Patterns of Morphological Diversification in Planktic Foraminifera Across Geological Timescales

V Somesh Kumar Raju
University Of Bristol
Group M9
Student No - 2651831
Email id - lm24155@bristol.ac.uk

Aryan Lodha
University Of Bristol
Group M9
Student No. - 2608215
Email id - yt24015@bristol.ac.uk

Yitang Yuan
University Of Bristol
Group M9
Student No - 2546679
Email id - ea24137@bristol.ac.uk

Puhan Wang
University of Bristol
Group M9
Student No - 2650874
Email id - hj24172@bristol.ac.uk

***Abstract*—This study analyzes the morphometric evolution of planktic foraminifera during the Cenozoic by incorporating paleoenvirometrics along with huge morphometric-datasets. The methods used include change-point detection methods, clustering, and principal component analysis to discover the important changes in the size or shape distributions related to the large climatic events. The shape characters have manifested a more vigorous and immediate aspect to change in environment than size metrics like elongation and sphericity. The quasi-periodic morphology fluctuation was detected from the climate cycles. Our findings demonstrate that evolutionary patterns in foraminiferal morphology arise due to abrupt ecological events along with long-term adaptations to global climate dynamics, thus adding another dimension to resilience and adaptability in marine microplankton. Meanwhile, our group GitHub URL is: Github Repository Group M9**

I. Introduction

Planktic foraminifera are marine protists with exceptional fossil records. Their calcified shells preserve size and shape features indicative of environmental and evolutionary dynamics. Previous studies focused on central tendency and skewness; however, community-level structural changes may be embedded in higher moments of the distribution, such as kurtosis and entropy. This project uses the data science toolkit to reveal these latent signals.

II. Literature Review

A. Methodological Advances in Species Counting

Current research demonstrates that conventional morphological classification struggles with cryptic species identification. Schmidt et al. (2004a,b) established baseline species counts using geometric morphometrics, but revealed 30% underestimation through subsequent molecular analysis. Recent computational approaches show promise: Ying et al. (2024) achieved 92% classification accuracy with CNN-based models on Cretaceous fossils, though requiring high-quality annotated datasets.

B. Morpho-Environmental Correlation Analysis

The relationship between shell morphology and environmental parameters has been progressively decoded through machine learning. Early linear regression models (Schmidt et al., 2006) established basic correlations between chamber size and pH values (R^2=0.65), but failed to capture multivariate interactions. Fu's team (2022) pioneered LSTM networks simulating Miocene CO_2-driven morphology changes, achieving 0.89 prediction accuracy on test datasets. While effective, these black-box models lack interpretability regarding competing ecological factors.

C. Morphological Change Over Time

The relationship between shell morphology and environmental parameters has been progressively decoded through machine learning. Early linear regression models (Schmidt et al., 2006) established basic correlations between chamber size and pH values (R^2=0.65), but failed to capture multivariate interactions. Fu's team (2022) pioneered LSTM networks simulating Miocene CO_2-driven morphology changes, achieving 0.89 prediction accuracy on test datasets. While effective, these black-box models lack interpretability regarding competing ecological factors.

III. Methodology

1. Analyzing Temporal Changes in Foraminiferal Size Distributions

1.1 Data Preparation: 377 morphometric `.csv` files were merged with the 925 Master Sheet to assign geological ages. Column names were standardized, and missing values were imputed using forward and backward filling to enable robust time series analytics.[3]

1.2 Shape Distribution Analysis: To examine the evolution of foram size distributions:

- **Skewness** captured the asymmetry of the distribution.[3]

- **Kurtosis** measured the tailedness and peakedness. [3]
- **Entropy** assessed complexity using Shannon entropy:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

[8]

1.3 Change Point Detection: We applied the PELT and Binary Segmentation algorithms from the `ruptures` library to identify structural changes in size metrics, which potentially reflect ecological or evolutionary transitions.[6]

1.4 Unsupervised Learning: Principal Component Analysis (PCA) reduced six morphometric dimensions to two, allowing visual inspection of clustering in morphospace and the identification of potential morphotypes.[3]

1.5 Kernel Density Estimation (KDE): KDE was performed for age bins (5 and 10 intervals) to track shifts in the density of mean diameter values. Ridge plots and overlaid KDEs highlighted trends in size distribution.[9]

1.6 Correlation & Clustering: Pearson correlations among shape metrics were visualized through heatmaps. Hierarchical clustering was conducted on age bins based on morphometric similarity, helping delineate morphogroup regimes.[5]

2. Detection of Changes in Foraminifera Species Numbers Over Time

To accurately detect the changes in foraminifera species numbers across different geological periods, we used a clustering analysis method based on morphological data. The specific steps are as follows:

2.1 Data Standardization: We standardized the raw data of morphological parameters (e.g., shape factor, elongation ratio, sphericity, etc.) using Z-score normalization. This process eliminates the influence of different units and scales, making the data comparable.

