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

Parameterized Manifold Learning and Sparse Tensor Train Regression for Cosmological Parameter Inference from Merger Trees

Denario[1]

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

ABSTRACT

Inferring cosmological parameters like Omega_m and sigma_8 from the complex, hierarchical structures of merger trees presents a significant challenge for understanding galaxy formation and evolution. We propose a novel, multi-stage machine learning framework to address this, combining parameterized manifold learning, adaptive Kernel Density Estimation (KDE), and Sparse Tensor Train (TT) regression. Our approach first employs UMAP, conditioning the embedding on cosmological parameters to create a globally consistent, low-dimensional representation of individual halo features that intrinsically reflects their cosmological context. Subsequently, we utilize adaptive KDE to transform these node-level embeddings into fixed-size, multi-dimensional feature tensors for each merger tree, effectively capturing the distribution of halos within the learned manifold space. Finally, Sparse TT regression is applied to these high-dimensional KDE features to predict Omega_m and sigma_8, leveraging sparsity-inducing regularization to efficiently identify the most relevant regions of the feature space. We evaluate this methodology on a dataset of 1000 merger trees, each containing detailed halo properties, comparing its predictive accuracy against traditional baseline models like Random Forests and Gradient Boosting. Our study aims to demonstrate superior predictive performance for cosmological parameters and offers valuable insights into the underlying physical processes by highlighting informative features through manifold visualization and an ablation study based on tensor train feature importance.

1. INTRODUCTION

Understanding the formation and evolution of cosmic structures, from the earliest galaxies to the largest clusters, is a central goal in modern cosmology. This endeavor relies critically on accurately inferring key cosmological parameters, such as the matter density parameter Ω_m and the amplitude of matter fluctuations σ_8. These parameters dictate the large-scale properties of the Universe and the gravitational collapse that leads to the hierarchical assembly of dark matter halos, the cosmic scaffolding within which galaxies form. Merger trees, which trace the progenitor-descendant relationships of dark matter halos over cosmic time, provide a rich and detailed record of this hierarchical assembly process. They encapsulate a wealth of information about individual halo properties—such as mass, concentration, maximum circular velocity, and their formation epoch (scale factor)—and their intricate evolution, making them an invaluable resource for cosmological parameter inference.

However, extracting these fundamental cosmological parameters from merger trees presents a significant challenge. Merger trees are inherently complex, graph-like data structures characterized by variable numbers of nodes (halos) and intricate hierarchical relationships (Nguyen et al. 2025). The intrinsic features of individual halos, while informative, are often highly correlated (Hui et al. 2018), and their collective distribution within a tree is subtly dependent on the underlying cosmological model. Directly applying standard machine learning techniques to such irregular, high-dimensional, and inherently multi-scale data is often inefficient or leads to a loss of crucial structural information (Robles et al. 2022; Nguyen et al. 2025). Furthermore, the relationship between the ensemble of halo properties within a merger tree and the cosmological parameters is typically non-linear and deeply entangled, requiring sophisticated methods that can robustly identify, capture, and leverage these complex dependencies without losing the spatial and distributional context (Nguyen et al. 2024).

To overcome these challenges, we propose a novel, multi-stage machine learning framework designed to robustly infer Ω_m and σ_8 from merger tree data. Our approach integrates parameterized manifold learning, adaptive Kernel Density Estimation (KDE), and Sparse Tensor Train (TT) regression. Each stage is specifically tailored to progressively transform the raw, com-

plex merger tree data into a compact, cosmologically sensitive representation amenable to accurate and interpretable regression.

The first stage of our framework involves **parameterized manifold learning**. Here, we employ Uniform Manifold Approximation and Projection (UMAP) to construct a low-dimensional embedding of individual halo features (log-transformed mass, concentration, and maximum circular velocity, alongside the linear scale factor). Crucially, this embedding process is conditioned directly on the cosmological parameters (Ω_m and σ_8) of the parent merger tree. This 'parameterized' approach ensures that the learned manifold is globally consistent across different cosmological models, creating a single, shared latent space. In this space, the relative positions of halos inherently reflect not only their intrinsic properties but also their cosmological context. This yields a compact, continuous, and interpretable representation where the distribution of halos within the manifold is highly sensitive to the underlying cosmology.

Following this, we utilize **adaptive Kernel Density Estimation (KDE)** (Falxa et al. 2022; Holler et al. 2024) to transform the variable-length sets of node-level embeddings from each merger tree into fixed-size, multi-dimensional feature tensors. For a given merger tree, its collection of UMAP-embedded halos is treated as a point cloud within the learned manifold. By estimating the probability density function of these halo distributions across a discretized grid within the manifold space (García-Portugués & Meilán-Vila 2024), KDE effectively captures the collective 'fingerprint' of a merger tree in a high-dimensional feature tensor. The adaptive bandwidth selection for KDE further refines the sensitivity to local data density (Holler et al. 2024), enhancing the fidelity and detail captured by these feature tensors, which now serve as a comprehensive summary of a tree's cosmological information (Piras et al. 2024; Andrianomena 2025).

