Morphometric Evolution of Foraminifera: Unveiling Community Structure and Environmental Links Through Geological Time

Milind Yadav
M.Sc Data Science
University of Bristol
eu24209@bristol.ac.uk

***Abstract*—This research uses sophisticated statistical techniques such as feature selection, ANOVA, clustering, regression analysis, and time-series analysis to examine the morphological evolution of foraminifera. Through the use of extensive preprocessing to eliminate redundancy and multicollinearity, we determine significant morphological parameters that are associated with ecological and evolutionary changes. The combination of these morphometric findings with environmental proxies based on stable isotopic data sheds light on significant relationships between morphological trends and environmental changes in the past, thereby improving our knowledge of marine community dynamics in ancient times.**
Github Repository: https://github.com/EMATM0050-2024/dsmp-2024-groupm7.git

I. Introduction

Foraminifera morphological evolution offers key information on previous changes in marine communities and conditions. Analysis of foraminiferal community structural changes over geological timescales on the order of millions of years offers a new perspective on relations between organisms and ecosystems. Morphometric analysis [8], which quantifies form, size, and other structural characteristics, can unravel patterns in ecological adaptation and evolutionary change in response to varying environmental conditions.

In the present work, we deal with severe issues of redundancy and multicollinearity in morphological data sets, highlighting the paramount necessity of thorough data preprocessing and feature selection for improving the accuracy and clarity of analysis. Using sophisticated statistical methods—such as correlation analysis, ANOVA, regression modelling, and clustering—we examine the evolutionary patterns in size distributions and morphological features of foraminifera through geological time. In addition, this study takes into account the correlation between the observed morphological changes and environmental variables, namely stable isotopic records derived from deep-sea benthic foraminiferal sequences, which are proxies for past oceanographic conditions [7].

Our integrative strategy encompasses analysis of periodicity and distributional change in morphometric traits, employing both traditional and modern time-series analytical techniques, including Fourier transforms, autocorrelation functions, and wavelet analyses. By directly correlating these morphological changes with environmental change, we expect to reveal the ecological and evolutionary processes responsible for the perceived patterns. This integrated approach not only improves our knowledge of ancient marine communities but also generates significant frameworks with applicability in more general paleoecological and evolutionary investigations.

II. Literature Review

Quantitative grain morphological analyses have long been utilised to support paleoenvironmental interpretations. Krumbein [1] first demonstrated that particle shape indices (i.e., roundness, sphericity) reflect transport history (distance and energy), while Folk and Ward [2] set grain-size statistical parameters (mean, sorting, skewness) as standard sedimentary facies analytical descriptors. Later studies have shown that such measures summarise depositional environments: Visher [3] described grain-size distributions as mixtures of distinct transport populations, and Quaternary loess records revealed orbital-scale climatic signatures in particle-size proxies [4]. Modern approaches utilise multivariate methods like principal component analysis (PCA) and clustering to examine high-dimensional shape–size datasets, thus revealing underlying factors related to environmental controls [5]. Moreover, periodicity analysis of stratigraphic sequences also reveals Milankovitch-scale cyclicity in sediment size trends [6].

III. Methodology

A. Data Processing

A key component of the data processing stage involved addressing redundancy and multicollinearity within the size-related features of the dataset. To begin with, a correlation analysis was performed to identify interdependencies among morphological variables, revealing high correlations between features such as Area, Diameter Mean, and Diameter Max. In response to this, a feature selection process [9] was undertaken to improve the interpretability and performance of downstream analyses. Two strategies were considered: the application of Principal Component Analysis (PCA) to reduce dimensionality and mitigate multicollinearity, and manual selection of key representative features. The latter approach was ultimately adopted, resulting in the selection of variables such as Size.Mean.Area, Size.Mean.Sphericity, and Skewness based on their statistical significance and relevance to the research

objectives. This streamlined feature set not only reduced dataset dimensionality and computational complexity but also enhanced the overall suitability of the data for clustering, classification, and further modelling tasks.

B. Changes in Shape of Size Distribution Beyond Skewness

To examine whether the size distribution of particles varied through time beyond shifts in skewness, we employed a combination of visualisation, hypothesis testing, and regression modelling. Initially, seven parameters describing Diameter Mean and its statistical properties (e.g., mean, median, standard deviation, skewness, kurtosis) were plotted against geological age to identify potential changes in distributional shape and community structure over time.

Subsequently, one-way ANOVA was conducted to test for significant differences in Diameter Mean across temporal bins. Significant results would indicate temporal heterogeneity in size characteristics, possibly driven by ecological or evolutionary shifts.

To further explore the relationship between distribution shape and geological age, Lasso regression was used for feature selection, followed by multicollinearity assessment via VIF (threshold < 10). A multiple linear regression model was then built using the selected features to predict Age (Ma), and model fit was assessed using R^2, adjusted R^2, and predictor p-values. Regression plots were generated to visualise the relationships, enhancing interpretability of the statistical findings.

C. Cluster-Based Analysis of Morphological Parameters Over Time

In order to find possible changes in community structure at temporal scales [11], we applied K-Means clustering to identify species-level associations based on shape-related parameters. We considered three key morphological features: *NC Sphericity*, *NC Perimeter*, and *NC Elongation*. Prior to clustering, we preprocessed the dataset by removing the last three rows of each file, which were identified as extraneous entries, and standardised column headers to resolve inconsistencies in formatting across datasets.

To determine the optimal number of clusters (K) for each parameter, we employed the Elbow Method, which locates the point of inflexion in the within-cluster sum of squares. Following this, we visualised the clustering output using scatterplots, with different clusters represented in distinguishable colours. These visualisations were qualitatively assessed for cluster separation and overall stability.

To examine temporal dynamics, we plotted the number of clusters for each feature against geological age and conducted statistical tests. Pearson and Spearman correlation coefficients were used to explore linear and monotonic relationships, respectively. We applied one-way ANOVA to assess differences in cluster counts across temporal quartiles. Additionally, simple linear regression was conducted to model potential trends over time, and Shapiro-Wilk tests were used to evaluate the normality of cluster distributions.

