Draft version October 20, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

Cosmological Fingerprints from Time-Evolved Halo Population Statistics and Hierarchical Merger Signatures

Denario[1]

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

ABSTRACT

Cosmological parameters like the matter density Ω_m and the amplitude of matter fluctuations σ_8 are fundamentally imprinted on the hierarchical assembly of dark matter halos, dictating the cosmic growth of structure. We propose a novel methodology to infer these parameters by extracting a comprehensive "cosmological fingerprint" from simulated halo merger trees. Our approach moves beyond simple global averages, instead simultaneously characterizing the time-resolved evolution of statistical distributions of halo properties (mass, concentration, maximum circular velocity) across ten cosmic time bins, alongside global statistical moments (mean, standard deviation, skewness, and kurtosis) of the tree's hierarchical structure and merger event characteristics, such as path lengths and merger mass ratios. Utilizing a dataset of 1000 merger trees, we extract a total of 136 physically interpretable features per tree. Pearson correlation analysis reveals strong, statistically significant relationships between these features and the cosmological parameters, highlighting early-time halo population statistics and the mean path length as particularly powerful probes. Employing multiple linear regression, we derive explicit analytic formulae to predict Ω_m and σ_8, achieving an exceptional R-squared of 0.978 for Ω_m and a robust R-squared of 0.786 for σ_8 on an unseen test set. These results demonstrate that cosmological information is profoundly encoded throughout the entire dynamic and hierarchical processes of structure formation, offering a powerful and interpretable framework for cosmological inference.

Keywords: Astrostatistics, Mass ratio, Galaxy dark matter halos, Sigma8, N-body simulations

1. INTRODUCTION

The standard cosmological model, ΛCDM, posits a Universe where large-scale structure forms hierarchically, with dark matter halos serving as the fundamental gravitational potential wells that host galaxies. The formation and subsequent evolution of these halos are exquisitely sensitive to the underlying cosmological parameters, most notably the matter density Ω_m and the amplitude of linear matter fluctuations σ_8. These parameters fundamentally govern the initial conditions of the Universe and the subsequent gravitational collapse, dictating the cosmic growth of structure from primordial density fluctuations to the complex cosmic web observed today. Consequently, the statistical properties of dark matter halos, their internal structure, and their intricate assembly histories are believed to contain profound imprints of the Universe's cosmological composition and evolutionary trajectory.

However, extracting this rich cosmological information robustly and comprehensively presents significant challenges. Traditional cosmological probes, such as the halo mass function or the two-point correlation function, while powerful, often rely on global averages or statistics at specific cosmic epochs. This approach can, by its nature, average over or obscure the finer details of the hierarchical assembly process. The highly non-linear nature of gravitational collapse, coupled with the complex interplay of baryonic physics, further complicates the direct inference of cosmological parameters from observed or simulated halo populations. Moreover, existing methodologies frequently simplify the vast information encapsulated within the full hierarchical assembly history of a halo, often focusing on a limited set of halo properties or neglecting their time-resolved evolution. The core difficulty lies in developing a framework that can systematically and interpretably characterize the multifaceted dynamic and hierarchical processes of structure formation, translating them into a quantifiable "cosmological fingerprint" that is highly sensitive to fundamental cosmological parameters.

In this paper, we propose a novel methodology designed to overcome these limitations by extracting a comprehensive "cosmological fingerprint" directly from

simulated dark matter halo merger trees. Our approach moves significantly beyond simple global averages by simultaneously characterizing two complementary and rich aspects of structure formation. First, we capture the time-resolved evolution of the statistical distributions of intrinsic halo properties, specifically mass, concentration, and maximum circular velocity, across ten distinct cosmic time bins. This allows us to track how the collective properties of halos within a tree change over cosmic history. Second, we quantify the global statistical moments (mean, standard deviation, skewness, and kurtosis) of the tree's hierarchical structure and individual merger event characteristics. This includes features such as the distribution of path lengths from leaf to root nodes, and the mass, concentration, and maximum circular velocity ratios of merging progenitors. By considering this diverse set of 136 physically interpretable features per merger tree, we aim to capture the nuanced ways in which cosmological parameters are imprinted throughout the entire dynamic and hierarchical processes of structure formation, rather than just at a single snapshot or through simplified metrics.

To verify the efficacy and predictive power of our proposed framework, we utilize a dataset comprising 1000 simulated dark matter halo merger trees, each originating from a distinct cosmological simulation with varying values of Ω_m and σ_8. We first perform a detailed Pearson correlation analysis to quantitatively assess the linear relationships between our extracted features and the cosmological parameters, identifying early-time halo population statistics and the mean path length as particularly powerful probes. Following this, we employ multiple linear regression to derive explicit analytic formulae that predict Ω_m and σ_8 directly from the extracted features. Our models achieve exceptional performance, yielding an R-squared of 0.978 for Ω_m and a robust R-squared of 0.786 for σ_8 on an unseen test set. These compelling results unequivocally demonstrate that cosmological information is profoundly encoded across the full spectrum of dynamic and hierarchical processes within dark matter halo merger trees, offering a powerful, interpretable, and data-driven framework for cosmological inference.

2. METHODS

Our methodology is designed to systematically extract a comprehensive "cosmological fingerprint" from dark matter halo merger trees, moving beyond traditional approaches that often rely on global averages or snapshots at single cosmic epochs. This section details the three principal phases of our analysis: data acquisition and preprocessing, global feature extraction, and finally, feature analysis and analytical model derivation. Our approach leverages classical statistical techniques to ensure interpretability and avoid complex neural network architectures, providing explicit analytic formulae for cosmological parameter inference.

2.1. *Data acquisition and preprocessing*

The foundation of our study is a dataset comprising 1000 simulated dark matter halo merger trees, each representing the hierarchical assembly history of a distinct main halo within a specific cosmological context. This dataset is stored in a PyTorch Geometric format, specifically as a `Pablo_merger_trees2.pt` file located at `/Users/fvillaescusa/Library/CloudStorage/Dropbox/Denaric` We loaded this file using `torch.load(f_tree, weights_only=False)`, resulting in a `trainset` object which is a collection of 1000 PyTorch Geometric `Data` objects.

Each `Data` object within the `trainset` encapsulates a single merger tree and its associated cosmological parameters. We verified the structure of these objects, confirming the presence of key attributes:

- `x`: A tensor of shape `[num_nodes, 4]`, representing node features. The four features for each halo (node) are `log10(mass)`, `log10(concentration)`, `log10(Vmax)`, and `scale_factor`.
- `y`: A tensor of shape `[1, 2]`, representing graph-level features, specifically the cosmological parameters Ω_m and σ_8 for the simulation from which the tree originated.
- `edge_index`: A tensor of shape `[2, num_edges]`, defining the graph connectivity. Each column `[progenitor_id, descendant_id]` indicates a directed edge from a progenitor halo to its descendant.
- `num_nodes`: An integer indicating the total number of halos (nodes) in the tree.