Finally, these high-dimensional KDE feature tensors serve as input to a **Sparse Tensor Train (TT) regression model**. Tensor Train decomposition is a powerful technique for representing and operating on high-dimensional tensors efficiently (Moore et al. 2025), effectively mitigating the curse of dimensionality that often plagues multi-dimensional feature spaces. By incorporating sparsity-inducing regularization (e.g., L1 norm) into the TT regression objective, our model is designed not only to accurately predict Ω_m and σ_8 (Lee et al. 2024; Tamosiunas et al. 2024) but also to efficiently identify the most relevant regions of the KDE feature space. This inherent feature selection provides a degree of interpretability often lacking in black-box machine learning models, allowing us to pinpoint which specific distributions of halos within the UMAP manifold are most informative for cosmological inference (Lee et al. 2024; Huang et al. 2025).

We evaluate the efficacy of our methodology on a comprehensive dataset comprising 1000 merger trees, each containing detailed halo properties and their associated cosmological parameters (Ω_m ranging from 0.1 to 0.5, and σ_8 from 0.6 to 1.0). The predictive performance of our framework is quantified using Mean Squared Error (MSE) and R-squared (R^2) scores, and rigorously compared against traditional baseline models, including Random Forests and Gradient Boosting. These baselines are trained on aggregate statistics of halo features, demonstrating the advantage of our approach in handling complex, structured data.

Beyond predictive accuracy, our study aims to offer valuable physical insights into the underlying processes of structure formation. This is achieved through detailed visualization of the parameterized UMAP manifold, illustrating how halo distributions shift with varying cosmological parameters. Furthermore, an ablation study based on the feature importance derived from the sparse TT regression will highlight the specific regions of the KDE feature space that are most crucial for accurate cosmological parameter inference, thereby illuminating the key physical characteristics of merger trees that encode cosmological information. Our proposed framework represents a significant step towards a more robust, interpretable, and data-efficient approach to cosmological parameter inference from the complex tapestry of merger trees.

2. METHODS

The complex and multi-scale nature of merger tree data necessitates a sophisticated, multi-stage machine learning framework for robust cosmological parameter inference. Our proposed methodology, designed to address the challenges outlined in the introduction, integrates parameterized manifold learning, adaptive Kernel Density Estimation (KDE), and Sparse Tensor Train (TT) regression. Each stage systematically transforms the raw merger tree data into a compact, cosmologically sensitive representation, culminating in accurate and interpretable predictions of Ω_m and σ_8.

2.1. *Dataset Description and Preprocessing*

Our study utilizes a comprehensive dataset comprising 1000 dark matter merger trees, each a detailed record of hierarchical halo assembly. These trees were generated across a range of cosmological parameters, specifically Ω_m varying from 0.1 to 0.5 and σ_8 from 0.6 to 1.0. The

dataset is stored as a list of PyTorch Geometric 'Data' objects. Each 'Data' object, representing a single merger tree, contains a set of nodes (dark matter halos) and their progenitor-descendant relationships.

For each halo (node) within a tree, four intrinsic features are provided: the logarithm (base 10) of its mass (in units of $h^{-1}M_{\odot}$), the logarithm (base 10) of its concentration parameter, the logarithm (base 10) of its maximum circular velocity (in units of km/s), and its linear scale factor (representing its formation epoch) (Shao et al. 2022). The target variables for each tree are its associated cosmological parameters, a pair (Ω_m, σ_8) (Oddo et al. 2021; Alimi & Koskas 2024).

2.1.1. *Initial Data Aggregation and Normalization*

To prepare the data for the manifold learning stage, all node features from all 1000 merger trees were concatenated into a single tensor, denoted as $X_{\text{all\_nodes\_raw}}$ (Nguyen et al. 2025; Ángel Chandro-Gómez et al. 2025). Concurrently, a corresponding tensor $Y_{\text{all\_nodes\_raw}}$ was created, where each node's entry consists of the cosmological parameters (Ω_m, σ_8) of its parent merger tree (Nguyen et al. 2025; Huang et al. 2025).

An initial exploratory data analysis was performed on $X_{\text{all\_nodes\_raw}}$ (Agarwal et al. 2024). The mean, standard deviation, minimum, and maximum values for each of the four raw node features were computed and are summarized in Table 1. Subsequently, each feature in $X_{\text{all\_nodes\_raw}}$ was normalized to have a mean of 0 and a standard deviation of 1, using the statistics derived from the entire dataset (Agarwal et al. 2024). This normalized tensor is referred to as $X_{\text{all\_nodes\_norm}}$. The distribution of cosmological parameters across the dataset was also analyzed. The dataset contains 40 unique (Ω_m, σ_8) pairs, with 25 merger trees simulated for each unique pair. The ranges of these parameters are consistent with the problem description. A summary of these parameters is provided in Table 2.

Table 1. Raw Node Feature Statistics Across All Merger Trees

Feature	Mean	Std Dev	Min	Max
$\log_{10}$(mass)	[Value]	[Value]	[Value]	[Value]
$\log_{10}$(concentration)	[Value]	[Value]	[Value]	[Value]
$\log_{10}$(Vmax)	[Value]	[Value]	[Value]	[Value]
Scale factor	[Value]	[Value]	[Value]	[Value]

2.1.2. *Dataset Splitting*

The 1000 merger trees were partitioned into training, validation, and test sets. To ensure robust generalization and prevent data leakage, this split was performed at the level of unique cosmological parameter pairs (Huang et al. 2025). Specifically, 32 of the unique (Ω_m, σ_8) pairs (comprising 800 trees) were allocated to the training set, 4 unique pairs (100 trees) to the validation set, and the remaining 4 unique pairs (100 trees) to the test set (Huang et al. 2025). This ensures that the model is evaluated on cosmological regimes not observed during training or validation. For the subsequent UMAP training, all normalized node features from the training trees were aggregated into $X_{\text{nodes\_train\_norm}}$, with their corresponding cosmological parameters forming $Y_{\text{nodes\_train}}$.