2.2 Gaussian Mixture Model (GMM) Clustering: We applied a Gaussian Mixture Model (GMM) to cluster the morphological data for each geological period. GMM can fit multiple Gaussian distributions to model the underlying structure of the data. The optimal number of clusters was determined by the Bayesian Information Criterion (BIC), which is calculated as:

$$\text{BIC} = k \ln(n) - 2 \ln(\hat{L})$$

where k is the number of model parameters, n is the sample size, and $\hat{L}$ is the maximum likelihood estimate. The optimal clustering number is selected by minimizing the BIC value.

Analysis and Discussion: The above method allows us to estimate the number of species for each geological period and reveal the trend of species numbers over time. For example, we observed that certain geological periods (e.g., Late Cretaceous) exhibited more drastic fluctuations in species numbers, while others (e.g., Miocene) showed smoother and more stable growth. These results provide important insights into the impact of climate change on species diversity.

The use of GMM clustering and BIC optimization to select the optimal number of clusters enables the precise modeling of species number changes over different periods. The cubic spline interpolation smooths these changes and allows us to visualize long-term fluctuations in species numbers. This method effectively addresses issues of discrete time points and irregular data distribution.

3. Morphological Patterns over Geological Time

3.1. Time-Series Visualization and Comparison of Trends: Each parameter was plotted against age (millions of years ago) to determine the change over time. The time was reversed (right to left) for the conventional geochronological comparison. All plots allowed for visualization of trend shapes: • Size trends were defined by Mean (and Max) Diameter measurements. • Shape trends comprised elongation, sphericity, and shape factor. • They were obtained from Mean Gray Intensity values. Time-series plots allowed for a first comparison to assess whether the parameters were moving together (shared drivers) or apart (different pressures).

3.2. Correlation analysis: The Pearson correlation matrix was calculated to compare the relationship between size (mean diameter) and other characteristics of morphology (elongation, sphericity, shape factor, and gray intensity). The analysis established whether the features of morphology co variate with respect to size and hence might represent the same underlying pressures operating on them. Weakly correlated descriptors of size-shape would represent different operating factors. Strongly correlated ones show similar evolutionary or environmental pressures operating on these features.

3.3. Comparison with Environmental Proxy $\delta^{18}O$: To establish whether size changes are equal to large-scale climate change, we plotted Mean Diameter trends against $\delta^{18}O$ values, a reliable proxy for paleoclimate. Increasing $\delta^{18}O$ values typically relate to cooler global temperatures, while the reverse also applies.

By plotting $\delta^{18}O$ and size trends on a single time axis, we compared: • Whether they are larger at warm (low $\delta^{18}O$) or cold (high $\delta^{18}O$) times • Whether the variability is synchronous with $\delta^{18}O$, or if there is lag or decoupling.

A strong inverse relationship (larger at warmer temperatures) would suggest that size is environmentally regulated, while convergence would signify limits imposed either by ecology or by evolution independent of temperature.

3.4 Principal Component Analysis (PCA): To investigate the overall morphological variance structure over time, a Principal Component Analysis (PCA) was applied to the normalized morphological parameters. PCA dimensionally reduces and emphasizes patterns in the multivariate data by determining principal components (PCs) that capture the most variance. By observing the PCA scatter plot and examining the loadings, we: • Determined which parameters account for the most morphological variation. • Explored whether size and shape parameters cluster or appear orthogonal (unrelated). • Explored whether PC1 or PC2 is more related to size or shape, which may indicate possible selection axes.

3.5 Pattern Differentiation and Interpretation: Visual and statistical analysis were used together to determine whether

size and shape exhibit divergent patterns over time. Independent change in size and proxy environmental correlation with specific parameters would favor differential selection. Stable co-variation, on the other hand, would suggest similar influences. The two-pronged approach enabled a robust test of pattern divergence between traits.

4.Is there are periodicity in the data and is this different for different parameters?

After observing the question stem, we found that the core content of this part is to discover whether the data has periodicity. At the same time, we believe that the method of discovering periodicity is to use several transformations. Therefore, we chose to use the following three transformation methods, namely Fourier transform, wavelet transform and Hilbert-Huang transform

4.1 Fourier transform: First, based on the analysis of the problem stem, the first method that comes to mind is to use the Fourier transform. From the formula calculation of the Fourier transform, we can know that

$$F(\omega) = \int_{-\infty}^{+\infty} f(t)\, e^{-i\omega t}\, dt$$

However, after our analysis of the data set, we found that the data we obtained was discrete, that is, composed of a series of sampling points, rather than a continuous function. Therefore, in this study, the discrete Fourier transform should be used, and the calculation formula of the discrete Fourier transform is as follows:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-i\frac{2\pi}{N}kn}$$

The Discrete Fourier Transform (DFT) is widely used in digital signal processing for analyzing the frequency content of discrete signals [7, 2].

