


Article

Representing Zooplankters: An Example from the Foraminifera

George H. Scott 

GNS Science, 1 Fairway Drive, Lower Hutt 5011, New Zealand; george.scott@gns.cri.nz

Abstract: Because of their excellent preservation record, testate zooplankters provide valuable proxy ocean climate data through the Quaternary–Recent. Commonly, specimen abundances are sought, which are time-consuming to collect manually and require taxonomic expertise. While machine learning models obviate these problems, it is questioned whether the current use of specimens selected by experts to train the models impartially captures the variation within the source populations. To illustrate the potential value of the latter and their relevance to the selection of representative specimens, the 2D outline shape of the planktonic foraminifer *Truncorotalia crassaformis* from four globally distributed, late-Quaternary–modern collections is examined. Large intra-sample variation is attributed to changes in the size and shape of the last-formed chamber, which often departs radically from its predecessors. Similar outlines occur in each collection, and no single axial shape is dominant when the aggregated data, aligned on their centroids and adjusted for size and position, are projected onto their principal components. Several partitions based on distance from the centroid of the standardized data are considered as sources of representative specimens, with that at $\pm 1.645\sigma$ (standard deviations, nominally 90%) suggested as suitable. This procedure obviates the need for expert-based consensus sampling; for greater environmental resolution, it can be applied to individual water mass samples. It assists, but does not fully resolve, the following basic diagnostic question: which characters separate *Truncorotalia crassaformis* from its relatives?

Keywords: foraminifera; machine learning; morphometrics; representative specimens; *Truncorotalia crassaformis*



Citation: Scott, G.H. Representing Zooplankters: An Example from the Foraminifera. *Geosciences* **2024**, *14*, 169. <https://doi.org/10.3390/geosciences14060169>

Academic Editors: Albert Galy and Jesus Martinez-Frias

Received: 26 February 2024

Revised: 21 May 2024

Accepted: 6 June 2024

Published: 14 June 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Which specimens faithfully represent a species is a question that confronts all taxonomists but has risen to new prominence with the application of machine learning (ML) methods [1] that enable the recognition of specimens programmatically. Presently, these methods commonly entail ‘supervised learning’, wherein algorithms are trained on ‘labelled’ data (e.g., images and measurements), to which Linnaean names are already attached, with the implication of independently checked (‘ground-truth’) identifications of representatives. From features in the training set, the algorithms construct models that allow unlabelled (=unidentified) specimens to be classified in accordance with the input representatives. This approach has been applied to the planktonic foraminifera [1–4] which, as mesoscopic testate zooplankters that inhabit mixed-layer through upper-bathyal depths, offer a rich source of ocean climate history [5]. For example, Hsiang et al. [6] used a panel of 24 expert taxonomists to identify taxa from images of planktonic foraminifers selected from 35 Atlantic core tops. Those images, for which there was 75% agreement amongst at least four experts, were retained as labelled training sets. This accommodated the significant inter-expert differences that arose from their variable experiences with the material [7]. Here, the problem of selecting representatives is examined from a population perspective: rather than probing the variable recall of experts on the identity of individual specimens, to examine the structure of the source populations and objectively isolate the potential representatives using those data. Expert judgement on the morphological limits of a source population remains, but reliance on their opinions about individuals is reduced.

A procedure for identifying labelled specimens suitable for ML applications is outlined for *Truncorotalia crassaformis* (Galloway and Wissler, 1927) [8].

2. Materials and Methods

2.1. Samples

To assess the utility of population analyses for the selection of representatives several Holocene–Recent collections of *Truncorotalia crassaformis* and Quaternary specimens from the type locality (Figure 1A) are studied. While a minuscule subset of data from the c. 5 myr history of the species, they sample it from different water masses in widely separated oceans which differ in their oceanographic setting, age, method of collection, and identifier. Discrimination of this species in the modern fauna is a current problem [9].