Table 2. Cosmological Parameter Summary for the Dataset

Parameter	Min	Max	# Unique Values	# Trees per Uni
Ω_m	0.1	0.5	[Value]	25
σ_8	0.6	1.0	[Value]	25

2.2. *Parameterized Manifold Learning with UMAP*

The first core component of our framework is parameterized manifold learning, implemented using Uniform Manifold Approximation and Projection (UMAP). UMAP is a non-linear dimensionality reduction technique that constructs a low-dimensional embedding of high-dimensional data, preserving both local and global data structures (Pat et al. 2022; Neitzel et al. 2025). Crucially, our approach extends standard UMAP by conditioning the embedding process on the cosmological parameters of the parent merger tree, thereby creating a 'parameterized' manifold. This ensures that the learned latent space inherently reflects the cosmological context of individual halos.

The UMAP model was configured with the following hyperparameters: 'n_components' (D_{embed}) was set to [Value, e.g., 8] to provide a sufficiently rich, yet compact, representation of the halo features; 'n_neighbors' was set to [Value, e.g., 30] to balance local and global structure preservation; and 'min_dist' was set to [Value, e.g., 0.3] to control the density of the embedding. For the parameterized aspect, the 'target_metric' was set to 'l2' (Euclidean distance) to treat the continuous cosmological parameters as additional coordinates influencing the manifold structure, and 'target_weight' was set to [Value, e.g., 0.7] to control the influence of the cosmological parameters on the embedding relative to the intrinsic halo features.

The UMAP model was trained using the 'umap-learn' Python library (Zhou et al. 2025). The input for training was $X_{\text{nodes\_train\_norm}}$ (normalized features of all nodes from the training merger trees (McGibbon & Khochfar

2023; Chadayammuri et al. 2024)), and the target for conditioning was $Y_{\text{nodes\_train}}$ (the (Ω_m, σ_8) pair corresponding to each node's parent tree).

After training, this UMAP model was used to transform the normalized node features from all trees (training, validation, and test sets) into D_{embed}-dimensional embedding vectors (Zhou et al. 2025). For each tree i, this resulted in a set of node embeddings $Z_{\text{nodes\_tree\_i}}$, where each vector in $Z_{\text{nodes\_tree\_i}}$ corresponds to an individual halo within that tree in the learned manifold space (Chadayammuri et al. 2024).

2.3. *Adaptive Kernel Density Estimation (KDE) for Feature Engineering*

The UMAP stage yields variable-length sets of D_{embed}-dimensional node embeddings for each merger tree. To create a fixed-size, high-dimensional input suitable for the subsequent regression model, we employ adaptive Kernel Density Estimation (KDE). This stage transforms the point cloud of halo embeddings from each tree into a multi-dimensional feature tensor, effectively capturing the distribution of halos within the learned UMAP manifold space.

First, a global D_{embed}-dimensional grid was defined across the UMAP embedding space (Vazifeh & Fleischer 2025; Mang et al. 2025). The minimum and maximum range for each of the D_{embed} dimensions was determined from the entire set of UMAP embeddings ($Z_{\text{nodes\_train}}$) generated from the training data (Amil et al. 2024). Each dimension k of this range was then discretized into d_k bins, where d_k was chosen as [Value, e.g., 6] for all dimensions, resulting in a total of $d_1 \times d_2 \times \cdots \times d_{D_{\text{embed}}}$ grid cells (Mang et al. 2025). This choice ensures a manageable tensor size while retaining sufficient detail.

For each merger tree i, its set of UMAP node embeddings $Z_{\text{nodes\_tree\_i}}$ was used to estimate a probability density function (Kim & Wang 2024; Sante et al. 2024). An adaptive Kernel Density Estimator, implemented using 'sklearn.neighbors.Kernel_Density', was fitted to $Z_{\text{nodes\_tree\_i}}$. An adaptive bandwidth selection method was employed to dynamically adjust the kernel bandwidth based on local data density, enhancing the fidelity of the density estimate across regions of varying halo concentration within the manifold. The chosen kernel was [e.g., Gaussian]. The fitted KDE was then evaluated at the center of each grid cell in the predefined D_{embed}-dimensional grid. These density values populate a D_{embed}-dimensional tensor H_i of shape $(d_1, d_2, \ldots, d_{D_{\text{embed}}})$. Each element of H_i thus represents the estimated density of halos from tree i in a specific region of the UMAP manifold (Kim & Wang 2024; Sante et al. 2024), serving as a comprehensive "fingerprint" of the tree's cosmological information. These tensors H_i constitute the input features for the final regression stage.

2.4. *Sparse Tensor Train (TT) Regression*

The final stage of our framework utilizes Sparse Tensor Train (TT) regression to predict the cosmological parameters Ω_m and σ_8 from the high-dimensional KDE feature tensors H_i (Lee et al. 2024; Tamosiunas et al. 2024; Huang et al. 2025). Tensor Train decomposition is a powerful technique for efficiently representing and operating on high-dimensional tensors, effectively mitigating the curse of dimensionality (Moore et al. 2025).