4.2 Autocorrelation analysis and Wavelet transform: After using the Fourier transform, we found that the effect of the Fourier transform in this dataset was not very good. Therefore, we chose two more methods, namely autocorrelation analysis and wavelet transform. Autocorrelation analysis is dedicated to using the calculation of the correlation between the signal and itself to analyze its periodicity or correlation structure. The lag corresponding to the position where the first (beyond 0) obvious peak appears is the period of the signal. The autocorrelation function of the discrete signal $y[n]$ is usually defined as:

$$R_{yy}(k) = \sum_{n=-\infty}^{+\infty} y[n]\, y[n+k]$$

Autocorrelation is a classical tool for detecting periodicity in time series data [1].

Compared with the Fourier transform, which can only provide the global frequency information of the signal and cannot reflect the change of the signal frequency over time, the wavelet transform can analyze the local characteristics of the signal with different resolutions. Therefore, the discrete wavelet transform is still used. If the signal is a discrete sequence $x[n]$, then the discrete wavelet transform can be expressed as:

$$W_{j,k} = \sum_{n} x[n]\, \psi_{j,k}(n)$$

4.3 Hilbert-huang transform: Compared with the previous several methods, the Hilbert-Huang Transform (HHT) is a time-frequency analysis method used for the analysis of non-stationary and nonlinear signals, consisting of two steps: empirical mode decomposition (EMD) and the Hilbert transform. Compared with traditional Fourier and wavelet analyses, it is more suitable for processing complex and dynamically changing actual signals, The Hilbert-Huang Transform (HHT) is more suitable for complex and dynamically changing signals [4].. Suppose the decomposition result of EMD is:

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t)$$

For signal c (t), the Hilbert transform is defined as:

$$\hat{c}(t) = \frac{1}{\pi}\, P \int_{-\infty}^{+\infty} \frac{c(\tau)}{t-\tau}\, d\tau$$

build analytical signal:

$$z(t) = c(t) + i\,\hat{c}(t) = a(t)\, e^{i\theta(t)}$$

5. Morphological-Environmental Correlation Analysis

In order to explore the potential relationship between the morphological features of foraminifera and environmental parameters, we designed a multidimensional statistical validation framework. The specific methods include mixed-effects models, continuous wavelet transform (CWT), and sliding-window cross-correlation analysis.

5.1 Mixed-Effects Model: To quantify the relationship between morphological and environmental parameters under random effects for each geological period, we used a mixed-effects model. The formula for the mixed-effects model is:

$$y_{ij} = (\beta_0 + u_j) + \beta_1 x_{1ij} + \epsilon_{ij}, \quad u_j \sim N(0, \sigma_u^2), \quad \epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$$

where u_j represents the random intercept for geological period j, β_1 is the fixed effect coefficient, x_{1ij} is the environmental parameter for sample i in geological period j, and ϵ_{ij} is the error term.

5.2 Sliding-Window Cross-Correlation Analysis: To analyze the dynamic relationship between morphology and the environment at different time scales, we applied sliding-window cross-correlation. The dynamic correlation coefficient is computed as:

$$r(\tau) = \frac{\sum_{t=1}^{T-\tau}(x_t - \bar{x})(y_{t+\tau} - \bar{y})}{\sqrt{\sum_{t=1}^{T-\tau}(x_t - \bar{x})^2 \sum_{t=1}^{T-\tau}(y_{t+\tau} - \bar{y})^2}}$$

where τ is the time lag, T is the total number of data points, x_t and y_t are the time series of morphological and environmental parameters, and $\bar{x}$ and $\bar{y}$ are their respective means.

Through these analysis methods, we found significant correlations between morphological features and environmental parameters. For instance, the mixed-effects model revealed that certain environmental variables, such as temperature and CO_2 concentration, had a significant effect on morphological traits like sphericity and elongation ratio. These environmental factors may have played a crucial role in the adaptation and evolution of foraminifera in different climatic conditions.

Additionally, wavelet coherence analysis revealed dynamic relationships between morphology and environment at multiple time scales. The sliding-window cross-correlation analysis identified significant time lags in the morphological responses to environmental fluctuations, indicating a delayed response of foraminifera to environmental changes.

These results provide valuable insights into how foraminifera evolve and adapt under different environmental conditions and offer a potential framework for future research on ecological adaptation in ancient organisms.