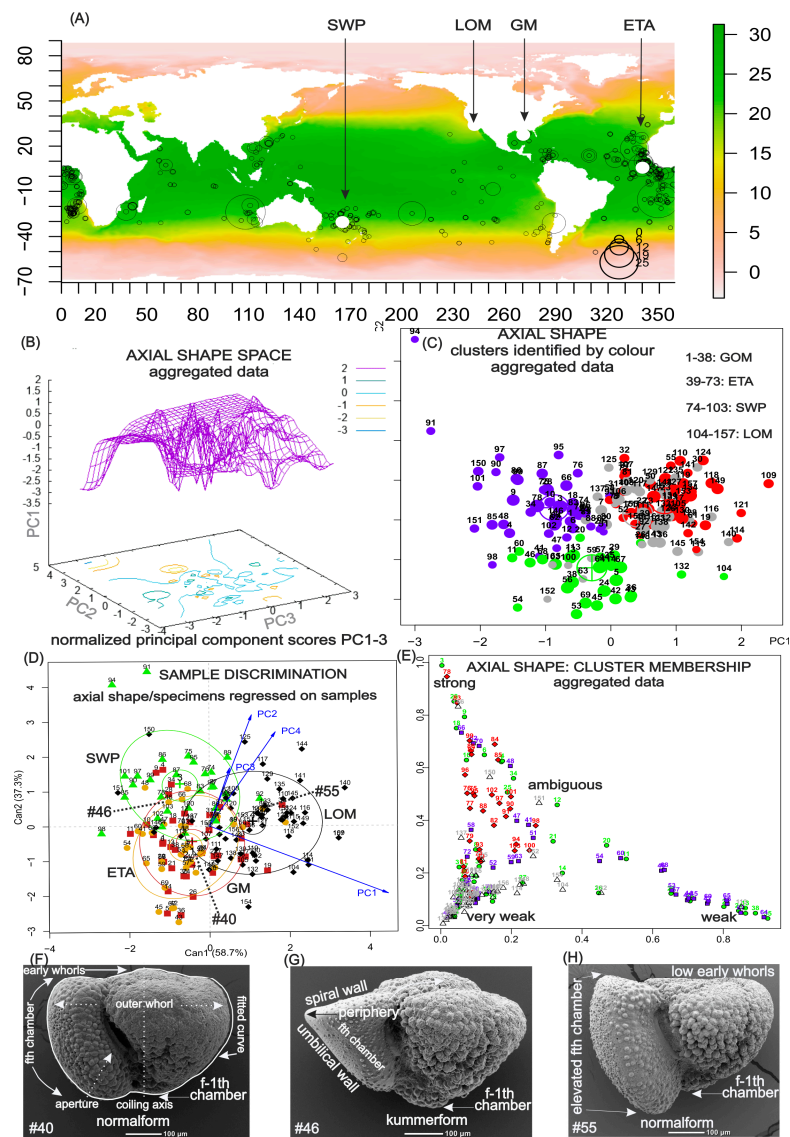


Figure 1. (A) Localities sampled in this study plotted on global sea surface temperatures from CARS 2009 [10]. Shown are the records of *Truncorotalia crassaformis* in the ForCenS database [11] with abundance $\geq 2\%$. (B) Principal component (PC1-3) projection using nearest-neighbour interpolation of the aggregated data on the axial outlines of specimens processed and normalized by the procrustes algorithm. (C) Fuzzy clusters in the data of (B) identified using the FKM algorithm [12]. (D) Generalized canonical discriminant analysis of a multivariate linear model of the collections using PC1-3 data of (B). (E) Strength of cluster membership in the analysis of (C). (F–H) Features of shell architecture in axial orientation and source of outline data.

DSDP Site 366A 1-1W-3-5 cm. (ETA): This site (05° 40.7 N; 19° 51 W; 2853 m) is on the Sierra Leone Rise in the eastern tropical Atlantic and lies under the Equatorial Counter Current. It is near Core 234 examined by [13]. From the model of [14], the age of the sampled horizon is <3 kyr. Thirty-five specimens from the >149 µm fraction were studied.

Gulf of Mexico sediment trap (GM): This trap is a time series (2008–2012) of foraminiferal and particulate flux at 700 m on the northern Gulf of Mexico continental shelf (27.5° N; 90.3° W; [15]). Thirty-eight specimens supplied by Dr. Caitlin Reynolds were studied from the GMT2-1 sediment trap (212 µm–425 µm fraction); they were collected between 21 and 27 April 2008.

DSDP Site 591A 1-1-1 cm (SWP): This site (31° 35.06 S; 164° 26.92 E; 2142 m) is on a southern spur of the eastern part of the Lord Howe Rise in the vicinity of the Tasman Front, a westward flowing branch of the East Australian Current. The age of the sample [16] is 9.4 kyr; 30 specimens were studied from the >149 µm fraction.

Lomita Marl, Los Angeles, California (LOM): The type locality of the species (Lomita Marl at Lomita Quarry, [17]) in urban Los Angeles is no longer exposed and specimens are rare; those analyzed by [17] are supplemented here by additional individuals supplied by Prof. James Ingle from other horizons at the type and adjacent Lomita Marl localities (157 specimens).

2.2. Methods

Although taxonomists take a holistic view of a specimen, experimental neuroscience shows that, subliminally, its outline is an important guide to its recognition [18,19] and cartoonists demonstrate its visual value. *Truncorotalia crassaformis* builds a shell by incremental addition of about 15 chambers that expand isometrically and are arranged in a low trochospiral of about 3 whorls. A view in the plane of the coiling axis (Figure 1F–H) encapsulates much of the ontogeny, including the rate of whorl translation (the height of early whorls), gross radial/axial dimensions, and the axial extension of late-formed chambers (the conical form). While this orientation captures only one plane of a 3D object, it provides a trait that relates to the hydrodynamic properties of the shell.