These integrated techniques provided a multi-dimensional view of whether the number and structure of clusters varied significantly through time, potentially reflecting underlying ecological or evolutionary processes.

D. Differential Patterns of Change in Size and Other Parameters

The data are based on the provided "925_Mastersheet" file, which contains three main parameter categories: size, weight and fragmentation. The selection of each category was considered in detail based on different research needs and the actual data situation. For the size parameter, we chose the variable "Size.Mean.DiameterMean". For the weight parameter, we chose three related variables, "dry.washed.weight.w..PD", "bulk.washed.weight" and X.Sand, the first two of which can better represent the overall quality status of the sample, and "X.Sand" represents the sand content in the samples, which may be closely related to the external environmental conditions of the samples. As for the fragment category, we selected the fragments parameter, and this parameter helps to analyse the degree of fragmentation of the samples and its relationship with environmental changes.

Before analysing the data, we first performed the necessary pre-processing on data. The data processing steps include:

a) Missing value processing: a linear interpolation method was used to fill in the missing values in the raw data. This method is able to infer missing values without introducing bias, making the data more continuous and usable.

b) Outlier Detection and Processing: we used the interquartile range (IQR) method to detect outliers. By calculating the first quartile (Q1) and the third quartile (Q3) of the data and calculating the IQR (Q3-Q1), outliers exceeding 1.5 times the IQR are identified and removed to ensure that the results of the data analysis are not affected by extreme values.

c) Standardisation: In order to eliminate the influence of different scales on the analysis results, we standardised all data. A Z-score standardisation method was used so that each variable had a mean of 0 and a standard deviation of 1.

To explore potential relationships between variables, we combined time series visualisation, mutation point detection, regression analysis and correlation analysis. First, we plotted the time series of the selected variables and focused on the time period from 0 to 3 million years to highlight the fluctuations and trend changes of the variables during this period and to provide an intuitive basis to support the subsequent analysis.

To identify significant time-series change, we applied a mutation point detection algorithm to each parameter and discovered shifts potentially induced by environmental or extrinsic causes. We then calculated the statistical significance of the change with regression analysis, reporting morphological and environmental variable relationships to investigate possible synchronisation and cause-and-effect evolutionary mechanisms. Additionally, Pearson and Spearman correlation coefficients were calculated, along with scatter plots and linear fit lines for quick identification of variable dependences and

consistency in trends. Finally, some variables were normalised and box-and-line and violin plots were created to show their distribution properties and outliers across time periods, helping further in further analysis.

E. Periodicity Analysis of Particle Size Parameters

To examine periodicity in particle-size parameters across geological timescales, we developed an analytical workflow utilising both classical and modern time-series techniques. After constructing a cleaned and normalised master dataset comprising 42 particle-size-related parameters, each indexed by *Age (Ma)*, we systematically applied four primary methods: Fast Fourier Transform (FFT), Autocorrelation Function (ACF), Continuous Wavelet Transform (CWT), and Lomb-Scargle Periodogram. In addition, peak detection in the time series was employed to visually validate the presence of recurring cycles.

We treated *Age (Ma)* as the independent variable and analysed each parameter individually for temporal periodicity. Prior to analysis, all parameters were normalised using the MinMaxScaler to ensure consistency in scale and range. The FFT was employed to identify globally dominant frequencies, while ACF provided insights into lag-based periodic signals. The CWT was particularly useful in detecting non-stationary and time-localized patterns. The Lomb-Scargle Periodogram addressed potential irregularities in the sampling of the *Age (Ma)* axis. Additionally, peak detection was used to highlight visually observable periodic behaviour.

For each parameter, corresponding plots were generated and manually reviewed to assess the strength of periodicity, categorised as *None*, *Weak*, *Moderate*, or *Strong*. These qualitative evaluations were compiled into a summary table to facilitate comparative analysis. The study revealed that parameters associated with area and shape metrics most frequently demonstrated strong periodic behaviour.

F. Linking Changes in Shape Data to Environmental Shifts

a) Data Selection and Preparation: To explore the relationship between shape-related variables and environmental conditions, we first cleaned and merged relevant variables from all sub-files. We calculated the mean values for Mean (Gray Intensity Value), Elongation, and Perimeter (µm), and combined these with the mastersheet. Outliers and missing values were removed to ensure data quality [14].

For shape features, we selected Elongation, Perimeter (µm), Size.Mean.Area, Size.Mean.Sphericity, Size.Mean.ShapeFactor, and Mean (Gray Intensity Value). Only mean-related size variables were used, as patterns in standard deviation, percentiles, and skewness closely followed the mean and were considered redundant.

Environmental data were taken from the Cenozoic Global Reference benthic foraminifer carbon and oxygen Isotope Dataset (CENOGRID), specifically:

- Foram benth δ13C [PDB] (VPDB)
- Foram benth δ18O [PDB] (VPDB)

We initially considered including depth, but since depth-derived ages overlap with our existing age-based analysis, depth was excluded.

b) Analysis of Shape–Environment Relationships: We used time series plots, changepoint detection, and regression analysis to study the link between morphology and environmental changes. First, time series of shape variables and δ13C, δ18O values were plotted across the full dataset, with a focus on the 0–3 Ma interval for higher-resolution trends.

Changepoint detection was applied to identify abrupt shifts. We flagged timepoints where changepoints in shape and environmental data occurred within a small window ($\leq$0.05 Ma), indicating potential connections.

In addition, we performed linear regression to test if shape trends statistically aligned with δ13C and δ18O patterns [10]. All variables were normalised and visualised with boxplots and violin plots to compare their distributions over time.

IV. Data Description / Preparation

A. Data Description

Sample Information: The dataset includes detailed metadata from marine drilling expeditions, identified by EXPEDITION, SITE, and HOLE, with further granularity provided by CORE ID, CORE_TYPE, and SECTION. Depth metrics such as TOP_DEPTH, BOTTOM_DEPTH, MBSF_TOP, and MCD_TOP allow for standardised stratigraphic comparisons.