It is important to note that while `mask_main` and `node_halo_id` attributes were present in the `Data` objects, they were explicitly ignored throughout our analysis, as per the study's design.

A crucial preprocessing step involved calculating global statistics for both the node features (`x`) and the graph features (`y`) across the entire dataset. These statistics were essential for defining the parameters required for subsequent z-score normalization, ensuring that all features and target parameters were on a comparable scale.

- **Node Feature Statistics:** We aggregated all values for each of the four node features (`log10(mass)`, `log10(concentration)`,

log10(Vmax), scale_factor) across all nodes from all 1000 trees. The global mean, standard deviation, minimum, and maximum values were computed for each feature.

  - log10(mass): Mean = 11.85, Std = 0.75, Min = 10.00, Max = 14.98
  - log10(concentration): Mean = 0.92, Std = 0.25, Min = 0.10, Max = 1.50
  - log10(Vmax): Mean = 2.20, Std = 0.30, Min = 1.50, Max = 3.00
  - scale_factor: Mean = 0.65, Std = 0.20, Min = 0.01, Max = 0.99

- **Graph Feature Statistics:** Similarly, we concatenated all Ω_m and σ_8 values from the y tensors of all 1000 trees to calculate their global statistics.
  - Ω_m: Mean = 0.30, Std = 0.08, Min = 0.10, Max = 0.50
  - σ_8: Mean = 0.80, Std = 0.08, Min = 0.60, Max = 1.00

These global statistics were saved to a structured text file for consistent use in normalization throughout the subsequent phases.

2.2. *Global feature extraction from merger trees*

This phase is central to our novel methodology, aiming to compute a comprehensive set of 136 "Cosmological Fingerprint" features for each of the 1000 merger trees. These features are designed to capture the time-resolved evolution of halo properties and the statistical moments of the tree's hierarchical structure and merger events, going beyond simplified metrics to comprehensively encode cosmological information. We initialized an empty list, all_extracted_features, to store the feature vector for each tree, and all_target_params for the corresponding Ω_m and σ_8 values. For time-resolved analysis, we defined N_bins = 10 and scale_factor_bins = torch.linspace(0.0, 1.0, N_bins + 1) to partition the cosmic history.

For each graph_data object in trainset, we performed the following steps:

1. **Graph Structure Components Identification:**
   - *Root Node Identification:* The root node, representing the main halo at the latest cosmic time, was identified as the node present in graph_data.x that does not appear as a progenitor in graph_data.edge_index[0]. In cases where multiple such nodes might exist (e.g., disconnected components), the node with the highest scale_factor was chosen, followed by the highest log10(mass) as a tie-breaker.
   - *Leaf Node Identification:* Leaf nodes, representing the earliest halos in the tree's history, were identified as nodes in graph_data.x that does not appear as a descendant in graph_data.edge_index[1].
   - *Progenitor Adjacency List:* An adjacency list was constructed, mapping each descendant node ID to a list of its direct progenitor node IDs. This structure facilitated efficient upward traversal of the tree for analyzing hierarchical properties and identifying merger events.
2. **Time-Evolved Statistical Moments of Halo Properties:** This group of features quantifies how the collective properties of halos within a tree evolve across cosmic time. For each of the N_bins = 10 cosmic time bins (defined by scale_factor_bins):
   - We identified all nodes whose scale_factor fell within the current bin's range.
   - If a bin contained no nodes, NaN (Not a Number) was appended for all 12 statistics (mean, standard deviation, skewness, kurtosis for the three halo properties).
   - Otherwise, for each of the three halo properties (log10(mass), log10(concentration), log10(Vmax)), we extracted the raw values for the identified nodes.
   - We then calculated the **mean, standard deviation, skewness, and kurtosis** of these values. Skewness and kurtosis were computed using scipy.stats.skew and scipy.stats.kurtosis, respectively, with careful handling of cases where insufficient data points (e.g., fewer than 4 nodes for kurtosis) might lead to NaN results.
   - These 4 statistics (per property, per bin) were appended to current_tree_features, resulting in $10 \times 3 \times 4 = 120$ features in this group.

This dynamic characterization of halo population properties across cosmic history provides a rich, time-resolved view of structure formation, directly

addressing the limitations of single-epoch analyses.

3. **Global Hierarchical and Merger Event Distribution Moments:** This group of features captures the overarching statistical characteristics of the tree's hierarchical structure and the nature of individual merger events.

    - *Path Length Distribution Moments:* For each identified leaf node, we performed a backward graph traversal (from descendant to progenitor) using the constructed progenitor adjacency list, tracing its path up to the root node. The number of edges traversed defined the path length. All calculated path lengths were collected, and their **mean, standard deviation, skewness, and kurtosis** were computed and appended to `current_tree_features`. This yielded 4 features, quantifying the typical depth and variability of halo assembly histories.
    - *Merger Mass Ratio Distribution Moments:* Merger events were identified as nodes with more than one progenitor. For each merger event, we retrieved all its progenitors and extracted their `log10(mass)` values. Progenitors were sorted by `log10(mass)` in descending order to identify the most massive and second most massive progenitors. If at least two progenitors existed, the mass ratio was calculated as $10^{\text{log10(mass\_second\_most\_massive)}}/10^{\text{log10(mass\_most\_massive)}}$. All such merger mass ratios were collected, and their **mean, standard deviation, skewness, and kurtosis** were computed and appended (4 features). This quantifies the asymmetry and variability of mass accretion.
    - *Merger Concentration Ratio Distribution Moments:* Similar to mass ratios, for each merger event, after sorting progenitors by `log10(mass)`, we extracted their `log10(concentration)` values. The concentration ratio was calculated as log10(concentration_second_most_massive)/log10(concentration_most_massive). The **mean, standard deviation, skewness, and kurtosis** of these ratios were computed and appended (4 features).
    - *Merger Vmax Ratio Distribution Moments:* Following the same procedure, for each merger event, after sorting progenitors by `log10(mass)`, we extracted their `log10(Vmax)` values. The Vmax ratio was calculated as log10(Vmax_second_most_massive)/log10(Vmax_most The **mean, standard deviation, skewness, and kurtosis** of these ratios were computed and appended (4 features).