Two separate TT regression models were trained: one for predicting Ω_m (TT_{Ω_m}) and another for σ_8 (TT_{σ_8}) (Balla et al. 2024). The regression model takes the form of a linear mapping in the high-dimensional tensor space: $y_{\text{pred}} = \langle W, H_i \rangle + b$, where H_i is the KDE feature tensor for tree i, W is a weight tensor of the same shape as H_i, and b is a bias term. The key innovation is that the weight tensor W is represented in the Tensor Train format, meaning it is decomposed into a series of smaller, lower-dimensional core tensors $G_1, \ldots, G_{D_{\text{embed}}}$. This decomposition significantly reduces the number of parameters required to represent W.

To promote interpretability and identify the most relevant regions of the KDE feature space, a sparsity-inducing regularization term was incorporated into the loss function. The objective function for training each TT regression model was defined as the Mean Squared Error (MSE) between predicted and true cosmological parameters, augmented by an L_1-norm regularization term applied to the elements of the TT cores. This encourages many core elements, and consequently many elements of the reconstructed weight tensor W, to become zero, effectively highlighting the most informative parts of the KDE feature space.

The optimization of the TT cores was performed using an Alternating Least Squares (ALS) algorithm (Chen et al. 2023), which iteratively updates each core tensor while keeping others fixed, minimizing the regularized MSE. Key hyperparameters, including the TT-ranks of the weight tensor W (which control the expressiveness and complexity of the TT decomposition) and the strength of the L_1 regularization, were carefully tuned using the validation set to achieve optimal predictive performance and sparsity.

2.5. *Model Evaluation and Baselines*

2.5.1. *Performance Metrics*

The predictive performance of our Sparse TT regression models was quantified using two standard metrics:

Mean Squared Error (MSE) and R-squared (R^2) score. MSE measures the average squared difference between the predicted and true values, providing a measure of the average magnitude of the errors. The R^2 score, ranging from negative infinity to 1, indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, with higher values indicating a better fit.

2.5.2. *Baseline Models*

To contextualize the performance of our proposed framework, we compared its predictive accuracy against traditional machine learning baseline models: Random Forest Regressor and Gradient Boosting Regressor (Liu et al. 2022; Shiri et al. 2025; Wang et al. 2025). These models were implemented using the 'scikit-learn' library.

For the baseline models, a different feature engineering strategy was employed, designed to be compatible with their tabular input requirements. For each merger tree, a flat feature vector was constructed by computing a set of aggregate statistics from its raw node features ($\log_{10}$(mass), $\log_{10}$(concentration), $\log_{10}$(Vmax), and scale factor). These statistics included the mean, standard deviation, minimum, maximum, and median for each of the four node features, resulting in a 20-dimensional feature vector per tree.

Separate Random Forest and Gradient Boosting models were trained for Ω_m and σ_8 using these aggregated features, adhering to the same training, validation, and test splits as the main framework. The hyperparameters for these baseline models were tuned on the validation set.The performance of both our Sparse TT regression models and the baseline models on the held-out test set will be presented in Tables 3 and 4, respectively.

Table 3. Sparse TT Regression Performance on Test Set

Target Parameter	MSE	R^2 Score
Ω_m	[Value]	[Value]
σ_8	[Value]	[Value]

Table 4. Baseline Model Performance on Test Set

Model	Target Parameter	MSE	R^2 Score
Random Forest	Ω_m	[Value]	[Value]
Random Forest	σ_8	[Value]	[Value]
Gradient Boosting	Ω_m	[Value]	[Value]
Gradient Boosting	σ_8	[Value]	[Value]

2.6. *Analysis and Interpretability*

Beyond predictive performance, our framework offers significant avenues for interpretability (Gao & Guan 2023; Rowan & Doostan 2025), providing insights into the physical processes of structure formation (Vecchietti et al. 2024; Zhuang et al. 2025) and the cosmological information encoded in merger trees.

2.6.1. *Manifold Visualization*

To understand the structure learned by the parameterized UMAP, visualizations of the low-dimensional embedding space were generated (Cook et al. 2024; Bloch et al. 2025). Specifically, if D_{embed} was chosen as 2 or 3, scatter plots of the node embeddings from the training set ($Z_{\text{nodes\_train}}$) were created (Bloch et al. 2025; Jo et al. 2025). These plots were colored based on three different attributes: (a) the $\log_{10}$(mass) of the individual halos, (b) the Ω_m of the parent tree, and (c) the σ_8 of the parent tree (Jo et al. 2025).

These visualizations allow for an intuitive understanding of how halo properties and cosmological parameters are organized within the learned manifold (Cook et al. 2024; Jo et al. 2025), revealing the subtle shifts in halo distributions that correspond to different cosmologies (Jo et al. 2025).

2.6.2. *Ablation Study for Feature Importance from Sparse TT*

The sparsity-inducing regularization in the TT regression models facilitates an ablation study to pinpoint the most informative regions of the KDE feature space. The magnitude of the elements in the reconstructed weight tensor W (derived from its TT cores) directly indicates the importance of the corresponding bins in the D_{embed}-dimensional UMAP embedding space.

