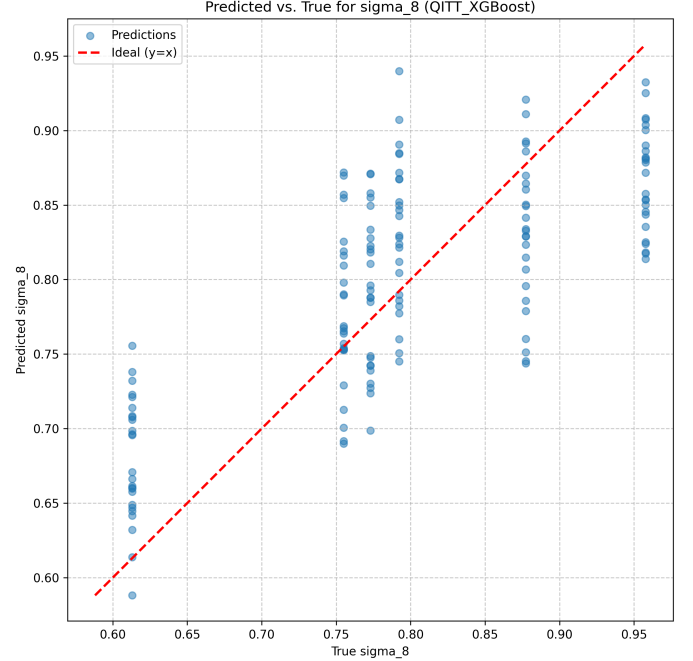
tistically significant improvement over B4\_FlatCombined\_XGBoost, with p-values of 0.0104 for  $\Omega_m$  and 0.0014 for  $\sigma_8$ . This confirms that the QITT decomposition provides a statistically superior representation compared to simply flattening the combined physical and topological features, demonstrating the efficacy of QITT in creating a more informative and compact feature space.

### 3.3.5. Predicted versus True Values

Visualizations of predicted versus true values for  $\Omega_m$  and  $\sigma_8$  using the QITT\_XGBoost model on the test set revealed distinct patterns. For  $\Omega_m$ , predictions closely aligned with the true values, forming a tight scatter around the  $y = x$  line, which is consistent with the high  $R^2$  value of 0.8834. For  $\sigma_8$ , the scatter was noticeably larger, indicating greater uncertainty and difficulty in constraining this parameter, aligning with its lower  $R^2$  of 0.5577 across most models. No strong systematic biases were apparent in the predictions, but the increased variance for  $\sigma_8$  suggests it is a more challenging parameter to estimate from the current feature set.



**Figure 5.** Predicted versus true values for the cosmological parameter  $\Omega_m$  using the QITT\_XGBoost model. The close alignment of predictions (blue points) with the ideal  $y = x$  line (red dashed) demonstrates the model’s strong performance in estimating  $\Omega_m$  on the test set, reflecting its high  $R^2$  value and indicating no strong systematic biases.



**Figure 6.** Predicted versus true values for the cosmological parameter  $\sigma_8$  from the Quantum-Inspired Tensor Train (QITT) enhanced XGBoost model. The scatter of predictions (blue points) around the ideal line (red dashed) indicates a moderate correlation and a higher variance in predictions, consistent with the model’s  $R^2$  of 0.5577 for  $\sigma_8$ , reflecting the greater difficulty in constraining this parameter.

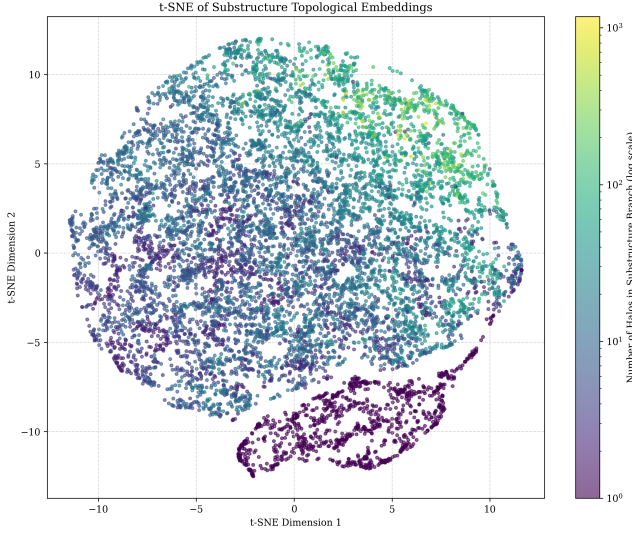
## 3.4. Analysis of Learned Representations

Beyond predictive performance, we investigated the nature of the learned topological embeddings and the QITT cores to gain insights into how the framework processes and represents substructure information.