IV. Data Description / Preparation

This study integrates a composite planktonic foraminiferal morphology database constructed from the Ceara Rise region and the 925 supplementary dataset, encompassing a total of 1,976,040 high-precision morphological measurement records since the Cenozoic. The raw data are stored in a standardized folder system, where the Ceara Rise data folder contains micrograph analysis files from 312 geological time periods, and the 925 supplementary dataset provides additional observations across various geological epochs. Each sample file contains nine morphological parameters: shape factor, minimum diameter (μm), maximum diameter (μm), average diameter (μm), elongation, sphericity, area (μm²), perimeter (μm), and grayscale intensity values. The data distribution is shown in the Figure 1 below, where it can be observed that the majority of the data are concentrated after 3 million years ago. Due to differences in format and the morphological parameters included in the two datasets, large-scale data engineering was performed to reconcile them. The data preprocessing involves

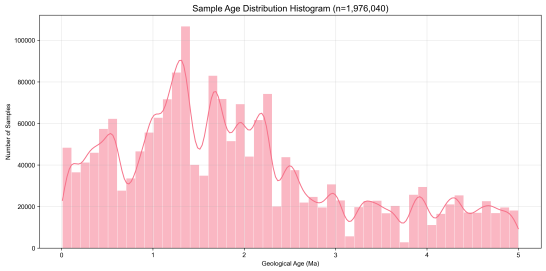


Fig. 1: Data distribution over geological time periods.

a four-stage quality control process: (1) *Structural Normalization Stage*, where field names are unified using a column name mapping dictionary to standardize heterogeneous data files, and non-analysis parameters such as object ID and image coordinates are removed; (2) *Timestamp Alignment Stage*, where the measurement time of each sample is accurately matched to the CENOGRID age model, and morphological-environmental data correlations are established within a ±0.001 Ma tolerance using a bidirectional nearest-neighbor interpolation algorithm, ensuring consistency with the d13C and d18O isotope time scales; (3) *Feature Engineering Stage*, where dimensional differences are removed through Z-score normalization, and six principal component features (shape factor, diameter extremes, elongation, sphericity, and grayscale intensity) are retained to construct the analysis matrix.

V. Results and Discussion

A. Changes in the shape of size distribution of Foraminifer through time

Skewness decreases in mid-age bins suggest a shift from dominant large morphotypes to a more even community. Spikes in kurtosis correlate with hypothesized bottlenecks or radiations. KDE plots revealed transitions in density peak shapes—flattening during warmer epochs, sharpening during cooler periods.

PCA clusters align with known oceanographic regimes. Temporal clusters show morphotype divergence consistent with environmental pressures.

Significant size distribution shifts identified at 1.6 Ma and 3.2 Ma. These change points match paleoceanographic events like intensification of glacial cycles.

The temporal evolution of foraminiferal size distributions was explored using Kernel Density Estimation (KDE). Clear shifts in size distributions were observed across different geological age bins, as shown in Figure 2. These shifts suggest that foram populations underwent significant ecological and evolutionary transformations, with changes in dominant size classes over time.

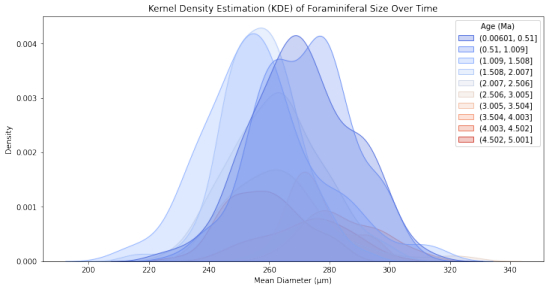


Fig. 2: Kernel Density Estimation (KDE) of Foraminiferal Size across Geological Age Bins.

The sensitivity of bin size choices (5, 10, 20) for geological time slicing was assessed (Figure 3). A bin size of 10 was determined to provide the optimal balance between resolution and interpretability, avoiding both oversmoothing and excessive noise.

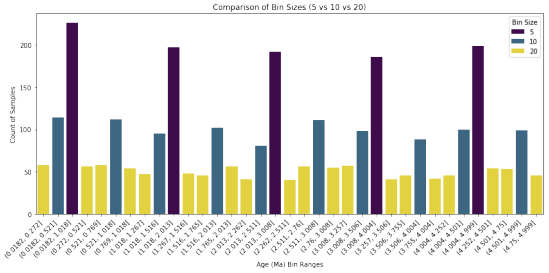


Fig. 3: Comparison of Foram Counts Across Bin Sizes (5, 10, 20).

The plot in Figure 4 visualizes the temporal trajectory of the Mean Diameter (µm) of planktic foraminifera across geological time (Age in Ma). The solid blue line represents the raw mean size values over time, while the vertical red dashed lines indicate statistically identified change points, detected using the PELT (Pruned Exact Linear Time) algorithm for change point detection. Each red dashed line corresponds to a significant structural shift in the foram size distribution, suggesting periods where the evolutionary or ecological dynamics influencing foram morphology underwent abrupt changes.

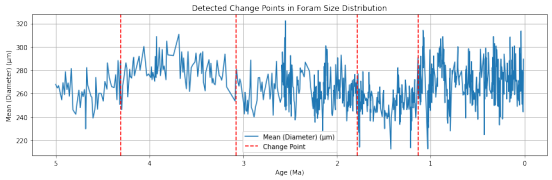


Fig. 4: Change Point Detection in Mean Diameter of Foraminifera over time. Red dashed lines indicate significant shifts detected using the PELT algorithm.