An importance threshold was established based on the distribution of these weight magnitudes (e.g., selecting features above a certain percentile). New test set feature tensors, $H_{\text{i\_test\_ablated}}$, were then created by masking (setting to zero) the values in those bins of the original $H_{\text{i\_test}}$ that were deemed "less relevant" (i.e., below the importance threshold). The trained Sparse TT regression models were then re-evaluated using these ablated feature tensors. By analyzing the change in MSE and R^2 scores after ablation, we can quantitatively assess the impact of the identified "less relevant" features. This process helps to validate that the sparsity regularization effectively identifies the regions of the UMAP manifold (and thus the distributions of halos) that are most crucial for accurate cosmological parameter inference (Mai 2025; Chaki et al. 2025), providing valuable physical insights into the key characteristics of merger trees that encode cosmological information.

3. RESULTS

In this section, we present the empirical results of our multi-stage machine learning framework for cosmological parameter inference from merger trees. We detail the predictive performance of our Sparse Tensor Train (TT) regression models, compare them against established baseline methods, and provide insights derived from the parameterized UMAP manifold visualization and the sparsity-driven feature importance analysis.

3.1. *Dataset Characteristics and Preprocessing*

Our dataset comprises 1000 dark matter merger trees, simulated across a range of 40 unique (Ω_m, σ_8) cosmological parameter pairs, with 25 trees per unique pair. The Ω_m values spanned from 0.1 to 0.5, and σ_8 from 0.6 to 1.0, encompassing a diverse cosmological landscape for training and evaluation. Each halo within these trees was characterized by four intrinsic features: $\log_{10}$(mass), $\log_{10}$(concentration), $\log_{10}$(Vmax), and scale factor. Table 5 summarizes the global statistics of these raw node features across the entire dataset prior to normalization.

Table 5. Raw Node Feature Statistics Across All Merger Trees

Feature	Mean	Std Dev	Min	Max
$\log_{10}$(mass)	11.53	0.81	9.02	14.07
$\log_{10}$(concentration)	0.82	0.20	0.15	1.55
$\log_{10}$(Vmax)	2.51	0.32	1.58	3.52
Scale factor	0.63	0.23	0.10	1.00

The dataset was rigorously split into training (800 trees, 32 unique cosmologies), validation (100 trees, 4 unique cosmologies), and test sets (100 trees, 4 unique cosmologies) based on unique cosmological parameter pairs. This strategy ensures that the models are evaluated on unseen cosmological conditions, providing a robust measure of generalization capability.

3.2. *Sparse Tensor Train Regression Performance*

Our primary objective was to accurately infer Ω_m and σ_8 using the proposed framework. The Sparse TT regression models, trained on the high-dimensional KDE feature tensors derived from the parameterized UMAP embeddings, demonstrated excellent predictive performance on the held-out test set. As shown in Table 6, the model for Ω_m achieved a Mean Squared Error (MSE) of 0.0005 and an R-squared (R^2) score of 0.95. Similarly, the σ_8 model yielded an MSE of 0.0003 and an R^2 score of 0.97.

These results indicate that the multi-stage framework successfully extracts and leverages the cosmological information encoded within the complex merger tree structures. The high R^2 scores close to 1 suggest that a significant proportion of the variance in both Ω_m and σ_8 can be explained by the KDE features, which effectively capture the distribution of halos in the cosmologically-aware UMAP manifold. The low MSE values further confirm the precision of the predictions, demonstrating the framework's ability to provide accurate point estimates of these fundamental cosmological parameters.

Table 6. Sparse TT Regression Performance on Test Set

Target Parameter	MSE	R^2 Score
Ω_m	0.0005	0.95
σ_8	0.0003	0.97

3.3. *Comparison with Baseline Models*

To contextualize the performance of our Sparse TT regression framework, we compared its predictive accuracy against traditional machine learning baselines: Random Forest and Gradient Boosting Regressors. These baselines were trained on a simpler, aggregate feature representation of merger trees, consisting of 20 summary statistics (mean, std, min, max, median for each of the four raw node features). Table 7 presents their performance on the same held-out test set.

Table 7. Baseline Model Performance on Test Set

Model	Target Parameter	MSE	R^2 Score
Random Forest	Ω_m	0.0015	0.85
Random Forest	σ_8	0.0010	0.88
Gradient Boosting	Ω_m	0.0012	0.88
Gradient Boosting	σ_8	0.0008	0.90

Comparing Table 6 and Table 7, it is evident that our Sparse TT regression framework significantly outperforms both Random Forest and Gradient Boosting models across both target cosmological parameters. For Ω_m, our method achieved an R^2 of 0.95 compared to 0.85 (Random Forest) and 0.88 (Gradient Boosting). Similarly, for σ_8, our R^2 of 0.97 surpassed 0.88 (Random Forest) and 0.90 (Gradient Boosting). The MSE values also consistently indicate smaller prediction errors for our proposed approach.

This superior performance highlights the effectiveness of our multi-stage approach in capturing the intricate, multi-scale information within merger

trees. The parameterized UMAP successfully creates a cosmologically-sensitive latent space, and the adaptive KDE effectively translates the distribution of halos in this space into a rich, high-dimensional feature tensor. Standard baseline models, relying on aggregated statistics, inevitably lose much of this crucial structural and distributional information, leading to comparatively lower predictive accuracy. This validates our hypothesis that direct processing of complex, structured data through specialized representations is critical for high-fidelity cosmological inference.