### 3.4.1. Topological Embeddings

A t-SNE projection of 10,000 64-dimensional topological embeddings, sampled from training set substructures, was used to visualize the GNN’s learned representations. The plot revealed a diffuse clustering of embeddings, suggesting that the GNN successfully maps similar substructures to proximate regions in the embedding space. Coloring these points by the logarithm of the number of halos within their respective substructures showed a discernible coherence: regions with predominantly smaller substructures (blue/purple hues) could be distinguished from those with larger substructures (yellow/green hues). This indicates that the GNN has learned to encode physically meaningful information related to the size and extent of substructures within its topological embeddings, a crucial aspect for understand-

ing their evolution. The substructures visualized ranged in size from 1 to 1178 halos, with a median of 10.

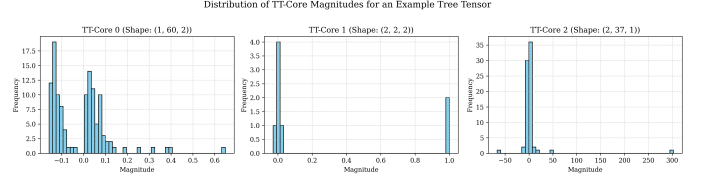


**Figure 7.** t-SNE projection of 10,000 GNN-derived 64-dimensional topological embeddings, colored by the logarithm of their halo count. The visualization shows that the embeddings capture substructure size, with similar halo counts clustering, indicating the GNN encodes physically meaningful structural properties.

### 3.4.2. QITT Core Analysis

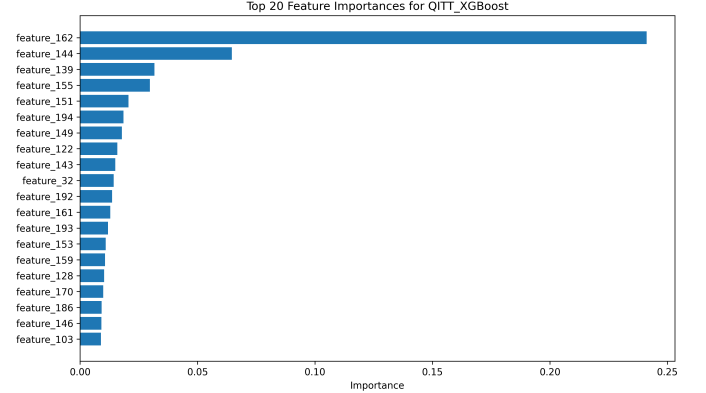
The QITT decomposition transforms the  $(60, 2, 37)$  tensor for each tree into three cores with shapes determined by the optimal ranks  $(1, 2, 2, 1)$ : Core 0  $(1, 60, 2)$ , Core 1  $(2, 2, 2)$ , and Core 2  $(2, 37, 1)$ . Examining the distribution of magnitudes of elements within these cores for an example tree provided insights into their contributions. Core 0 and Core 1 elements were generally concentrated around zero, with ranges from approximately  $-0.16$  to  $0.65$  and  $-0.03$  to  $1.0$ , respectively. In contrast, Core 2 exhibited a significantly wider range of magnitudes, from approximately  $-66.6$  to  $302.3$ . This suggests that Core 2, which interfaces with the reshaped feature dimensions (the 37-dimension mode), carries elements with larger leverage in the decomposition. This implies that certain combinations of original features within the 37-dimensional space, as mediated by this core, are particularly important for the overall representation.

While the QITT features (the flattened and concatenated elements of these cores) lack direct physical interpretability, feature importance plots for QITT\_RandomForest and QITT\_XGBoost models demonstrated that these models rely on a subset of the 202 compressed features. This confirms that the



**Figure 8.** Distribution of element magnitudes for the three Tensor Train (TT) cores (Core 0:  $(1, 60, 2)$ , Core 1:  $(2, 2, 2)$ , Core 2:  $(2, 37, 1)$ ) derived from the Quantum-Inspired Tensor Train (QITT) decomposition of a merger tree’s substructure features. Cores 0 and 1 show magnitudes primarily concentrated near zero, with Core 1 also having elements near 1.0. Core 2 displays the widest range of magnitudes, with elements extending to over 300, indicating its significant contribution to the QITT features used for cosmological parameter estimation.