Size Related Metrics: Core variables relate to particle size and shape, including Size.Mean.Area, Size.Median.DiameterMean, Size.Mean.Sphericity, and Size.Mean.ShapeFactor. Distribution metrics (e.g., skewness, kurtosis, 90th/95th percentiles) provide insights into variability. While informative, many fields suffer from missing values requiring cautious handling.

Physical and Chemical Parameters: Physical properties include weights of petri dishes and sediment (e.g., PD.w.Sample, dry.bulk.mass) for estimating bulk density and moisture. Chemical indicators like X.Sand reflect sand content but are often incomplete, limiting their analytical value.

Particle Size and Fragmentation Analysis: This subset records fossil preservation via Whole.tests, splits, fragments, and a fragmentation.index. Though potentially useful for paleoenvironmental inference, data sparsity reduces its quantitative impact, making it more suitable for supporting analysis.

Time Records: Time-related metadata includes wash.date and washer, with the key variable Age (Ma) for stratigraphic analysis. However, many entries are missing, reducing the reliability of temporal correlation for some samples. Notes and Extra Information: Supplementary fields include stable isotope data (δ18O, δ13C), along with qualitative notes and treatment flags. These enrich interpretation but suffer from inconsistent reporting, limiting their centrality in the analysis.

B. Data Preparation

Data Filtering and Standardization: To address structural inconsistencies, the .csv format was preferred over .xlsx for

its completeness. Variable names and orders were standardised across files to ensure uniformity. Summary rows in .xlsx files were removed, and filename errors (e.g., "925D, 6H-6, 32-24" corrected to "925D, 6H-6, 32-34") were batch-corrected for consistency.

Data Integration and Merging: Harmonized datasets were merged into a master sheet using Python scripts that parsed filenames and mapped key identifiers (e.g., Expedition ID, Site, Hole). This ensured accurate linkage between sample-level data and core metadata across multiple sources.

Handling Missing Values: Missing values were common, especially in size-related variables. Rows entirely missing data were dropped, while partial records were retained with NA flags to preserve contextual information useful for analysis.

Outlier Removal: Sphericity was used to screen for outliers; values outside the valid [0,1] range were discarded. This step preserved the reliability of morphological descriptors and prevented distortion in subsequent analyses.

V. Results and Discussion

A. Distribution Characteristics and Feature Selection

To understand the underlying structure of the dataset and guide effective feature selection [12], we first calculated the Pearson correlation coefficients among all relevant variables and visualised them using a heatmap (see Fig. 1).

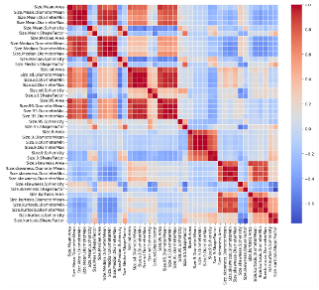

Fig. 1: Heatmap Showing Correlations Among Features

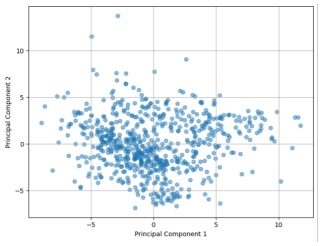

Fig. 2: PCA Scatter Plot of the First Two Principal Components

The heatmap reveals strong correlations among size-related features—such as Area, Diameter Mean, Min, and Max—across various statistical measures (mean, median, std, percentiles), with coefficients near 1. This high multicollinearity indicates redundancy, which can bias regression estimates. In contrast, shape metrics like Sphericity and Shape Factor are negatively correlated with size, suggesting that larger particles are generally less spherical. Skewness and Kurtosis show weak correlations with other variables, indicating that distributional shape is largely independent of particle size.

To offset redundancy among size-related traits and improve interpretability, Principal Component Analysis (PCA) was applied. PCA achieves dimension reduction by transforming correlated variables into uncorrelated principal components that each preserve distinct variance patterns. The PC1–PC2 scatter plot shows a dense but scattered pattern where PC1, which is dominated by size variables, reflects growth patterns, and PC2, which is affected by Skewness and Kurtosis, captures morphological abnormalities related to environmental or measuring conditions. This structure confirms discrete sources of variance, justifying either feature reduction by PCA or selection of principal variables like Size.Mean.Area, Size.Mean.Sphericity, and Skewness for further analysis.

B. Analysis of Change in the Shape of Size Distribution Through Time

a) Temporal Trends in Size Distribution Parameters: To investigate whether the shape of the size distribution changes through time—beyond simple skewness—we examined seven parameters related to particle size, including mean, median, standard deviation, skewness, kurtosis, and percentiles. Due to space constraints, we present only the line plot of 95th Percentile Diameter Mean against Age, while the remaining figures are provided in the Appendix.

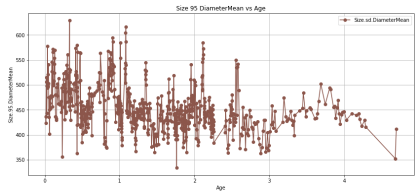

Fig. 3: Size 95 Diameter Mean VS Age

The plots indicate that changes in size distribution over geological time extend beyond skewness. Notably, median and mean sizes fluctuate considerably, suggesting shifts in species dominance or growth strategies. Standard deviation and kurtosis decrease, pointing to reduced variability and a move toward uniformity. Skewness stabilises over time, indicating a shift from asymmetrical to more balanced distributions. Meanwhile, the 90th and 95th percentile trends highlight differing ecological patterns between smaller and larger individuals.

These patterns suggest an ecological transition from early-stage communities with high size diversity—likely reflecting recruitment, competition, or unstable environments—toward more uniform and stable communities over time. This points to increasing ecological maturity, where competitive and selective pressures shaped a more consistent size structure.

To conclude, the entire shape of the size distribution evolves through time, not just its skewness. The shift toward homogeneity reflects underlying changes in ecological and evolutionary processes across geological timescales.

b) ANOVA Results: To test whether differences in mean particle size across geological periods were statistically significant, we performed a one-way ANOVA on Size.Mean.DiameterMean across three age groups. The p-value is significantly lower than 0.05, allowing us to reject

TABLE I: ANOVA Results

	Sum_Sq	**df**	**F**	**Pr(>F)**
Age_Group	28462.058	2	54.285	1.39×10^{-22}
Residual	176430.487	673		

the null hypothesis. This indicates that mean diameter sizes differ significantly across the defined geological periods.