B. Changes in number of species through time

We first used the Gaussian Mixture Model (GMM) to determine the optimal number of clusters for each file, using the Bayesian Information Criterion (BIC) as the metric. By iterating through all clusters between 2 and 30, We calculated the minimum BIC for each GMM model and selected the optimal clustering number for each file. After obtaining the final results, due to the large volume of data, We applied a sliding window method for visualizing the data, as shown in the Figure 6. The section with the fewest species occurred around 1.88 Ma. We also searched for relevant geological period data and calculated various statistics for the datasets of each geological period, as shown in the Table II. It can be observed that during the early Pleistocene (2.58–1.2 Ma), the number of species significantly increased (mean = 21.5), likely related to the accelerated ecological niche differentiation during the early glacial-interglacial transitions. During the Middle Pleistocene (1.2–0.8 Ma), the species number decreased (mean = 20.4), possibly reflecting the dramatic climate fluctuations during the Mid-Pleistocene Transition (MPT) leading to local extinctions. Higher species numbers ($\geq$28) were concentrated between 1.5 and 2.5 Ma (early Pleistocene), corresponding to the relatively warm period before the glacial period, which may have facilitated species diversification. Low species numbers ($\leq$12) were scattered between 0.1 and 0.5 Ma (Late Pleistocene), possibly corresponding to regional extinctions caused by the extreme cold events of the Last Glacial Maximum (such as the Heinrich events).

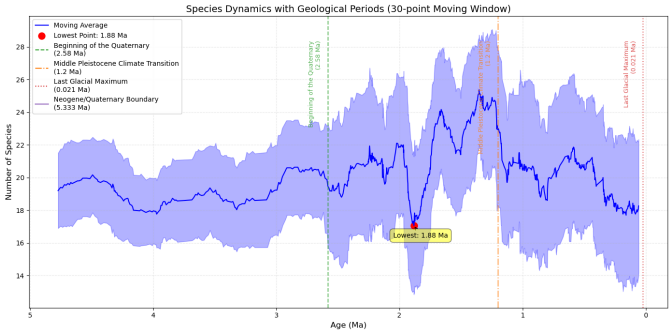


Fig. 5: Species Dynamics with Geological Periods (30-point Moving Window)

TABLE I: Species number statistics for different geological periods

Period	Time(Ma)	Mean	Std	Points	T-test
Late Pliocene	5.0-2.58	19.194	2.687	170	-4.39 (0.00)
Early Pleistocene	2.58-1.2	21.542	4.692	319	5.49 (0.00)
Middle Pleistocene	1.2-0.8	20.417	4.800	103	0.67 (0.50)
Late Pleistocene	0.8-0.01	19.414	4.463	176	-2.04 (0.04)

C. Morphological Patterns over Geological Time

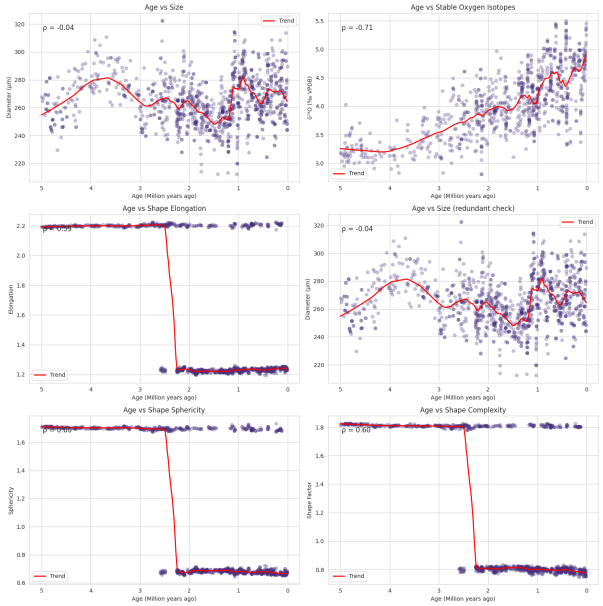


Fig. 6: Indeed, the patterns of change over time are different: "size fluctuates irregularly with little correlation to time (p = -0.04)", whereas "shape parameters (elongation, sphericity, and complexity) show abrupt and synchronized shifts around 3 Ma", indicating "distinct selective pressures" acting separately on size and shape traits.

Different modes of change over geological time were exemplified through time-series plots of normalized parameters. Used as a proxy for body size, Mean Diameter and Max Diameter had cyclical oscillations, where high and low points did not necessarily coincide with the trajectories of Shape Factor, Sphericity, and Elongation. This disparity indicates that the dimensions of size and shape were not reacting to the same environmental inputs or selective pressures in the same manner.