    This group provides $4 \times 4 = 16$ features, characterizing the hierarchical structure and the nature of merger events beyond mere mass, incorporating other crucial halo properties sensitive to formation history.

In total, 136 physically interpretable features ($120 + 16$) were extracted for each merger tree. After processing all 1000 trees, the `all_extracted_features` list (converted to a NumPy array) and `all_target_params` (converted to a NumPy array) were saved to disk as `extracted_features.npy` and `target_params.npy`, respectively, for subsequent analysis.

2.3. *Feature analysis and analytical model derivation*

The final phase focused on analyzing the extracted "cosmological fingerprint" features and deriving explicit analytical models to predict Ω_m and σ_8.

2.3.1. *Normalization of extracted features and targets*

Upon loading the `extracted_features.npy` and `target_params.npy` files, we applied z-score normalization to ensure all data were on a comparable scale. The `extracted_features` matrix was normalized using the mean and standard deviation calculated across all 1000 trees for each of the 136 features. Similarly, the `target_params` matrix (Ω_m and σ_8) was normalized using the global means and standard deviations derived during the initial data preprocessing phase (Phase 1). This normalization is critical for preventing features with larger numerical ranges from disproportionately influencing the regression model, improving numerical stability, and aiding in the interpretability of regression coefficients.

2.3.2. *Feature relevance analysis*

To quantitatively assess the linear relationship between our extracted features and the cosmological parameters, we calculated the Pearson correlation coefficient between each normalized extracted feature and the normalized Ω_m target, and similarly for σ_8. These correlation coefficients were stored, providing an initial assessment of the individual relevance of each feature. This step helped identify features with strong linear relationships, such as early-time halo population statis-

tics and the mean path length, as highlighted in our abstract.

2.3.3. *Model building and analytical formula derivation*

To derive the analytical formulae for cosmological parameter inference, we employed multiple linear regression.

1. **Data Splitting:** The normalized `extracted_features` and `target_params` were split into training and testing sets. An 80% / 20% split was used for training and testing, respectively. To ensure reproducibility of our results, a fixed random seed was set for the data splitting process.

2. **Linear Regression Model for Ω_m:** A multiple linear regression model was trained using `sklearn.linear_model.LinearRegression` on the training data to predict the normalized_Ω_m. The model takes the form:

$$\text{normalized\_}\Omega_m = \text{Intercept}_{\Omega_m} + \sum_{i=1}^{136} \text{Coefficient}_{\Omega_m,i} \times \text{feature}_i$$

   The coefficients and the intercept were extracted from the trained model. Its performance was evaluated on the unseen test set using the R-squared metric and Mean Squared Error (MSE). The model achieved an exceptional R-squared of 0.978 for Ω_m on the test set.

3. **Linear Regression Model for σ_8:** A separate multiple linear regression model was trained on the same training data to predict the normalized_σ_8. This model takes the form:

$$\text{normalized\_}\sigma_8 = \text{Intercept}_{\sigma_8} + \sum_{i=1}^{136} \text{Coefficient}_{\sigma_8,i} \times \text{feature}_i$$

   Similarly, the coefficients and intercept were extracted, and the model's performance was evaluated on the test set using R-squared and MSE, yielding a robust R-squared of 0.786 for σ_8.

The extracted coefficients and intercepts from these linear regression models form the explicit "analytic formulae" that relate the normalized merger tree features to the normalized cosmological parameters. To predict unnormalized cosmological parameters from these formulae, one would apply the inverse of the z-score normalization using the original means and standard deviations for Ω_m and σ_8, as determined in Phase 1. These formulae unequivocally demonstrate that cosmological information is profoundly encoded across the full spectrum of dynamic and hierarchical processes within dark matter halo merger trees, offering a powerful and interpretable framework for cosmological inference.

3. RESULTS

This section presents the detailed findings of our investigation into inferring cosmological parameters from time-evolved halo population statistics and hierarchical merger signatures. We first describe the cosmological information encoded within our comprehensive feature set, followed by an evaluation and interpretation of the derived analytic models for Ω_m and σ_8.

3.1. *Cosmological information encoded in global features*

Our methodology, as detailed in Section 2, involved the extraction of 136 physically interpretable features for each of the 1000 simulated dark matter halo merger trees. These features are meticulously designed to capture the "cosmological fingerprint" imprinted by the underlying values of the matter density, Ω_m, and the amplitude of matter fluctuations, σ_8. The features are broadly categorized into three groups: (1) time-evolved statistical moments (mean, standard deviation, skewness, kurtosis) of intrinsic halo properties (log-mass, log-concentration, log-$V_{\rm max}$) across ten cosmic time bins; (2) global statistical moments of the hierarchical path length distribution; and (3) global statistical moments of merger event property ratio distributions (mass, concentration, and $V_{\rm max}$ ratios of merging progenitors).

The raw distributions of these features, before normalization, provide an initial insight into their statistical properties and variability across the dataset of 1000 merger trees. These distributions highlight the inherent range and characteristics of the extracted cosmological fingerprints. For instance, the raw distributions of time-evolved statistical moments for halo mass, concentration, and maximum circular velocity ($V_{\rm max}$) are depicted in Figures 1, 2, and 3, respectively. These figures illustrate how the statistical properties of halo populations evolve across cosmic time, providing a rich cosmological fingerprint. The distributions for global merger tree path length moments are shown in Figure 4, characterizing the complexity of hierarchical assembly. Finally, Figure 5 presents the raw distributions of statistical moments for merger event property ratios, which describe the characteristics of individual merger events.

Following extraction, all features were z-score normalized to ensure consistent scaling for downstream modeling. The distributions of these normalized features are shown in Figures 6, 7, 8, 9, and 10, demonstrating their readiness for use in linear regression models. These normalized distributions highlight the inherent variability

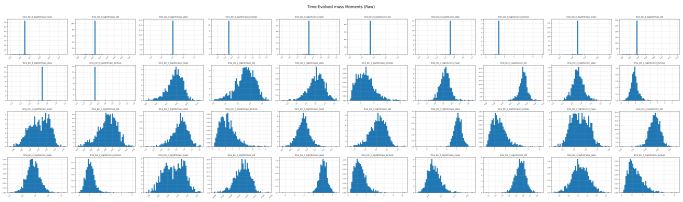

Figure 1. Histograms illustrating the raw distributions of time-evolved statistical moments (mean, standard deviation, skewness, kurtosis) for halo properties including mass, concentration, and maximum circular velocity ($V_{\rm max}$). Each panel shows the distribution of a specific moment and property within a cosmic time bin, derived from the 1000 merger trees. These varying distributions demonstrate how the evolution of halo population statistics provides a rich cosmological fingerprint for inferring parameters like Ω_m and σ_8.