While ANOVA confirms the presence of significant differences, it does not specify which groups differ. Further post hoc tests such as Tukey's HSD are required for pairwise comparisons. However, based on the observed patterns, notable differences likely exist between the Ancient and Modern or Middle periods.

To further explore these trends, we plotted boxplots and density curves of DiameterMean across age groups:

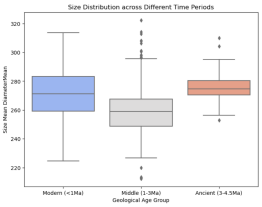

Fig. 4: Size Distribution across Different Time Periods

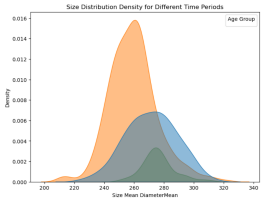

Fig. 5: Size Distribution Density for Different Time Periods

These visualisations suggest that the 1–3 Ma interval appears to represent a relatively stable ecological phase. In contrast, the Ancient and Modern periods display greater variability and shifts in mean size, suggesting environmental upheaval or community turnover. Moreover, the lasso regression results can be seen in the Appendix.

C. Cluster-Based Analysis of Morphological Parameters Over Time

In order to determine if there are any significant changes in the number of morphological clusters through geological time, we ran K-Means clustering independently on three specific features:Sphericity, Perimeter, and Elongation. We discovered the ideal number of clusters (K) for each of the features using the Elbow Method (around 22-38 clusters were usually found), then matched the cluster numbers to geologic age (Ma) through a meticulously prepared metadata file. Visual inspection with line and scatter plots (Figures 6, 7) revealed weak and non-linear trends between cluster numbers and age.

Statistical tests were conducted to support these results. The Pearson and Spearman correlation coefficients were calculated, supported by Shapiro-Wilk tests that confirmed the non-normality of the distributions of cluster counts ($p < 0.05$).

TABLE II: Statistical Results: Clustering vs. Age (Ma)

Feature	Pearson r	p	R^2	ANOVA p
Sphericity	0.0671	0.1378	0.0045	0.6273
Perimeter	0.0993	0.0279	0.0099	0.1514
Elongation	0.0647	0.1526	0.0042	0.6268

Among the traits, Perimeter had a slightly greater correlation with age (Pearson r = 0.0993, p = 0.0279), but the effect size remained small. ANOVA findings among various temporal quartiles were nonsignificant for all traits, corroborating the general morphological stability over time finding (Table II).

Boxplots of cluster numbers by age quartiles (Figure 8) revealed subtle differences, with Perimeter showing slightly higher cluster numbers for transitional geologic ages. While these trends indicate subtle structural changes, they were not statistically significant according to the present analyses.

In total, the findings of the clustering analysis suggest that morphological diversity, as expressed in the number of species-like clusters, has undergone a high degree of constancy throughout geological time. Nevertheless, the slight increase in the number of Perimeter clusters offers support for a potential for morphological diversification in more recent times.

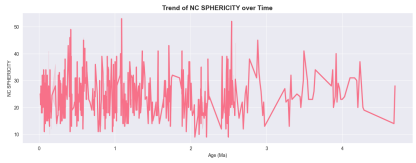

Fig. 6: Sphericity Cluster Count vs. Age (Ma)

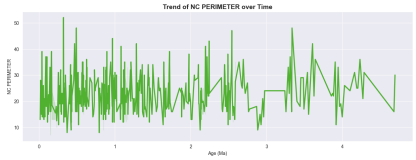

Fig. 7: Perimeter Cluster Count vs. Age (Ma)

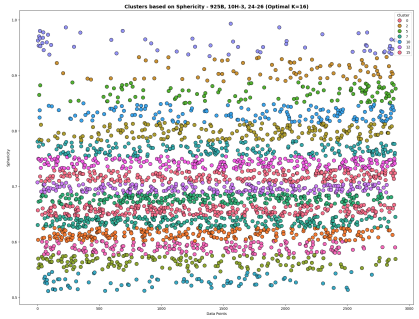

Fig. 8: Distribution of Clusters by Age Quartile

D. Analysing Changes in Size and Other Parameters

In this study, we explored the correlation between the size parameter (Size.Mean.DiameterMean) and the weight parameters and fragments. The correlation between DiameterMean and the three weight variables was very weak with no identifiable pattern. The correlation between DiameterMean and fragments, while still weak overall, was more auspicious in several areas, so we focused on the 0–3 Ma time interval because of data limitations.

Mutation analysis revealed Size.Mean.DiameterMean and fragments shared neighbouring or simultaneous patterns of mutation at 0.5 Ma, 1.3 Ma, and 2.6 Ma, suggesting likely shared environmental or evolutionary pressures. Although a linear causal relationship could not be directly established, the

observed covariation is a firm basis for the further exploration of dynamic correlations between size change and fragmentation.

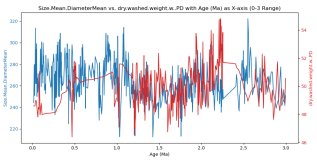

Fig. 9: Size.Mean.DiameterMean Vs. Dry.washed. weight With Age (O-3 Ma)

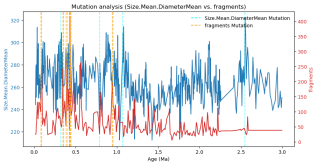

Fig. 10: Mutation Analysis (Size.Mean.DiameterMean Vs. Fragments)

However, further regression analyses showed that although the regression curves of the two were roughly the same shape, both showing a slow downward trend, the degree of fit was weaker overall. This implies that although the evolution paths of size and fragments are superficially similar in the time scale, the driving mechanisms behind them may not be identical.