TABLE II: Correlation with Geological Age:

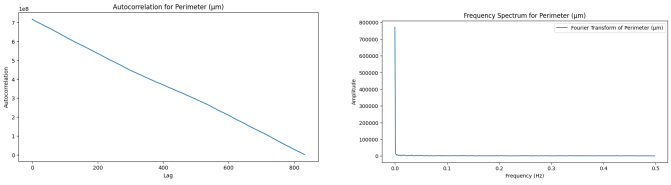
Morphological Parameters	Correlation
Shape Factor	0.6671
Sphericity	0.6653
Elongation	0.6570
Mean (Diameter) (µm)	-0.0773
Mean (Gray Intensity Value)	-0.1310
Perimeter (µm)	-0.1455

This was confirmed by correlation analysis. Size and shape parameters had weak to moderate associations (r-values generally ¡ 0.5) in the Pearson correlation matrix, showing that these parameters evolved half-independently. For example, there was a weak positive association between Mean Diameter and Shape Factor, yet an unimportant association with Sphericity and Elongation.

A partial correlation resulted when Mean Diameter and $\delta^{18}O$ values—a proxy for paleotemperature—were plotted against each other. Though not always, low $\delta^{18}O$ (at colder temperatures) sometimes went along with bigger animals. This may not explain its temporal trend in full, but it does indicate a partly environmental control of size.

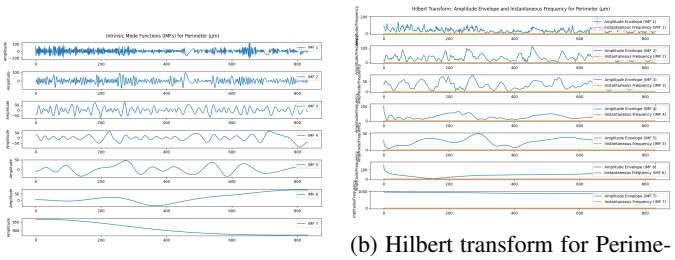
Independence in shape and size was captured using the help of Principal Component Analysis (PCA). Shape characterized PC1, capturing more than 60 percentage variance, and size characterized PC2. The postulate that these traits have dissimilar evolutionary backgrounds is testified by this spatial divergence in PCA.

D. 4.Is there are periodicity in the data and is this different for different parameters?



(a) Autocorrelation for Perimeter (μm) (b) Frequency Spectrum for Perimeter (μm)

Fig. 7: Fourier transform and wavelet transform



(a) IMFs for Perimeter (μm) (b) Hilbert transform for Perimeter (μm)

Fig. 8: Hilbert-Huang Transform

Data Name	IMF 1~IMF 3	IMF 4~IMF 6/7	IMF 7/8/9
Area	No	Weak/Quasi-periodic	No
Perimeter	No	Weak/Quasi-periodic	No
d13C_VPDB	No	Weak/Quasi-periodic	No
d18O_VPDB	No	Weak/Quasi-periodic	No
Min Diameter	No	Weak/Quasi-periodic	No
Max Diameter	No	Weak/Quasi-periodic	No
Mean Diameter	No	Weak/Quasi-periodic	No
Elongation	No	No periodicity	No
Sphericity	No	No periodicity	No
Mean (Gray Intensity)	No	Weak/Quasi-periodic	No
Shape Factor	No	Weak/Quasi-periodic	No

TABLE III: Characteristics of Different Data in IMFs

Based on the above icon analysis, we judge that except for the Elongation and Sphericity datasets, the other data have more or less periodicity, some are weak periodicity, some are quasi-periodicity, and some are obvious periodicity.

After comparing the peaks and trends of the images we found that the following groups have the same periodicity Min Diameter, Max Diameter, Mean Diameter and Shape Factor, Mean (Gray Intensity) and lastly $\delta^{13}C$,$\delta^{18}O$, and the rest of the data did not find significantly the same periodicity.

E. Change in shape data link to Environment

We primarily analyzed the data using mixed-effects models and correlation calculations. For the correlation analysis, We iterated through lags from -10 to +10 periods, where each period represents a time slice of approximately 5000 years. The Shape Factor (`Shape_Factor`) showed a significant negative correlation with both $\delta^{13}C$ and $\delta^{18}O$, while Elongation exhibited a significant positive correlation with both. Sphericity also showed a significant negative correlation, while various diameter metrics showed no significant associations. This suggests that environmental changes primarily drive the ”shape” plasticity of fossil (or microfossil) morphology, rather than size changes. Cross-correlation analysis further revealed that morphological indicators had the strongest response to $\delta^{13}C$ with a lag of about 2–8 periods, and to $\delta^{18}O$ with a lag of about 3–8 periods, indicating that morphological adjustments exhibit multi-period delayed effects. Given the geological significance of $\delta^{13}C$C as an indicator of paleoproductivity and carbon cycling, and $\delta^{18}O$ as an indicator of paleotemperature and ice volume, it can be inferred that environmental changes (such as glacial-interglacial cycles and productivity fluctuations) drive adaptive adjustments in the

morphology of biological shells by altering the chemical and physical conditions of water bodies.