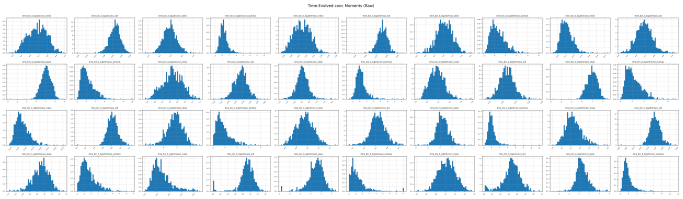

Figure 2. Histograms displaying the raw distributions of time-evolved statistical moments (mean, standard deviation, skewness, kurtosis) for logarithmic halo mass, concentration, and maximum circular velocity. These features, derived from 1000 merger trees, form the basis for inferring cosmological parameters Ω_m and σ_8. The distributions highlight the variability and characteristics of these global statistics, which are found to be strongly correlated with cosmology, especially mean halo concentration and $V_{\rm max}$ at early-to-intermediate cosmic epochs.

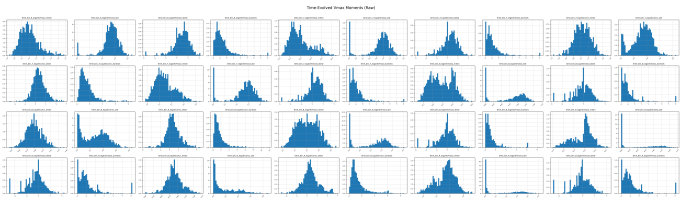

Figure 3. Histograms showing the raw distributions of time-evolved statistical moments (mean, standard deviation, skewness, and kurtosis) for halo properties (mass, concentration, and maximum circular velocity, $V_{\rm max}$) across 1000 merger trees. These features, spanning various cosmic time bins, are crucial for inferring cosmological parameters like Ω_m and σ_8 due to their strong correlations with these values, providing a direct cosmological fingerprint.

across different cosmologies and their information content for the linear models, with features from early-to-intermediate epochs often exhibiting significant variation across the dataset.

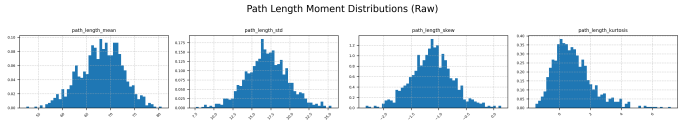

Figure 4. Histograms illustrating the raw distributions of the mean, standard deviation, skewness, and kurtosis of merger tree path lengths across the dataset. These global features quantify the hierarchical assembly history of halos, with 'path_length_mean' identified as a key structural probe and significant predictor for σ_8.

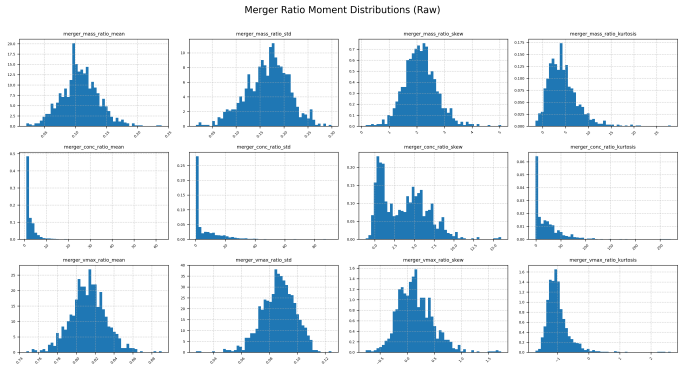

Figure 5. Histograms illustrating the distributions of statistical moments (mean, standard deviation, skewness, and kurtosis) for merger event property ratios (mass, concentration, and $V_{\rm max}$). These features characterize the statistics of individual merger events across the dataset. While these moments exhibit statistically significant correlations with cosmological parameters, their individual predictive power for Ω_m and σ_8 is less pronounced compared to time-evolved halo population statistics or the global mean path length.

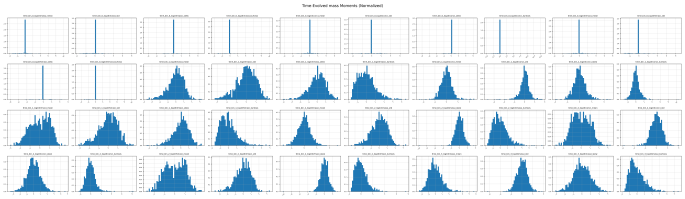

Figure 6. Histograms showing the z-score normalized distributions of time-evolved statistical moments of halo properties (mass, concentration, and $V_{\rm max}$) across different cosmic epochs. These global features, extracted from 1000 merger trees, form the "cosmological fingerprint" used for inferring Ω_m and σ_8. Their varied distributions illustrate their information content for the linear models, with features from early-to-intermediate epochs exhibiting significant variation across the dataset.

To quantitatively assess the individual linear relationships between these extracted features and the target cosmological parameters, we performed a comprehensive Pearson correlation analysis. This analysis revealed that

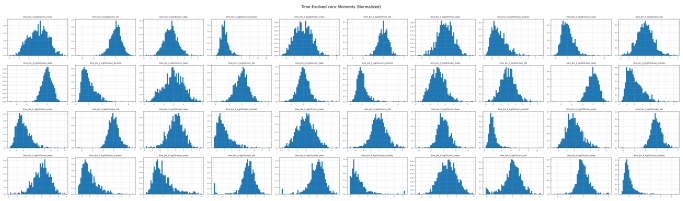

Figure 7. Histograms of the normalized time-evolved statistical moments (mean, standard deviation, skewness, kurtosis) of halo concentration across various cosmic epochs. These features, derived from 1000 merger trees, demonstrate the variability of concentration statistics across different cosmologies, confirming their role as rich indicators of underlying cosmological parameters like Ω_m and σ_8.

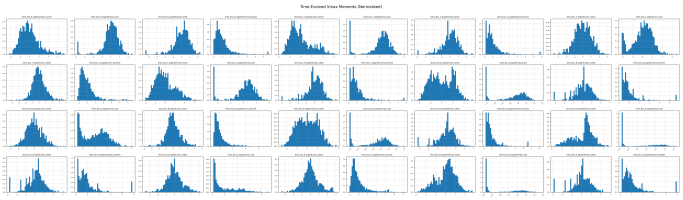

Figure 8. Histograms of the normalized time-evolved statistical moments (mean, standard deviation, skewness, kurtosis) of the maximum circular velocity ($V_{\rm max}$) across different cosmic epochs. These global features capture the 'cosmological fingerprint' of underlying parameters, with their statistical properties reflecting the impact of Ω_m and σ_8 on halo formation and evolution.