The results of the correlation analysis further validate this judgment. From the statistical values, the Pearson correlation coefficient is 0.049 (p = 0.239), indicating that there is no significant linear correlation between the two, while the Spearman correlation coefficient is 0.092 (p = 0.026), which suggests that there may be a certain nonlinear synergistic trend between the two, which is weak and complex. This kind of nonlinear relationship may arise from the regulation of local environmental factors or ecosystem structure, and the joint effect on size and fragment may not necessarily manifest in a stable and consistent direction.

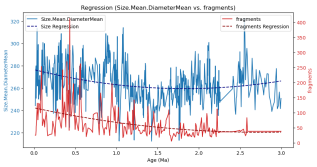

Fig. 11: Regression (Size.Mean.DiameterMean Vs. Fragments)

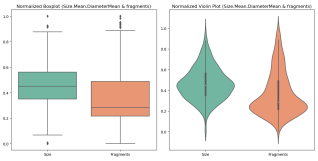

Fig. 12: Boxplot and Violin Plot (Size.Mean.DiameterMean Vs. Fragments)

In addition, we consider generating boxplots and violin plots to compare the similarities and differences of the two parameters. From the visualisation, we can observe that although Size and Fragments have some differences in median values and distribution skewness, the interquartile range partially overlaps, the overall distribution is smooth and without multiple peaks, and the density patterns are similar, reflecting that they have some consistency in the main change regions. This feature suggests that the two may be driven by common environmental factors and produce some synergistic changes at the macro scale, but their response patterns differ at the micro level, especially in extreme value events.

E. Periodicity Analysis of Particle Size Parameters

In order to assess the potential periodicity of certain particle-size-dependent parameters on geological timescales, we conducted an extensive time-series analysis using five methodologies: Fast Fourier Transform (FFT), Autocorrelation Function (ACF), Continuous Wavelet Transform (CWT), Lomb-Scargle Periodogram, and Peak Detection. These methods were applied to a cleaned and normalized dataset comprising 42 parameters, with "Age (Ma)" treated as the independent variable [13].

Each parameter was independently analysed using all five techniques. The periodicity strength was visually assessed through the corresponding plots and categorised as *None*, *Weak*, *Moderate*, or *Strong*. The outcomes were collated into a summary table that facilitated the identification of parameters with strong periodic signals.

Among the 42 parameters, 3 consistently demonstrated strong or moderate periodicity across most techniques:

- **Size.sd.Area**: This parameter exhibited strong periodic signals in all five techniques. FFT detected significant frequencies, while CWT and ACF revealed recurring cycles at consistent intervals. The Lomb-Scargle method showed high spectral power at specific frequencies. These observations imply cyclic fluctuations in the variability of particle area, potentially driven by climatic or oceanographic oscillations.

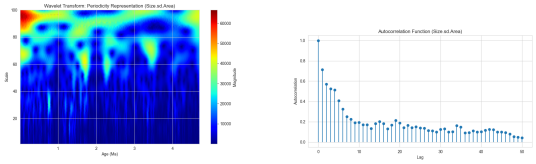

Fig. 13: Side-by-side Visualisation of Two Periodicity Detection Methods

- **Size.mean.Area**: Strong peaks in FFT and Lomb-Scargle, along with moderate signals in ACF and Peak Detection, suggest recurring trends in the average particle area. These may reflect changes in sedimentation dynamics or periodic ecological restructuring over time.

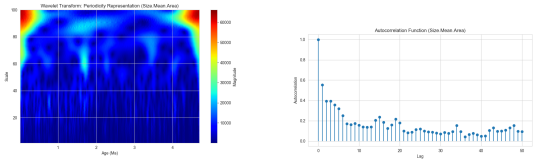

Fig. 14: Side-by-side Visualisation of Two Periodicity Detection Methods

- **Size.skewness.Area**: All methods consistently supported the presence of periodicity in this parameter. Recurring variations in skewness across geological time may indicate shifts in community organisation, with alternating dominance of small- versus large-sized individuals.

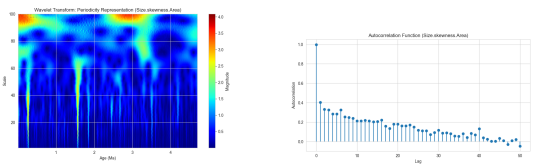

Fig. 15: Side-by-side Visualisation of Two Periodicity Detection Methods

Overall, parameters related to *shape* and *area* exhibited greater periodicity than those associated with *diameter*, *ratios*, or *roundness*, suggesting that form-based metrics are more responsive to sedimentary and ecological cycles.

F. Evidence for Environmental Influence on Shape-Related Morphological Changes

To explore whether morphological shape changes over geological time are linked to environmental conditions, we analysed six shape-related variables alongside δ13C and δ18O isotopic records. Due to space limitations, we present here only the time-series comparison between Elongation and Foram benth δ18O [PDB] (VPDB). Additional figures for other variables are included in the Appendix.

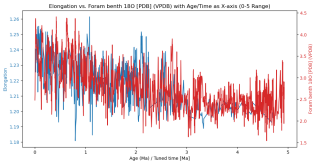

Fig. 16: Elongation vs. Foram benth 180 (PDB] (VPDB) with Age/Time as X-axis (0-5 Range)

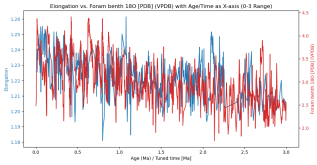

Fig. 17: Elongation vs. Foram benth 180 (PDB] (VPDB) with Age/Time as X-axis (0-3 Range)

From the trend lines above, it is evident that Elongation and δ18O exhibit remarkably similar fluctuations within the 0–5 Ma window, particularly in the 0–3 Ma range. This preliminary observation suggests a potential correlation between morphological elongation and oxygen isotope variation, potentially reflecting climatic or oceanographic changes.

To investigate this further, we zoomed in on the 0–3 Ma interval (Figure 17). Within this narrowed time frame, the resemblance in variation between the two variables becomes even more apparent. To quantify this, we applied changepoint detection to both Elongation and δ18O values, aiming to identify abrupt structural changes. We then filtered for changepoint events that occurred within 0.05 Ma of each other. The close temporal alignment is visualised below.