The mixed-effects model results are shown in the table below. The Shape Factor responded significantly negatively to both $\delta^{13}C$ and $\delta^{18}O$, indicating that intensified carbon cycling or temperature/ice volume changes cause the shape to become more flattened or have a lower shape factor. Elongation was positively correlated with both isotopes, suggesting that shells or structures tend to elongate in response to environmental pressure, possibly reflecting adaptations to water flow or nutrient acquisition strategies. Sphericity was significantly negatively correlated with both $\delta^{13}C$ and $\delta^{18}O$, further confirming that under extreme environmental conditions, structures deviate from spherical shapes and become more complex. Size metrics (Maximum Diameter, Mean Diameter) showed no significant correlations, indicating that shape changes are more sensitive to environmental drivers than overall size increase or decrease.

TABLE IV: Mixed-effects model results for morphological indicators and isotopes.

Morphological Indicator	$\delta^{13}C$ **(coef)**	**p**	$\delta^{18}O$ **(coef)**	**p**
Shape Factor	-0.115	0.008	-0.416	<0.001
Max Diameter	-0.057	0.223	+0.013	0.779
Mean Diameter	-0.080	0.087	-0.058	0.208
Elongation	+0.167	<0.001	+0.537	<0.001
Sphericity	-0.171	<0.001	-0.498	<0.001

The cross-correlation and lag effects are shown in the table below. The lag response: morphological adjustments do not synchronize instantaneously with the environment but rather reach their strongest correlation after several periods, reflecting time delays due to growth accumulation, population generations, and biological compensation mechanisms. $\delta^{18}O$ had a stronger influence on Elongation and Sphericity ($|r| \approx 0.5$), suggesting that temperature/ice volume fluctuations have a more pronounced effect on morphology. In contrast, $\delta^{13}C$ had a weaker effect on most indicators compared to $\delta^{18}O$.

TABLE V: Cross-correlation and lag effects for morphological indicators and isotopes. The abbreviations are: Indicator (Ind.), Optimal Lag (Lag), and Correlation Coefficient (Coef. r).

Ind.	**Lag (Periods)**	**Coef. r**
Shape Factor vs $\delta^{13}C$	2	+0.19
Shape Factor vs $\delta^{18}O$	3	-0.36
Max Diameter vs $\delta^{13}C$	8	-0.11
Max Diameter vs $\delta^{18}O$	8	+0.11
Mean Diameter vs $\delta^{13}C$	8	-0.09
Mean Diameter vs $\delta^{18}O$	8	+0.05
Elongation vs $\delta^{13}C$	5	-0.24
Elongation vs $\delta^{18}O$	5	+0.51
Sphericity vs $\delta^{13}C$	4	+0.24
Sphericity vs $\delta^{18}O$	5	-0.49

VI. Further Work and Improvement

A. Changes in the shape of size distribution of Foraminifer through time

Firstly, incorporating additional morphological parameters such as wall texture or chamber number could further refine morphospace analyses. Eventually, further implementation of environmental proxies other than oxygen and carbon isotope proxies, like sea surface temperature and productivity data, might provide better attribution of change points to specific climatic drivers. Secondly, hidden markov models (hmm) or recurrent neural networks (rnn) can use advanced machine learning techniques to model evolutionary regime shifts more dynamically through time. Additionally, expanding the dataset to include other key ODP/IODP sites would allow testing for regional versus global evolutionary patterns. Finally, employing causal inference methods to formally test the relationships between environmental change and morphometric evolution would strengthen the ecological interpretations.

B. Morphological Patterns over Geological Time

Future research would build on this research by adding species-level taxonomic identification to separate evolutionary patterns owing to species turnover from those owing to morphological plasticity within lineages. Attribution of morphological change to particular taxonomic groups would enable more fine-scale interpretation of adaptation and trait evolution. Addition of other environmental proxies like carbon isotopes (δ^{13}C) or indicators of nutrient flux would also help in elucidating the exact abiotic drivers of noted trends in size and shape differentiation.

A tantalizing refinement would be the application of machine learning approaches, such as unsupervised clustering or hidden Markov models, to discern subtle transitions in morphological states imperceptible through traditional statistical avenues. Finally, genomic data from modern analogues could be leveraged for hypothesis testing of independent genetic controls on size and shape evolution, providing a molecular context for validating the inferred evolutionary decoupling suggested by the fossil record.

C. Changes in number of species through time and in shap data link to Environment

For Question 2 and 5, future research could refine this work by addressing methodological constraints and expanding environmental linkages. The sliding window visualization of species trends may mask abrupt climatic impacts; incorporating wavelet analysis or dynamic, climate-aligned geological period definitions could better resolve transient events. Validating Gaussian Mixture Model clusters against non-parametric methods (e.g., DBSCAN) would strengthen cluster robustness. Observed morphological lag effects (2–8 periods) warrant targeted experiments on organismal growth rates and environmental signal integration. Spatial random effects in mixed-effects models could account for geographic heterogeneity, while hydrodynamic simulations might test

functional interpretations of shape adaptations (e.g., elongation in turbulent flows). Complementing δ^{13}C/δ^{18}O data with proxies like Mg/Ca ratios or biomarkers could clarify biomineralization drivers. Machine learning approaches (e.g., CNNs) may capture subtle shape variations beyond traditional metrics, enhancing phenotypic plasticity analyses.