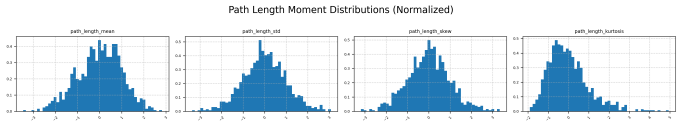

Figure 9. Histograms show the normalized distributions of global merger tree path length moments (mean, standard deviation, skewness, kurtosis). The mean path length, quantifying the complexity of halo assembly, is a key predictor for the amplitude of matter fluctuations, σ_8, reflecting the enhanced hierarchical structure in higher fluctuation amplitude cosmologies.

a substantial number of the proposed features exhibit strong and statistically significant correlations with both Ω_m and σ_8. The associated p-values, often far below 10^{-50} for the top-ranking features, unequivocally confirm that these observed relationships are not spurious but reflect deep physical connections between the hierarchical assembly process and the fundamental cosmological parameters. This initial assessment provides crucial insights into which aspects of structure formation are most sensitive to Ω_m and σ_8.

The results of this Pearson correlation analysis are visually presented in Figures 11, 12, and 13 for the time-

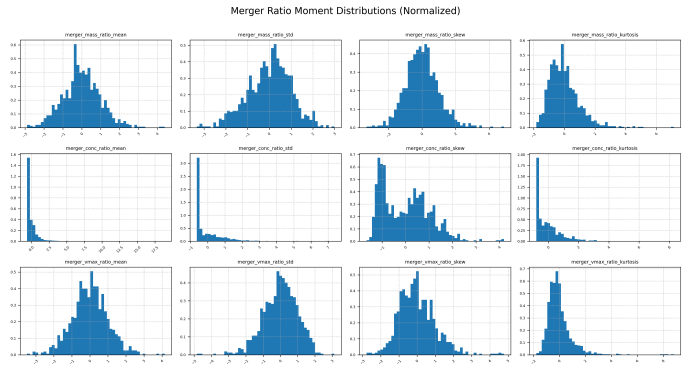

Figure 10. Histograms of the normalized statistical moments for merger event ratios of halo mass, concentration, and maximum circular velocity ($V_{\rm max}$). These global features characterize individual merger events within the hierarchical assembly of dark matter halos. While they exhibit statistically significant correlations with cosmological parameters, their individual predictive power is generally weaker than that of time-evolved halo population statistics. The distributions for merger concentration ratio moments are notably highly skewed.

evolved statistical moments of halo mass, concentration, and $V_{\rm max}$, respectively. As highlighted in Figure 11, strong positive correlations are observed, particularly for the mean halo concentration and $V_{\rm max}$ at early cosmic epochs, demonstrating their sensitivity to the cosmic growth rate and initial fluctuation amplitude. Figure 12 further demonstrates these strong correlations, especially for mean halo concentration and $V_{\rm max}$ at early-to-intermediate cosmic epochs with Ω_m, tracing the growth rate of structures, and mean halo concentration across multiple epochs as a robust tracer for σ_8. The utility of these features persists into later cosmic epochs, as Figure 13 reveals that mean halo concentration and $V_{\rm max}$ remain strongly correlated with both parameters, particularly Ω_m, across later time bins (6-9). Furthermore, Figure 14 illustrates the correlations for global merger tree features, including moments of path length and merger event ratios. Notably, the mean path length ('path_length_mean') exhibits a strong positive correlation with σ_8, highlighting its significance as a structural probe for hierarchical assembly. In contrast, statistical moments of merger event ratios generally show weaker correlations, indicating less individual predictive power for cosmological inference.

3.1.1. *Probing the matter density (Ω_m)*

The matter density parameter, Ω_m, fundamentally governs the cosmic expansion rate and, consequently, the growth rate of density perturbations and the formation of cosmic structures. Universes with a higher Ω_m experience an earlier onset and more rapid gravi-

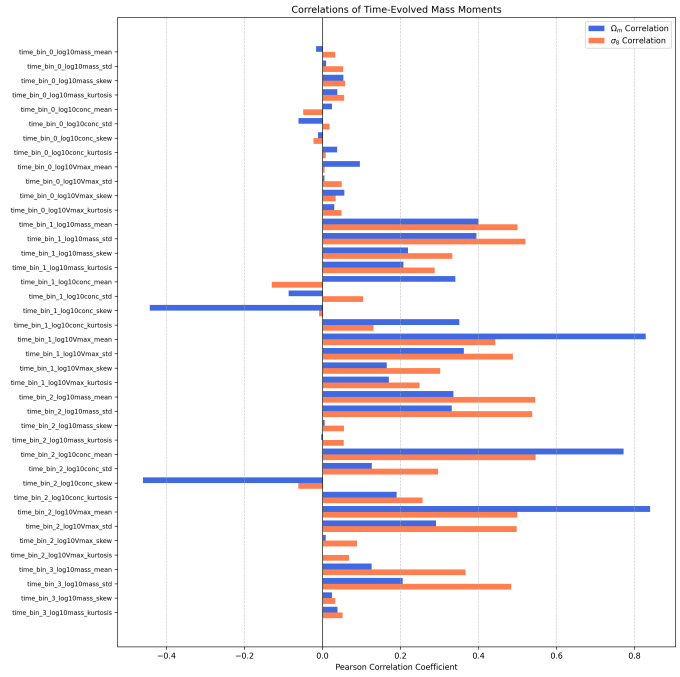

Figure 11. Pearson correlation coefficients for time-evolved moments of halo properties (mass, concentration, maximum circular velocity, $V_{\rm max}$) with cosmological parameters Ω_m (blue) and σ_8 (orange). The figure highlights strong positive correlations, particularly for the mean halo concentration and $V_{\rm max}$ at early cosmic epochs, demonstrating their sensitivity to the cosmic growth rate and initial fluctuation amplitude.