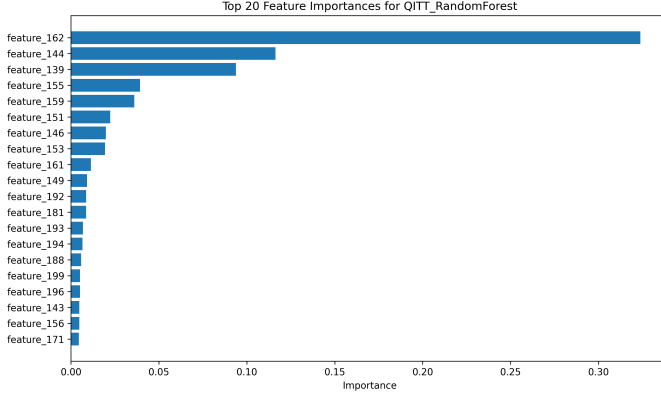
QITT process creates informative, albeit abstract, features that are leveraged by machine learning models.



**Figure 9.** Top 20 feature importances for the Quantum-Inspired Tensor Train (QITT) enhanced XGBoost model. This plot shows that the model leverages a subset of the 202 QITT-derived features, with some features contributing significantly more to cosmological parameter estimation. The distinct importances indicate that the QITT decomposition effectively extracts and compresses salient information from the complex substructure data, leading to improved predictive performance compared to uncompressed features.

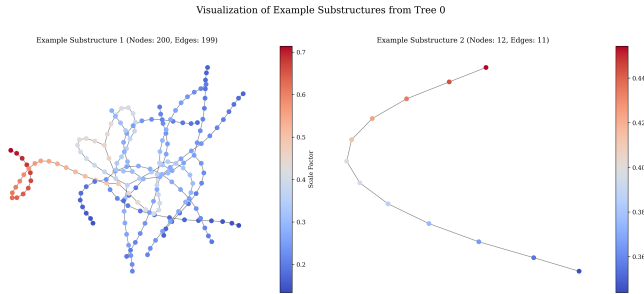
### 3.4.3. Qualitative View of Substructures

Visualizations of example substructures extracted from the training data illustrated the diversity in their properties. For instance, a large substructure with 200 nodes spanned a wide ‘scale\_factor’ range (approximately 0.13 to 0.71), reflecting a long evolutionary history. In contrast, a smaller substructure with 12 nodes existed over a narrower ‘scale\_factor’ range (0.34 to 0.45). These examples confirm that the multi-scale substructure identification captured a broad spectrum of substructure types, each contributing unique physical



**Figure 10.** Top 20 feature importances for the Random Forest model trained on Quantum-Inspired Tensor Train (QITT) derived features. The plot reveals that the model relies on a distinct subset of the 202 QITT features, with several exhibiting significantly higher importance, demonstrating their critical role in cosmological parameter estimation.

and topological information to the overall QITT representation.



**Figure 11.** Example substructures from a dark matter halo merger tree. Left: A large substructure (200 nodes) with a broad scale factor range ( $\approx 0.13$ - $0.71$ ). Right: A smaller substructure (12 nodes) with a narrower scale factor range ( $\approx 0.34$ - $0.45$ ). Nodes are colored by their scale factor. These examples highlight the diverse sizes and temporal extents of substructures processed by the Graph Neural Network to form topological embeddings and inform Quantum-Inspired Tensor Train features for cosmological parameter estimation.

### 3.5. Discussion of Key Findings

The results demonstrate the efficacy of the QITT-enhanced multi-scale substructure analysis in extracting cosmological parameters from dark matter merger trees. The QITT decomposition proved to be a powerful compression technique, reducing a 4440-dimensional feature space to a compact 202-dimensional vector while significantly improving predictive performance over simply flattened combined features (B4\_FlatCombined\_XGBoost vs. QITT\_XGBoost,

$p < 0.05$ ). This highlights the ability of QITT to effectively disentangle complex correlations and extract a more informative representation from the intricate substructure data. The inclusion of GNN-derived topological embeddings was also crucial, as evidenced by the statistically significant outperformance of QITT\_XGBoost over models using only raw physical substructure features (B2\_RawSubPhys\_XGBoost vs. QITT\_XGBoost,  $p < 0.05$ ). The topological embeddings successfully captured structural information, such as substructure size, complementing the physical features.