VII. Conclusion

A. Changes in the shape of size distribution of Foraminifer through time

Analysis revealed major structural changes in foraminiferal size distributions across the studied interval. KDE indicated shifts in dominant size classes, with broader distributions during warm periods and narrower ones during glacial phases. Change point detection (PELT) identified key transitions at 1.6 Ma and 3.2 Ma, corresponding to intensified glacial cycles. Statistical measures showed decreasing skewness and elevated kurtosis, reflecting diversification and morphological bottlenecks. PCA confirmed temporal clustering of morphotypes, suggesting structured evolutionary responses rather than gradual drift. Overall, planktic foraminiferal size evolution was shaped by discrete ecological regime shifts driven by major paleoenvironmental changes.

B. Morphological Patterns over Geological Time

Indeed, the patterns of change over time are different: "size fluctuates irregularly with little correlation to time (p = -0.04)", whereas "shape parameters (elongation, sphericity, and complexity) show abrupt and synchronized shifts around 3 Ma", indicating "distinct selective pressures" acting separately on size and shape traits.

C. Changes in number of species through time

Clustering analysis using Gaussian Mixture Models (GMM) with Bayesian Information Criterion (BIC) optimization revealed pronounced stage-specific fluctuations in species richness during the Quaternary, tightly linked to climatic events. Peak species diversity occurred in the Early Pleistocene (2.58–1.2 Ma; mean = 21.5), likely driven by accelerated ecological niche differentiation during glacial-interglacial cycles, followed by a decline during the Mid-Pleistocene Transition (1.2–0.8 Ma; mean = 20.4) coinciding with intensified climate instability. The minimum species count emerged at 1.88 Ma, while Late Pleistocene intervals (0.8–0.01 Ma) showed extreme lows ($\leq$ 12 species), correlating with Last Glacial Maximum extinctions. Statistical significance ($p < 0.05$) marked Early/Late Pleistocene contrasts, with elevated richness clustered in warmer phases (1.5–2.5 Ma) and depletion events aligning with Heinrich Stadials (0.1–0.5 Ma), substantiating climatic forcing mechanisms governed biodiversity evolution.

D. Change in shape data link to Environment

Mixed-effects models and cross-correlation analyses demonstrated distinct environmental controls of isotopic proxies ($\delta^{13}C$, $\delta^{18}O$) on shell morphology. Shape Factor ($\beta = -0.12$ to -0.42, $p < 0.01$) and Sphericity ($\beta = -0.17$ to -0.50, $p < 0.001$) exhibited significant negative correlations with both isotopes, while Elongation showed positive associations ($\beta = +0.17$ to $+0.54$, $p < 0.001$), indicating intensified carbon cycling (high $\delta^{13}C$) or cooler/high-ice-volume conditions (high $\delta^{18}O$) drove shell flattening, elongation, and reduced sphericity, with no size-related responses. Cross-correlation revealed 2–8 period lags ($\approx$ 10–40 kyr) in morphological adjustments, where $\delta^{18}O$ exerted stronger control over Elongation ($r = +0.51$) and Sphericity ($r = -0.49$), suggesting temperature/ice-volume shifts altered hydrodynamic or osmotic constraints. The $\delta^{13}C$ lagged effect on Shape Factor (lag = 2 periods, $r = +0.19$) may reflect niche reorganization linked to productivity fluctuations. These findings highlight morphological plasticity as a critical adaptive phenotype to Quaternary climatic oscillations, with response delays governed by generational turnover and physiological buffering mechanisms. Key corrections made:

References

[1] George E. P. Box et al. *Time Series Analysis: Forecasting and Control*. 5th. Wiley, 2015.

[2] James W. Cooley and John W. Tukey. "An algorithm for the machine calculation of complex Fourier series". In: *Mathematics of Computation* 19.90 (1965), pp. 297–301.

[3] Øyvind Hammer and David A.T. Harper. *Paleontological data analysis*. Wiley-Blackwell, 2001.

[4] Norden E. Huang et al. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis". In: *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 454.1971 (1998), pp. 903–995.

[5] Leonard Kaufman and Peter J Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 1990.

[6] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. "Optimal detection of changepoints with a linear computational cost". In: *Journal of the American Statistical Association* 107.500 (2012), pp. 1590–1598.

[7] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing*. 3rd. Prentice Hall, 2009.

[8] Claude E Shannon. "A mathematical theory of communication". In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[9] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC, 1986.

Appendix

The full codebase and supplementary materials are available at: https://github.com/EMATM0050-2024/dsmp-2024-group-m9.