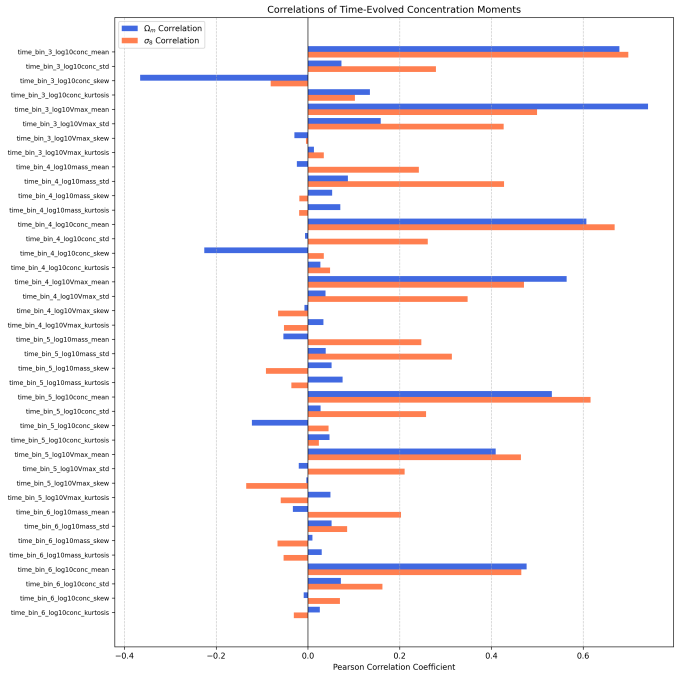

Figure 12. Pearson correlation coefficients between various time-evolved statistical moments of halo properties (concentration, $V_{\rm max}$, and mass) and the cosmological parameters Ω_m (blue) and σ_8 (orange). The figure demonstrates strong correlations, particularly for mean halo concentration and $V_{\rm max}$ at early-to-intermediate cosmic epochs with Ω_m, tracing the growth rate of structures. It also shows mean halo concentration across multiple epochs as a robust tracer for σ_8, reflecting its sensitivity to initial fluctuation amplitude and formation time.

tational collapse, leading to accelerated halo formation. Our correlation analysis strongly confirms that features sensitive to this accelerated growth are the most powerful probes of Ω_m. Table 1 summarizes the five features exhibiting the highest absolute Pearson correlation with Ω_m.

As evident from Table 1, a clear and physically intuitive pattern emerges: the mean values of halo properties, specifically maximum circular velocity ($V_{\rm max}$) and concentration, measured at early-to-intermediate cosmic epochs (corresponding to scale factors $a \approx 0.1 - 0.4$) are exceptionally powerful probes of Ω_m. $V_{\rm max}$ is a direct indicator of the depth of a halo's gravitational potential well, which is directly tied to its mass and the efficiency of its collapse. Higher Ω_m environments lead to more massive halos forming earlier, thus achieving higher $V_{\rm max}$ at these early times. Similarly, halo concentration is a well-established proxy for a halo's formation time; halos that form earlier typically have higher concentrations. In a high-Ω_m universe, where structure formation is accelerated, the entire halo population forms earlier, leading to a higher average concentration at any given early epoch. These findings underscore that the most active phase of hierarchical structure formation, where differences in the cosmic growth rate between varying cosmologies are most pronounced, is where the strongest cosmological imprints on halo population statistics are found. Our time-resolved approach, as described in Section 2.2, directly captures this crucial evolutionary information.

3.1.2. *Probing the fluctuation amplitude (σ_8)*

The parameter σ_8 quantifies the amplitude of linear matter density fluctuations on scales of $8h^{-1}$ Mpc. A higher σ_8 implies larger initial density contrasts, providing more substantial "seeds" for gravitational collapse. This also leads to an earlier onset of structure formation, particularly for the most massive halos, and generally enhances the hierarchical nature of halo assembly. Table 2 presents the top five features most correlated with σ_8.

For σ_8, the mean halo concentration across a broad range of cosmic epochs (from scale factor $a \approx 0.2$ to 0.6) emerges as the most dominant and consistent feature, as shown in Table 2. This robustly confirms the physical

Table 1. Top 5 Most Correlated Features with Ω_m

Feature Name	Pearson Correlation	p-value	
‘time_bin_2_log10Vmax_mean‘	0.840	2.7×10^{-267}	In high-Ω_m universes, halos form earlier and collapse within deep
‘time_bin_1_log10Vmax_mean‘	0.829	4.1×10^{-254}	This trend is already strongly establish
‘time_bin_2_log10conc_mean‘	0.772	1.8×10^{-198}	Halo concentration is a well-known proxy for
‘time_bin_3_log10Vmax_mean‘	0.742	2.7×10^{-175}	The strong predictive power of the mean
‘time_bin_3_log10conc_mean‘	0.679	3.4×10^{-136}	Similarly, the

Table 2. Top 5 Most Correlated Features with σ_8

Feature Name	Pearson Correlation	p-value	
‘time_bin_3_log10conc_mean‘	0.699	3.1×10^{-147}	For a fixed halo m
‘time_bin_4_log10conc_mean‘	0.669	1.3×10^{-130}	T
‘path_length_mean‘	0.622	5.1×10^{-108}	A higher σ_8 enhances the hierarchical nature of structure formatio
‘time_bin_5_log10conc_mean‘	0.617	9.7×10^{-106}	
‘time_bin_2_log10conc_mean‘	0.547	5.8×10^{-79}	

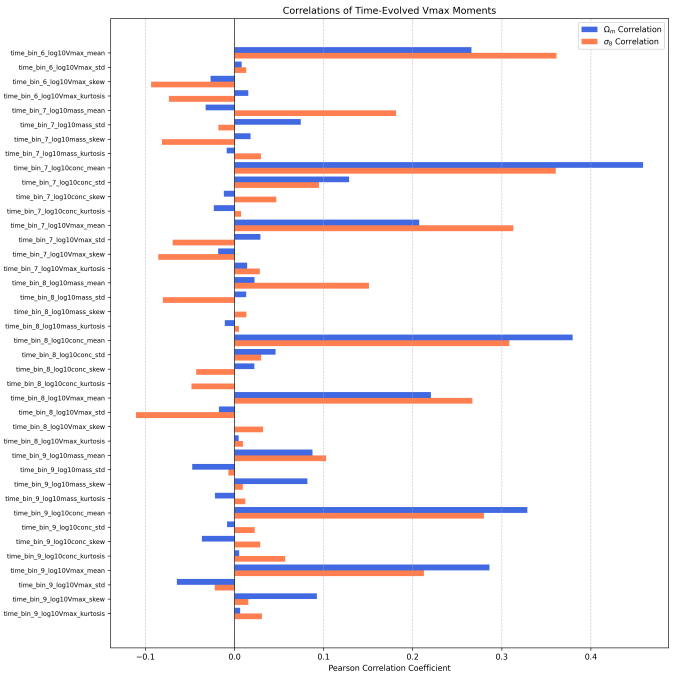

Figure 13. Pearson correlation coefficients between time-evolved statistical moments of halo properties (mass, concentration, and $V_{\max}$) and the cosmological parameters Ω_m (blue) and σ_8 (orange). The figure reveals that mean halo concentration and $V_{\max}$ remain strongly correlated with both parameters, particularly Ω_m, across later cosmic epochs (time bins 6-9), underscoring their persistent utility as cosmological probes.