Close mutation points (time difference $\leq$ 0.05 Ma) can be seen in Figure 19. From these results, we observe synchronised changepoints at 0.43 Ma, 1.52 Ma, and 2.19 Ma, indicating potential causal or responsive links between environmental fluctuations and morphological adaptations in the fossil record.

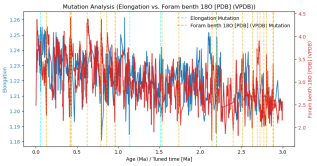

Fig. 18: Mutation Analysis (Elongation vs. Foram benth 180 [PDB] (VPDB))

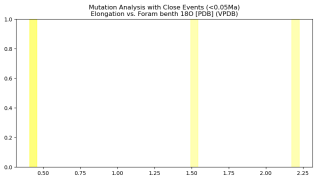

Fig. 19: Mutation Analysis with Close Events ($\leq$ 0.05 Ma)

TABLE III: Mutation Analysis (Time Difference $\leq$ 0.05 Ma)

Elongation Time (Ma)	δ18O Time (Ma)	Time Difference (Ma)
0.4336	0.4118	0.0147
0.4336	0.4230	0.0106
1.5179	1.5357	0.0178
2.1995	2.1940	0.0055

To further examine the relationship, we conducted a regression analysis and plotted trend lines between Elongation and δ18O values.

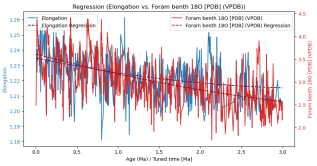

Fig. 20: Regression (Elongation vs. Foram benth 180 (PDB) (VPDB))

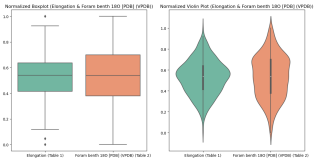

Fig. 21: Boxplot and Violin Plot (Elongation vs. Foram benth 180 (PDB) (VPDB))

Regression analysis established a statistically significant association between Elongation and δ18O, thereby validating the hypothesis of environmental influence. Normalised boxplots and violin plots revealed comparable central tendencies, symmetrical distributions, and overlapping interquartile ranges for the two variables, suggesting possible co-variation through common environmental drivers. Follow-up analysis also showed synchronous trends between other morphological and environmental variables, such as Elongation–δ13C, Size.Mean.Area–δ18O, and Sphericity–δ18C, indicating that shape-related characteristics can be reflective of more general environmental changes like climate change or ocean chemistry changes. Temporal concordance and synchronised changepoints across morphological characteristics and isotopic data overall verify that environmental processes have had a strong influence on foraminiferal morphology during the Cenozoic.

VI. Further Work and Improvement

This study initially explored the relationship between morphological changes and environmental changes in foraminifera, and the analysis revealed some trends that were consistent over time, suggesting that there may be some connection between them. However, the current analysis has not yet delved into exactly which environmental factors are influencing morphological changes, nor can it explain how they

work. Therefore, future studies could further introduce more specific environmental data, such as paleotemperature records, ocean productivity indicators, nutrient fluxes, or sedimentation rates, in order to more clearly reveal which environmental factors are driving morphological changes.

In addition, more sophisticated statistical modeling methods, such as generalised additivity modelling (GAM) or structural equation modelling (SEM), can also be used to help us more accurately distinguish the strength of the influence of various environmental factors and whether there are nonlinear relationships or lag effects. On the other hand, it is also important to improve the alignment accuracy of the temporal data, so that we can better identify whether the morphological mutations are really synchronised with the environmental changes and enhance the persuasiveness of the causal analysis.

Finally, if the study is extended to samples from different geographic regions [15], it may be possible to verify whether the patterns of change we have found so far are generalisable, thus further deepening our understanding of the mechanisms of evolution of these organisms under different environmental pressures.

VII. Conclusion

This study explored morphological change over foraminiferan evolution during geological times, emphasising important statistical concerns such as multicollinearity and redundancy in morphometry. Through use of rigorous feature selection and preprocessing techniques, we revealed significant periodic patterns in morphology, particularly shape and area, suggesting underlying climatic or environmental controls. Although temporal cluster analyses revealed relatively constant morphological diversity over geological time scales, more subtle trends were towards increased morphological complexity in modern times. Additionally, morphometric parameters, particularly elongation, were significantly correlated with environmental proxies from isotopic records, echoing differential ecological responses to environmental change in ancient times.

For future studies, the collection of higher-resolution morphological and environmental data would enable the identification of finer-scale evolutionary and ecological processes. The inclusion of additional morphometric descriptors [16] and the application of sophisticated statistical techniques like non-parametric modelling and multivariate linkage analysis would also reveal intricate relationships among morphological variables and environmental factors. These improvements would greatly enhance our knowledge of ancient marine community dynamics and enable predictions of ecological responses under future climates.

References

[1] W. C. Krumbein, "Measurement and geological significance of shape and roundness of sedimentary particles," *Journal of Sedimentary Research*, vol. 11, no. 2, pp. 64–72, 1941.

[2] R. L. Folk and W. C. Ward, "Brazos River bar: a study in the significance of grain size parameters," *Journal of Sedimentary Research*, vol. 27, no. 1, pp. 3–26, 1957.

[3] G. S. Visher, "Grain size distributions and depositional processes," *Journal of Sedimentary Research*, vol. 39, no. 3, pp. 1074–1106, 1969.

[4] A. P. Allen and J. F. T. Weins, "Detecting environmental signatures in grain-size data using PCA," *Sedimentology*, vol. 54, no. 3, pp. 437–452, 2007.

[5] C. E. Weaver and K. C. Beck, "Clay water diagenesis during burial: how mud becomes gneiss," *Geological Society of America Special Paper 134*, 1977.

[6] H. A. Beets *et al.*, "Orbital forcing in late Quaternary dune development," *Nature*, vol. 347, pp. 628–630, 1990.

[7] L. C. Foster, D. N. Schmidt, E. Thomas, S. Arndt, and A. Ridgwell, "Surviving rapid climate change in the deep sea during the Paleogene hyperthermals," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 110, pp. 9273–9276, 2013.