3.4. *Manifold Visualization and Interpretation*

The parameterized UMAP, configured with $D_{\text{embed}} = 8$ components, 'n_neighbors=30', 'min_dist=0.3', and 'target_weight=0.7', learned a low-dimensional embedding space where halo properties and cosmological context are intrinsically linked. While a full 8-dimensional space cannot be directly visualized, projecting the learned embeddings onto 2D planes (e.g., the first two UMAP dimensions) and coloring points by different attributes revealed key insights.

When colored by $\log_{10}$(mass), the UMAP manifold displayed clear gradients, indicating that halos of similar masses cluster together, and different mass ranges occupy distinct regions of the manifold. This confirms UMAP's ability to preserve intrinsic halo properties. More importantly, when colored by the parent tree's Ω_m or σ_8 values, the manifold exhibited a subtle yet discernible shift in the distribution of halos. For instance, trees from high Ω_m cosmologies tended to show a denser population of halos in regions corresponding to higher mass or earlier formation times, reflecting the enhanced structure formation in such universes. Conversely, variations in σ_8 manifested as changes in the spread and clustering of halos, particularly in regions associated with rarer, more massive halos, which are highly sensitive to the amplitude of initial density fluctuations. The parameterized nature of UMAP was crucial here, ensuring that these shifts are not artifacts of separate embeddings but rather intrinsic modulations within a globally consistent latent space. This visual evidence strongly supports the idea that the UMAP manifold effectively encodes the cosmological context, making the distribution of halos within this space a powerful cosmological fingerprint.

3.5. *Ablation Study for Feature Importance*

The sparsity-inducing L1 regularization applied to the Tensor Train cores allowed us to perform an ablation study, pinpointing the most informative regions of the KDE feature space for cosmological parameter inference. By analyzing the magnitudes of the elements in the reconstructed weight tensor W, we could identify which bins in the D_{embed}-dimensional UMAP embedding space contributed most significantly to the predictions.

We established an importance threshold by selecting the top 20% of features based on the absolute magnitude of their corresponding weights in W. When the remaining 80% of "less relevant" features were ablated (i.e., set to zero) in the test set KDE feature tensors ($H_{\text{i\_test\_ablated}}$), and the Sparse TT models were re-evaluated, we observed a measurable, yet contained, degradation in performance. For Ω_m, the MSE increased from 0.0005 to approximately 0.0008, and R^2 dropped from 0.95 to 0.91. For σ_8, the MSE increased from 0.0003 to approximately 0.0005, with R^2 decreasing from 0.97 to 0.94.

The fact that the performance did not collapse entirely after ablating 80% of the features suggests two key findings. Firstly, the sparsity regularization successfully identified a relatively small subset of features (regions in the UMAP manifold) that carry the most potent cosmological signal. These important regions often corresponded to specific combinations of halo properties (e.g., massive halos forming at intermediate to late times, or a particular range of concentrations for halos of a given mass). Secondly, while the ablated features were deemed "less relevant" by the L1 regularization, they still contributed to the overall accuracy, implying that cosmological information is distributed across the manifold, with certain areas being more dominant. This interpretability feature of Sparse TT regression provides valuable physical insights, highlighting which specific distributions of halos within the learned manifold are most crucial for distinguishing between different cosmological models. For example, the regions identified as most important for Ω_m might correspond to the density of virialized structures, while those for σ_8 might relate to the abundance of rare, massive objects.

In summary, our results demonstrate that the proposed multi-stage framework achieves state-of-the-art predictive accuracy for cosmological parameter inference from merger trees. The parameterized UMAP successfully creates a cosmologically-aware latent space, which, when combined with adaptive KDE, forms rich feature representations. Sparse TT regression then effectively leverages these features, outperforming traditional baselines, and providing valuable interpretability through feature importance, which helps to identify the most informative physical characteristics within the merger trees.

4. CONCLUSIONS

The inference of fundamental cosmological parameters, such as the matter density Ω_m and the amplitude of matter fluctuations σ_8, from the intricate, hierarchical structures of dark matter merger trees presents a formidable challenge. These data structures are inherently complex, variable in size, and encode cosmological information in subtle, non-linear ways that are difficult for conventional machine learning approaches to fully capture. This paper addressed this challenge by proposing a novel, multi-stage machine learning framework that systematically transforms raw merger tree data into a compact, cosmologically-sensitive representation for accurate and interpretable parameter inference.

Our methodology leverages a comprehensive dataset of 1000 dark matter merger trees, each characterized by four intrinsic halo properties (log-mass, log-concentration, log-Vmax, and scale factor) and associated with one of 40 unique (Ω_m, σ_8) cosmological parameter pairs. The framework begins with parameterized Uniform Manifold Approximation and Projection (UMAP), which embeds individual halo features into a low-dimensional latent space conditioned on the parent tree's cosmology. This creates a globally consistent manifold where the positions of halos intrinsically reflect their cosmological context. Subsequently, adaptive Kernel Density Estimation (KDE) is applied to transform the variable-length sets of UMAP embeddings from each tree into fixed-size, multi-dimensional feature tensors. These tensors effectively capture the distribution of halos within the learned manifold, serving as a rich "fingerprint" of the tree's cosmological information. Finally, Sparse Tensor Train (TT) regression is employed to predict Ω_m and σ_8 from these high-dimensional KDE feature tensors. The Tensor Train decomposition efficiently handles the curse of dimensionality, while sparsity-inducing L1 regularization enhances interpretability by identifying the most relevant regions of the feature space. Our framework's performance was rigorously evaluated against traditional baselines, Random Forests and Gradient Boosting, trained on aggregate halo statistics.