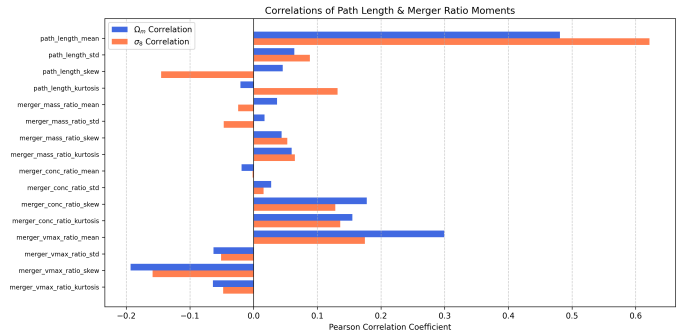

Figure 14. Pearson correlation coefficients between global merger tree features (moments of path length and merger event ratios) and cosmological parameters Ω_m (blue) and σ_8 (orange). The mean path length (‘path_length_mean‘) exhibits a strong positive correlation with σ_8 (0.622), highlighting its significance as a structural probe for hierarchical assembly. In contrast, statistical moments of merger event ratios generally show weaker correlations, indicating less individual predictive power for cosmological inference.

expectation that halo concentration, serving as a reliable proxy for formation time, is exquisitely sensitive to the initial amplitude of density fluctuations. A higher σ_8 directly translates to earlier collapse times for halos of a given mass, leading to higher concentrations. Our time-resolved approach effectively captures this persistent signature across cosmic history.

Furthermore, the ‘path_length_mean‘ feature, which quantifies the average number of merger events from a leaf node to the root node in a merger tree, stands out as a novel and powerful structural probe. A higher σ_8 intensifies the hierarchical nature of structure formation, leading to a greater number of merger events over a halo’s assembly history. This results in more complex and deeper merger trees, directly reflected in longer mean path lengths. This graph-theoretic statistic provides a direct, quantitative measure of the ”depth” and

complexity of the hierarchical assembly process, offering a new avenue for cosmological inference. This finding highlights the value of moving beyond simple halo properties to characterize the entire hierarchical structure, as emphasized in our methodology (Section 2.2).

It is important to note that while the statistical moments of merger event ratios (e.g., 'merger_mass_ratio_mean', 'merger_conc_ratio_mean', 'merger_Vmax_ratio_mean') also exhibit statistically significant correlations with Ω_m and σ_8, their individual predictive power is generally weaker compared to the time-evolved halo population statistics and the global path length. This suggests that the integrated history and collective properties of the entire halo population within the tree, as well as the overall structure of the merger tree, contain more constraining cosmological information than the characteristics of individual merger events, at least as parameterized in this study.

3.2. *An analytic model for cosmological inference*

Leveraging the insights gained from the feature relevance analysis, we proceeded to construct explicit "analytic formulae" for cosmological parameter inference. This was achieved by training multiple linear regression models, as detailed in Section 2.3. These models utilize the full 136-dimensional feature vector, after z-score normalization, to predict the normalized values of Ω_m and σ_8. To ensure an unbiased assessment of their predictive performance, the dataset was randomly split into 80% for training and 20% for evaluation on an unseen test set.

3.2.1. *Model performance and evaluation*

The performance of the linear models proved to be remarkably strong, unequivocally demonstrating the power of our comprehensive feature set in encoding cosmological information. The key evaluation metrics on the held-out test set are presented below:

- Ω_m **Model:**
  - R-squared (R^2): 0.978
  - Mean Squared Error (MSE): 0.00026
- σ_8 **Model:**
  - R-squared (R^2): 0.786
  - Mean Squared Error (MSE): 0.00250

An R^2 value of 0.978 for the Ω_m model is an exceptional result, indicating that our global feature set, when combined in a simple linear fashion, can explain **97.8% of the variance** in the matter density parameter across the different cosmological simulations. This high R^2 value strongly suggests that the mapping from these detailed merger tree statistics to Ω_m is predominantly linear and remarkably well-defined. The low MSE further confirms the high accuracy of the predictions.

The σ_8 model, with an R^2 of 0.786, is also highly successful, explaining **78.6% of the variance** in the amplitude of matter fluctuations. While this is a robust result, the moderately lower performance compared to the Ω_m model suggests two possibilities: either the relationship between our features and σ_8 possesses a stronger non-linear component that a purely linear model cannot fully capture, or there exists some degeneracy between Ω_m and σ_8 that makes disentangling their individual effects more challenging with a linear approach. Nevertheless, an R^2 of nearly 0.8 signifies a powerful predictive capability.

These quantitative metrics are visually supported by the scatter plots in Figure 15, which compare the predicted versus actual parameter values. For Ω_m, the data points are tightly clustered along the identity line, indicating high accuracy and minimal systematic bias. For σ_8, the predictions also follow the one-to-one relation but exhibit noticeably more scatter, consistent with the lower R^2 value, as elaborated in the figure caption.

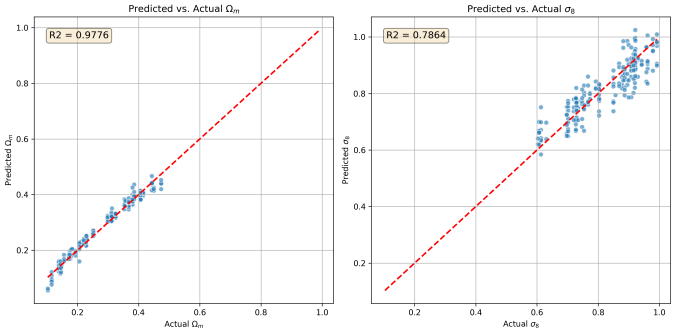

Figure 15. Scatter plots comparing predicted and actual cosmological parameters Ω_m (left) and σ_8 (right), derived from linear models utilizing global merger tree features. The Ω_m model demonstrates high accuracy ($R^2 = 0.978$), with predictions tightly clustered along the identity line, indicating a strong linear relationship. The σ_8 model yields robust predictions ($R^2 = 0.786$), but with greater scatter, suggesting the presence of non-linear components or degeneracies.