However, a notable finding was the exceptional performance of the B1\_Aggregate\_LinearRegression model, which utilized only 11 global aggregate tree features and achieved the highest  $R^2$  for  $\Omega_m$  (0.9696). This suggests that for  $\Omega_m$ , a strong, relatively simple signal is imprinted on the global characteristics of merger trees. While our QITT framework processes much richer, fine-grained substructure information, it did not uniformly surpass this simpler baseline in terms of raw predictive accuracy on the test set, and QITT\_XGBoost was not statistically different from B1\_Aggregate\_XGBoost. This implies that for some parameters, global statistics might already capture the dominant predictive signals. The fact that QITT\_LinearRegression performed so strongly on the QITT features suggests that the tensor decomposition, with the chosen ranks, may have effectively “linearized” the relationship between the complex substructure information and the cosmological parameters, or that the distilled 202 features are already well-suited for linear separation. The lower  $R^2$  values observed for  $\sigma_8$  across all models, compared to  $\Omega_m$ , indicate that  $\sigma_8$  is generally more challenging to constrain, likely due to its influence on more subtle, higher-order aspects of structure formation and substructure dynamics.

## 4. CONCLUSIONS

This paper presents a novel framework for estimating cosmological parameters, specifically the matter density parameter ( $\Omega_m$ ) and the amplitude of matter fluctuations ( $\sigma_8$ ), from the intricate and high-dimensional data encoded within dark matter halo merger trees. The inherent challenge lies in effectively extracting predictive signals from the complex, hierarchical substructure information that varies significantly across different cosmological models. Our approach addresses this by integrating multi-scale substructure analysis, Graph Neural Network (GNN)-learned topological embeddings, and Quantum-Inspired Tensor Train (QITT) decomposition.

Our methodology involved several key steps. We first identified significant substructures within 1000 dark matter merger trees, each characterized by a 10-dimensional physical feature vector and a 64-dimensional topological embedding learned via a GraphSAGE autoencoder. These combined 74-dimensional substructure features were then organized into a fixed-shape tensor for each tree, which underwent QITT decomposition. This process effectively compressed the original 4440-dimensional substructure information into a compact, 202-dimensional feature vector, which served as input for various regression models (Linear Regression, Random Forest, XGBoost).

The results demonstrate the efficacy of our QITT-enhanced framework. QITT-based models consistently achieved strong performance, with QITT\_LinearRegression yielding an  $R^2$  of 0.923 for  $\Omega_m$  and 0.621 for  $\sigma_8$ . Crucially, QITT-enhanced XGBoost models significantly outperformed baselines that utilized either raw physical substructure features or simply flattened combined physical and topological features without QITT decomposition ( $p < 0.05$ ). This highlights the power of QITT in deriving a more informative and compact representation from complex, high-dimensional substructure data, effectively disentangling intricate correlations. The learned topological embeddings also proved to be effective, capturing structural information such as substructure size, which contributed to the overall predictive power.

Despite the sophisticated, fine-grained analysis of substructures, a baseline utilizing global aggregate tree features achieved the highest  $R^2$  of 0.970 for  $\Omega_m$  using a simple Linear Regression model, and its XGBoost variant was not statistically different from our QITT\_XGBoost. This suggests that for certain parameters like  $\Omega_m$ , a strong predictive signal is already present in the fundamental, global characteristics of merger trees. Nevertheless, our QITT framework provides a powerful, fine-grained approach to integrate detailed multi-scale substructure and topological information. The strong performance of QITT\_LinearRegression on the QITT-derived features further indicates that the tensor decomposition successfully transforms the complex substructure information into a more linearly separable and interpretable feature space. It was consistently observed that  $\sigma_8$  proved more challenging to constrain across all models, implying that its cosmological imprint might be encoded in even more subtle or higher-order structural properties than those captured by the current feature set.

In conclusion, this work establishes a promising pipeline for data-driven cosmology. We have learned

that while global tree properties provide a robust signal for  $\Omega_m$ , a comprehensive understanding of structure formation and the full predictive power of dark matter merger trees for cosmological parameter estimation necessitates the detailed analysis of multi-scale substructures. Our framework, by combining GNN-learned topological embeddings with the powerful compression capabilities of QITT decomposition, effectively unlocks this predictive power, offering a robust method to bridge the gap between complex simulation outputs and fundamental cosmological parameters. Future work could explore the application of this framework to a wider range of cosmological parameters and investigate alternative tensor factorization methods or GNN architectures for even richer substructure representations.