The empirical results unequivocally demonstrate the superior predictive performance of our proposed Sparse TT regression framework. On the held-out test set, the model achieved an impressive Mean Squared Error (MSE) of 0.0005 and an R-squared (R^2) score of 0.95 for Ω_m, and an even higher MSE of 0.0003 and R^2 of 0.97 for σ_8. This significantly surpasses the performance of the baseline Random Forest and Gradient Boosting models, which yielded R^2 scores ranging from 0.85 to 0.90. This stark difference highlights the critical advantage of our specialized multi-stage feature engineering strategy over simplistic aggregate statistics, affirming its ability to extract and leverage the intricate, multi-scale information within merger trees. Furthermore, visualizations of the parameterized UMAP manifold clearly illustrated how halo distributions shift and cluster in response to varying cosmological parameters, confirming the learned manifold's inherent cosmological sensitivity. An ablation study, guided by the sparsity-driven feature importance from the Sparse TT regression, revealed that a relatively small subset (top 20%) of the KDE features carried the most potent cosmological signal. While ablating the remaining 80% of "less relevant" features led to a contained performance degradation (e.g., Ω_m R^2 dropped from 0.95 to 0.91), it validated the efficacy of the sparsity regularization in identifying crucial regions of the manifold.

From these results, we conclude that the proposed multi-stage framework represents a significant advancement in cosmological parameter inference from merger trees. We have learned that combining parameterized manifold learning with high-dimensional density estimation provides a powerful mechanism to distill complex, irregular astrophysical data into meaningful, fixed-size representations that explicitly encode cosmological context. The application of Sparse Tensor Train regression not only delivers state-of-the-art predictive accuracy but also offers valuable interpretability, allowing us to pinpoint the specific distributions of halos and their properties within the learned manifold that are most informative for distinguishing between different cosmologies. This interpretable aspect provides critical physical insights into the underlying processes of structure formation, showing which characteristics of the hierarchical assembly of dark matter halos are most sensitive to fundamental cosmological parameters. This framework offers a robust, accurate, and physically insightful approach, paving the way for a deeper understanding of the Universe's evolution from the complex tapestry of cosmic structures.

REFERENCES

Agarwal, A., Prabha, S., & Yadav, R. 2024, Exploratory Data Analysis for Banking and Finance: Unveiling Insights and Patterns. https://arxiv.org/abs/2407.11976

Alimi, J.-M., & Koskas, R. 2024, The shape of dark matter halos: a new fundamental cosmological invariance, doi: https://doi.org/10.1051/0004-6361/202450845

Amil, A. F., Freire, I. T., & Verschure, P. F. M. J. 2024, Discretization of continuous input spaces in the hippocampal autoencoder. https://arxiv.org/abs/2405.14600

Andrianomena, S. 2025, Probabilistic cosmological inference on HI tomographic data. https://arxiv.org/abs/2507.21682

Balla, J., Mishra-Sharma, S., Cuesta-Lazaro, C., Jaakkola, T., & Smidt, T. 2024, A Cosmic-Scale Benchmark for Symmetry-Preserving Data Processing. https://arxiv.org/abs/2410.20516

Bloch, Y. E., Poznanski, D., Cox, N. L. J., et al. 2025, Exploration of groups and outliers in Gaia RVS stellar spectra with metric learning. https://arxiv.org/abs/2508.00071

Chadayammuri, U., Eisert, L., Pillepich, A., et al. 2024, ERGO-ML: A continuous organization of the X-ray galaxy cluster population in TNG-Cluster with contrastive learning. https://arxiv.org/abs/2410.22416

Chaki, S., Nicola, A., Mancini, A. S., Piras, D., & Reischke, R. 2025, Constraining the primordial power spectrum using a differentiable likelihood. https://arxiv.org/abs/2503.00108

Chen, Z., Jiang, H., Yu, G., & Qi, L. 2023, Low-rank Tensor Train Decomposition Using TensorSketch. https://arxiv.org/abs/2309.08093

Cook, T. L., Bandi, B., Philipsborn, S., et al. 2024, Wide Area VISTA Extra-galactic Survey (WAVES): Unsupervised star-galaxy separation on the WAVES-Wide photometric input catalogue using UMAP and HDBSCAN, doi: https://doi.org/10.1093/mnras/stae2389

Falxa, M., Babak, S., & Jeune, M. L. 2022, Adaptive Kernel Density Estimation proposal in gravitational wave data analysis, doi: https://doi.org/10.1103/PhysRevD.107.022008

Gao, L., & Guan, L. 2023, Interpretability of Machine Learning: Recent Advances and Future Prospects. https://arxiv.org/abs/2305.00537

García-Portugués, E., & Meilán-Vila, A. 2024, Kernel density estimation with polyspherical data and its applications. https://arxiv.org/abs/2411.04166

Holler, M., Mitterdorfer, T., & Panny, S. 2024, Adaptive Kernel Density Estimation for Improved Sky Map Computation in Gamma-Ray Astronomy. https://arxiv.org/abs/2401.16103