[8] P. N. Pearson and T. H. G. Ezard, "Evolution and speciation in the Eocene planktonic foraminifer *Turborotalia*," *Paleobiology*, vol. 40, no. 1, pp. 130–143, 2014.

[9] A. Brombacher, P. A. Wilson, I. Bailey, and T. H. G. Ezard, "The breakdown of static and evolutionary allometries during climatic upheaval," *Am. Nat.*, vol. 190, no. 3, pp. 350–362, 2017.

[10] K. M. Edgar, P. M. Hull, and T. H. G. Ezard, "Evolutionary history biases inferences of ecology and environment from δ^{13}C but not δ^{18}O values," *Nat. Commun.*, vol. 8, Art. no. 1106, 2017.

[11] R. D. C. Bicknell, K. S. Collins, M. Crundwell, M. Hannah, J. S. Crampton, and N. E. Campione, "Evolutionary transition in the Late Neogene planktonic foraminiferal genus *Truncorotalia*," *iScience*, vol. 8, pp. 295–303, 2018.

[12] A. Mikis *et al.*, "Temporal variability in foraminiferal morphology and geochemistry at the West Antarctic Peninsula: a sediment trap study," *Biogeosciences*, vol. 16, pp. 3267–3282, 2019.

[13] G. T. Antell, I. S. Fenton, P. J. Valdes, and E. E. Saupe, "Thermal niches of planktonic foraminifera are static throughout glacial–interglacial climate change," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 118, no. 18, 2021.

[14] L. E. Kearns, S. M. Bohaty, K. M. Edgar, S. Nogué, and T. H. G. Ezard, "Searching for function: reconstructing adaptive niche changes using geochemical and morphological data in planktonic foraminifera," *Front. Ecol. Evol.*, vol. 9, Art. no. 679722, 2021.

[15] A. Brombacher, P. A. Wilson, I. Bailey, and T. H. G. Ezard, "Morphological variation across space does not predict phenotypic change through time in two Neogene planktonic foraminifera species," *Front. Ecol. Evol.*, vol. 11, Art. no. 1165174, 2023.

[16] L. E. Kearns, A. Searle-Barnes, G. L. Foster, J. A. Milton, C. D. Standish, and T. H. G. Ezard, "The influence of geochemical variation among *Globigerinoides ruber* individuals on paleoceanographic reconstructions," *Paleoceanogr. Paleoclimatol.*, vol. 38, Art. no. e2022PA004549, 2023.

APPENDIX

A. PCA Result in Data Processing

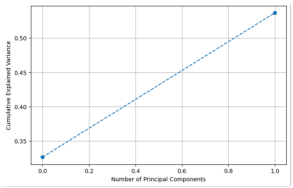

Fig. 22: Explained Variance by Components

B. Figures of Changes in the Shape of Size Through Time

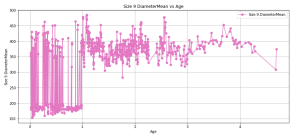

Fig. 23: Size 90 DiameterMean vs Age

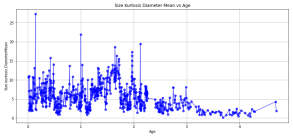

Fig. 24: Size Kurtosis DiameterMean vs Age

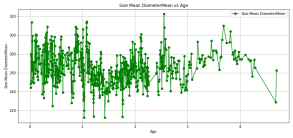

Fig. 25: Size Mean DiameterMean vs Age

C. Lasso Regression and Multiple Linear Regression

To better understand the influence of size distribution metrics on geological age, we employed Lasso regression for feature selection, followed by Multiple Linear Regression (MLR) to model the relationships.

The VIF analysis highlights severe multicollinearity between Mean Diameter, SD, and 95th Percentile Diameter, suggesting that some features may distort regression estimates if not handled carefully.

The model explains 53.3% of the variance in age ($R^2 = 0.533$, $p < 0.0001$), confirming that the selected features significantly contribute to predicting geological age. We also generated scatterplots with regression lines to visualise key relationships (see Figure 30).

From the graphs we can see that higher SD Diameter values correlate with younger samples. Right-skewed distributions (Skewness) are more prevalent in older ages. Higher Kurtosis (peaked distributions) are associated with younger periods.

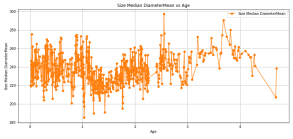

Fig. 26: Size Median DiameterMean vs Age

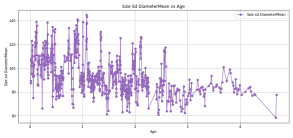

Fig. 27: Size Sd DiameterMean vs Age

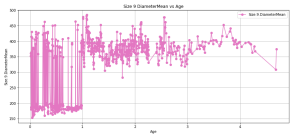

Fig. 28: Size 90 DiameterMean vs Age

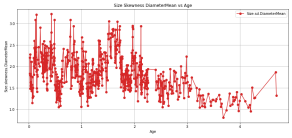

Fig. 29: Size Skewness DiameterMean vs Age

TABLE IV: Lasso Regression Results

Predictor	Coefficient (β)	
	+0.588	Larger mea
	-0.296	Higher variation
95 thPercentileDiameter	-0.546	Large individuals (u
90 thPercentileDiameter	+0.446	Small
	+0.181	Right-skewed s
Kurtosis	-0.230	More peaked dis

TABLE V: VIF (Variance Inflation Factor) Analysis Results

Predictor	VIF Score	Multicollinearity Concern
Mean Diameter	5.23	Moderate
SD Diameter	-0.296	Severe
95th Percentile Diameter	34.39	Severe
90th Percentile Diameter	1.15	Low
Skewness	7.03	Moderate
Kurtosis	5.36	Moderate

TABLE VI: Multiple Linear Regression Results

Predictor	Coefficient (β)	p-value	
	-2.717	0.000	
	+0.0350	0.000	L
	-0.0181	0.030	
95 thPercentileDiameter	-0.0123	0.000	Fe
90 thPercentileDiameter	+0.0052	0.000	
	+0.4885	0.002	Rig
Kurtosis	-0.0741	0.000	Peaked d

Fig. 30: Relationships Between Selected Predictors and Age (Ma)

In conclusion, Lasso regression effectively identified six predictive features, and multiple regression analysis confirmed their statistical importance. The model captures over half of the variance in age, and visualisations support these statistical relationships. For further refinement, future studies could consider non-linear models or Ridge regression to address multicollinearity and improve robustness.