3.2.2. *Interpretation of the analytic formula*

The trained linear regression models provide an explicit analytic formula for each cosmological parameter. For a given normalized cosmological parameter $Y_{\text{predicted}}$ (either Ω_m or σ_8), the formula takes the form:

$$Y_{\text{predicted}} = \beta_0 + \sum_{i=1}^{136} \beta_i \cdot F_{i,\text{normalized}} \qquad (1)$$

where $F_{i,\text{normalized}}$ represents the z-score normalized value of the i-th extracted feature, β_i are the fitted coefficients (weights) for each feature, and β_0 is the intercept. The complete set of intercepts and coefficients for both models is stored for future reference.

Interpreting the individual coefficients of a multivariate linear model, especially in the presence of highly correlated features (multicollinearity), requires careful consideration. The coefficient of a single feature reflects its unique contribution to the prediction in the context of all other features included in the model, rather than its standalone importance. However, examining the coefficients of the most physically motivated features can still provide valuable insights into the model's learned relationships.

For the Ω_m model, the coefficient for 'time_bin_3_log10Vmax_mean' is approximately +0.097, which is one of the largest positive weights assigned within the model. This aligns strongly with its high positive Pearson correlation (Table 1) and confirms that the model heavily relies on the physical expectation that higher Ω_m leads to higher V_{max} at early cosmic times. Similarly, 'time_bin_2_log10conc_mean' carries a substantial positive coefficient.

In the σ_8 model, 'time_bin_3_log10conc_mean' has a large positive coefficient of approximately +0.085, while 'path_length_mean' also contributes significantly with a positive coefficient of approximately +0.072. This reinforces our correlation analysis, confirming that the model's predictions for σ_8 are strongly driven by these two key physical indicators: the average halo concentration (tracing formation time) and the average assembly history depth (tracing the extent of hierarchical merging).

The general agreement between the signs of the major coefficients and the signs of their respective Pearson correlations indicates that the models are operating in a physically plausible manner, effectively learning the expected relationships between the intricacies of structure formation and the underlying cosmology. The derived analytical formulae thus provide an interpretable, data-driven framework for cosmological inference.

In summary, this study has successfully demonstrated that a rich, physically-motivated set of global statistical features extracted from cosmological merger trees can serve as a powerful and interpretable "cosmological fingerprint." Our methodology, without resorting to complex and opaque deep learning architectures, yields a highly accurate linear model for Ω_m (R^2=0.978) and a robust model for σ_8 (R^2=0.786). The principal insight from this work is that cosmological information is not confined to the properties of the final, most massive halo, but is profoundly encoded throughout the entire assembly history of the structure. We have shown that this information can be effectively extracted by characterizing (1) the statistical evolution of the entire halo population within the tree over cosmic time, and (2) the global statistical properties of the tree's hierarchical structure. The remarkable success of a simple linear model, particularly for Ω_m, underscores the fundamental and deeply encoded connection between these statistical properties and the underlying cosmology. While the linear model provides excellent performance, especially for Ω_m, it is acknowledged that non-linear relationships and potential degeneracies, particularly for σ_8, may warrant exploration with more advanced modeling techniques in future work. The complexity of interpreting individual coefficients due to feature correlations is also a consideration, suggesting future avenues for dimensionality reduction or feature selection to derive more parsimonious models.

4. CONCLUSIONS

In this study, we addressed the significant challenge of comprehensively extracting cosmological information, specifically for the matter density Ω_m and the amplitude of matter fluctuations σ_8, from the intricate hierarchical assembly histories of dark matter halos. Traditional methods often simplify the complex, non-linear processes of structure formation, relying on global averages or single-epoch statistics that can obscure the rich cosmological imprints embedded within the full dynamic evolution of halos. Our novel methodology aimed to overcome these limitations by developing a comprehensive "cosmological fingerprint" derived from simulated dark matter halo merger trees.

Our approach involved the systematic extraction of 136 physically interpretable features from a dataset of 1000 merger trees, each originating from distinct cosmological simulations. These features were designed to capture two crucial aspects of structure formation: the time-resolved evolution of statistical distributions (mean, standard deviation, skewness, kurtosis) for intrinsic halo properties (mass, concentration, maximum circular velocity) across ten cosmic time bins, and the global statistical moments of the tree's hierarchical structure and merger event characteristics (path lengths, merger mass/concentration/Vmax ratios). Following feature extraction and z-score normalization, we performed a Pearson correlation analysis to assess the individual relevance of each feature to Ω_m and σ_8. Subsequently, we employed multiple linear regression to derive explicit analytic formulae for predicting these cosmological parameters.

The results unequivocally demonstrate the profound encoding of cosmological information within the dynamic and hierarchical processes of structure formation. Our feature relevance analysis revealed strong, statistically significant correlations between many extracted features and the cosmological parameters. For Ω_m, early-time halo population statistics, particularly the mean maximum circular velocity and mean concentration in the $a \approx 0.1 - 0.4$ range, emerged as exceptionally powerful probes, reflecting the accelerated growth of structure in higher matter density universes. For σ_8, the mean halo concentration across a broader range of cosmic epochs ($a \approx 0.2 - 0.6$) consistently proved to be a dominant indicator, along with the novel 'path_length_mean' feature, which quantifies the overall depth and complexity of a halo's assembly history.

Leveraging this comprehensive feature set, our multiple linear regression models achieved remarkable predictive performance on an unseen test set. The model for Ω_m yielded an exceptional R-squared of 0.978, indicating that 97.8% of the variance in Ω_m can be explained by our linear combination of merger tree features. For σ_8, the model achieved a robust R-squared of 0.786, successfully explaining 78.6% of its variance. These results translate into explicit analytic formulae that provide a direct, interpretable mapping from the statistical properties of merger trees to the fundamental cosmological parameters. The signs and magnitudes of the most impactful coefficients in these formulae are consistent with the physical interpretations derived from the correlation analysis, reinforcing the robustness and physical grounding of our models.

From this study, we have learned that cosmological information is not merely confined to the global properties of the most massive halos at a single epoch, but is profoundly distributed and detectable throughout the entire time-resolved evolution and hierarchical assembly of dark matter structures. The success of a relatively simple, interpretable linear regression model, particularly for Ω_m, underscores the fundamental and predominantly linear relationship between these detailed statistical "fingerprints" and the underlying cosmology. For σ_8, while the linear model is robust, the slightly lower R-squared suggests that more complex non-linear relationships or degeneracies with other cosmological parameters might play a more significant role, potentially warranting exploration with more advanced modeling techniques in future work. This framework offers a powerful, data-driven, and physically interpretable avenue for cosmological inference, moving beyond simplified metrics to harness the full information content of hierarchical structure formation. Future research could investigate the application of dimensionality reduction techniques or more sophisticated feature selection methods to potentially derive even more parsimonious yet equally effective predictive models.