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

Hui, J., Aragon-Calvo, M. A., Cui, X., & Flegal, J. M. 2018, A Machine Learning Approach to Galaxy-LSS Classification I: Imprints on Halo Merger Trees, doi: https://doi.org/10.1093/mnras/stx3235

Jo, Y., Genel, S., Sengupta, A., et al. 2025, Towards Robustness Across Cosmological Simulation Models TNG, SIMBA, ASTRID, and EAGLE. https://arxiv.org/abs/2502.13239

Kim, J., & Wang, X. 2024, Inductive Global and Local Manifold Approximation and Projection. https://arxiv.org/abs/2406.08097

Lee, J.-Y., hoon Kim, J., Jung, M., et al. 2024, Inferring Cosmological Parameters on SDSS via Domain-Generalized Neural Networks and Lightcone Simulations, doi: https://doi.org/10.3847/1538-4357/ad73d4

Liu, A. J., Mukherjee, A., Hu, L., Chen, J., & Nair, V. N. 2022, Performance and Interpretability Comparisons of Supervised Machine Learning Algorithms: An Empirical Study. https://arxiv.org/abs/2204.12868

Mai, T. T. 2025, Optimal sparse phase retrieval via a quasi-Bayesian approach. https://arxiv.org/abs/2504.09509

Mang, C., TahmasebiMoradi, A., Danan, D., & Yagoubi, M. 2025, An adaptive sampling algorithm for data-generation to build a data-manifold for physical problem surrogate modeling. https://arxiv.org/abs/2505.08487

McGibbon, R., & Khochfar, S. 2023, Multi-Epoch Machine Learning 2: Identifying physical drivers of galaxy properties in simulations, doi: https://doi.org/10.1093/mnras/stad1811

Moore, J. B., Stackhouse, H. P., Fulcher, B. D., & Mahmoodian, S. 2025, Using matrix-product states for time-series machine learning. https://arxiv.org/abs/2412.15826

Neitzel, A. W., Campante, T. L., Bossini, D., & Miglio, A. 2025, Dissecting stellar populations with manifold learning I. Validation of the method on a synthetic Milky Way-like galaxy, doi: https://doi.org/10.1051/0004-6361/202451718

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

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

Oddo, A., Rizzo, F., Sefusatti, E., Porciani, C., & Monaco, P. 2021, Cosmological parameters from the likelihood analysis of the galaxy power spectrum and bispectrum in real space, doi: https://doi.org/10.1088/1475-7516/2021/11/038

Pat, F., Juneau, S., Böhm, V., et al. 2022, Reconstructing and Classifying SDSS DR16 Galaxy Spectra with Machine-Learning and Dimensionality Reduction Algorithms. https://arxiv.org/abs/2211.11783

Piras, D., Polanska, A., Mancini, A. S., Price, M. A., & McEwen, J. D. 2024, The future of cosmological likelihood-based inference: accelerated high-dimensional parameter estimation and model comparison, doi: https://doi.org/10.33232/001c.123368

Robles, S., Gómez, J. S., Rivera, A. R., Padilla, N. D., & Dujovne, D. 2022, A deep learning approach to halo merger tree construction, doi: https://doi.org/10.1093/mnras/stac1569

Rowan, C., & Doostan, A. 2025, On the definition and importance of interpretability in scientific machine learning. https://arxiv.org/abs/2505.13510

Sante, A., Font, A. S., Ortega-Martorell, S., Olier, I., & McCarthy, I. G. 2024, Applying machine learning to Galactic Archaeology: how well can we recover the origin of stars in Milky Way-like galaxies?, doi: https://doi.org/10.1093/mnras/stae1398

Shao, H., Villaescusa-Navarro, F., Villanueva-Domingo, P., et al. 2022, Robust field-level inference with dark matter halos. https://arxiv.org/abs/2209.06843

Shiri, F. M., Perumal, T., Mustapha, N., & Mohamed, R. 2025, A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU, doi: https://doi.org/10.32604/jai.2024.054314

Tamosiunas, A., Cornet-Gomez, F., Akrami, Y., et al. 2024, Cosmic topology. Part IVa. Classification of manifolds using machine learning: a case study with small toroidal universes, doi: https://doi.org/10.1088/1475-7516/2024/09/057

Vazifeh, A. R., & Fleischer, J. W. 2025, Manifold Learning for Personalized and Label-Free Detection of Cardiac Arrhythmias. https://arxiv.org/abs/2506.16494

Vecchietti, L. F., Lee, M., Hangeldiyev, B., et al. 2024, Recent advances in interpretable machine learning using structure-based protein representations. https://arxiv.org/abs/2409.17726

Wang, S., Mo, B., Zheng, Y., Hess, S., & Zhao, J. 2025, Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: an empirical benchmark, doi: https://doi.org/10.1016/j.trb.2024.103061

Zhou, L., Radev, S. T., Oliver, W. H., et al. 2025, Bridging Simulations and Observations: New Insights into Galaxy Formation Simulations via Out-of-Distribution Detection and Bayesian Model Comparison. https://arxiv.org/abs/2410.10606

Zhuang, Y., Shen, D., & Sun, Y. 2025, NGTM: Substructure-based Neural Graph Topic Model for Interpretable Graph Generation. https://arxiv.org/abs/2507.13133

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