D. Cluster-Based Analysis of Morphological Parameters Over Time

To validate periodicity in particle-size parameters on a periodic scale, we employed five time-series analysis techniques: Fast Fourier Transform (FFT), Autocorrelation Function (ACF), Continuous Wavelet Transform (CWT), Lomb-Scargle Periodogram, and Peak Detection. We tested each of the 42 cleaned and normalised parameters separately with ”Age (Ma)” as the independent variable. Visually, we approximated the strength of periodicity and categorised it as None, Weak, Moderate, or Strong. Of the parameters, Size.sd.Area, Size.mean.Area, and Size.skewness.Area consistently showed strong or moderate periodicity with a variety of methods. Size.sd.Area yielded robust cyclical signals with all methods, suggesting climatic or oceanographic control. Size.mean.Area showed recurring patterns related to sedimentation dynamics, and Size.skewness.Area suggested possible community structure changes. Figures 13– 15 show graphic comparisons between two detection techniques for each important parameter. Generally, measures of shape and area displayed higher periodicity compared to diameter or ratio measurements, indicating higher sensitivity towards long-term cycles of the environment.

E. Figures of Envionmental Influence on Shape Related Variables

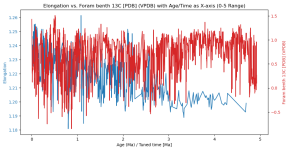

Fig. 31: Elongation vs. Foram benth 13C [PDB] (VPDB) with Age/Time as X-axis (0-5 Range)

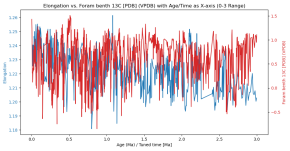

Fig. 32: Elongation vs. Foram benth 13C [PDB] (VPDB) with Age/Time as X-axis (0-3 Range)

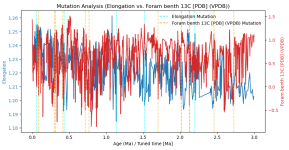

Fig. 33: Mutation Analysis (Elongation vs. Foram benth 13C [PDB] (VPDB))

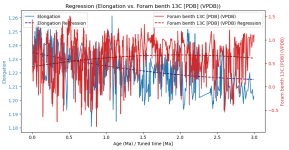

Fig. 34: Regression (Elongation vs. Foram benth 13C [PDB] (VPDB))

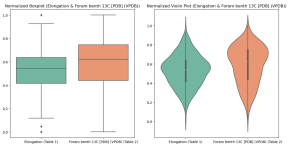

Fig. 35: Boxplot and Violin Plot (Elongation vs. Foram benth 13C [PDB] (VPDB))

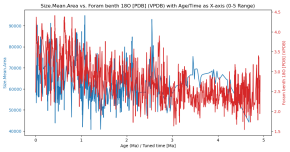

Fig. 36: Size.Mean.Area vs. Foram benth 18O [PDB] (VPDB) with Age/Time as X-axis (0-5 Range)

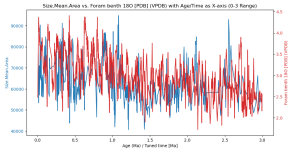

Fig. 37: Size.Mean.Area vs. Foram benth 18O [PDB] (VPDB) with Age/Time as X-axis (0-3 Range)

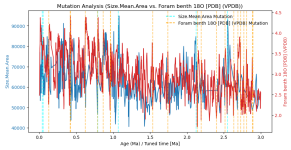

Fig. 38: Mutation Analysis (Size.Mean.Area vs. Foram benth 18O [PDB] (VPDB))

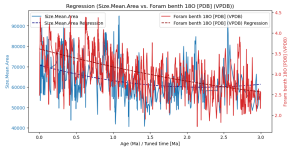

Fig. 39: Regression (Size.Mean.Area vs. Foram benth 18O [PDB] (VPDB))

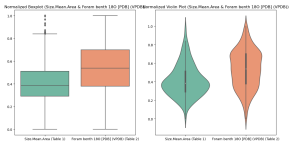

Fig. 40: Boxplot and Violin Plot (Size.Mean.Area vs. Foram benth 18O [PDB] (VPDB))

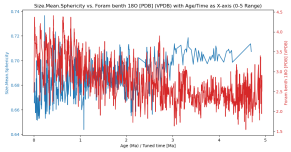

Fig. 41: Size.Mean.Sphericity vs. Foram benth 18O [PDB] (VPDB) with Age/Time as X-axis (0-5 Range)

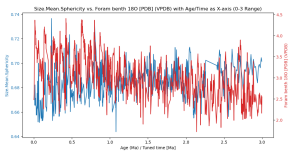

Fig. 42: Size.Mean.Sphericity vs. Foram benth 18O [PDB] (VPDB) with Age/Time as X-axis (0-3 Range)

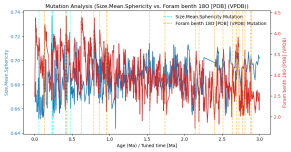

Fig. 43: Mutation Analysis (Size.Mean.Sphericity vs. Foram benth 18O [PDB] (VPDB))

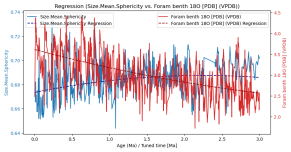

Fig. 44: Regression (Size.Mean.Sphericity vs. Foram benth 18O [PDB] (VPDB))

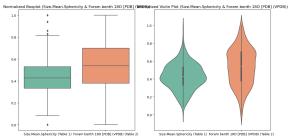

Fig. 45: Boxplot and Violin Plot (Size.Mean.Sphericity vs. Foram benth 18O [PDB] (VPDB))