Specimen outlines were manually captured (tpsDig2 version 2.31, <https://life2.bio.sunysb.edu/morph/soft-dataacq.html> (accessed on 5 June 2024)) from SEM images as 180 equally spaced coordinates; use of binarizing algorithms is often more convenient. Raw data were processed using generalized procrustes analysis (GPA, R package shapes <https://cran.r-project.org/web/packages/shapes/index.html> (accessed on 5 June 2024)), which aligns specimens on their centroids and removes the size and positional differences. Principal component (PC) projections of the high-dimensional data for specimens onto 3 orthogonal axes retained 78% of the shape information and allowed for the inspection of the specimen configuration in low-dimensional Euclidean space. The `convhulln` function (R package geometry (<https://cran.r-project.org/web/packages/geometry/index.html> (accessed on 5 June 2024))) was used to construct the smallest convex hull for the samples in the PC1-3 space; the Delaunay triangulation of the points was plotted as a wireframe; clustering used the FKM algorithm [12]. Generalized canonical discriminant analysis (R package candisc, <https://cran.r-project.org/web/packages/candisc/index.html> (accessed on 5 June 2024)) provided a multivariate linear model of the samples using PC data.

‘Normalform’ (Figure 1F,H) refers to the specimens in which, through ontogeny, chambers increase in size with little change in shape. ‘Kummerform’ refers to specimens (Figure 1G) in which the last-formed (fth) chamber is dimensionally smaller than its predecessor (f-1th). Refer to Supplementary Materials for outline data and an R script for viewing them.

3. Results

The sampling design allows the data to be analyzed as either an aggregated dataset or as individual collections. The former approach, wherein inter-collection variation is neglected in the choice of representatives, is widely used in ML studies and is followed here. Viewed

on standardized PC1-3 coordinates, the shape landscape of the aggregated data is complex and diverse, consisting of multiple peaks and troughs within 2σ of the centroid and some distant individuals (Figure 1B). There is no well-defined central shape. Four groups based solely on shape (Figure 1C,E) are identified by unsupervised clustering [12], but they are weakly separated and each includes specimens from all samples. Certain morphotypes are distributed across the source populations; none are dominant. Canonical discriminant analysis (Figure 1D) of the PC data regressed on their source samples shows that most specimens are within the 68% data ellipse of their respective sample. Their partial intersection is a principal feature of the projection. While the aggregated data analyses suggest that a single, highly variable morphogroup is sampled, this analysis indicates that it may consist of overlapping subgroups. Eastward surface flows from the tropical Atlantic likely account for the closely adjacent plots of means and 68% data ellipses of the ETA and GM.

The aggregated data (Figure 1B) disregard specimen sources, thus emulating the sampling designs of [4,6]. Given the complex distribution of PC1-3 scores and the absence of a dominant morphotype, the data are pragmatically partitioned into three groups of differing suitability for selection of representatives based on distances from the centroid.

Partition 1—Specimens near the centre of the aggregated collections (Figure 2).

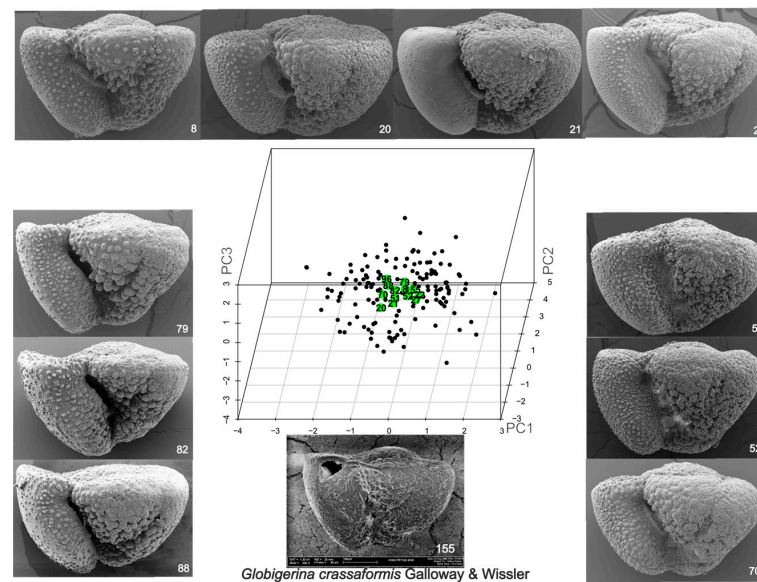


Figure 2. PC1-3 projection of aggregated data on axial outlines of specimens processed and normalized by the procrustes algorithm. Randomly selected specimens from the $\pm 0.5\sigma$ partition from the centroid are shown in green. Specimen #155 is the original holotype, now destroyed.

The centroid of the aggregated shape data (PC1-3, standardized scores) is located in a region where specimens are sparse. Only 14 specimens are included within $\pm 0.5\sigma$, whereas, if the aggregated data sampled a multivariate normal population, about 60 might be expected. However, this region does include specimens from all collections, with three to four from ETA, GM, and SWP, and one from LOM. As a group, all are normalforms (Figure 1F,H), or close thereto, but there are strong differences in the curvature of the outline at the periphery (Figure 1G) of some (e.g., #21 and #79). Indeed, the axial shape of #79 emulates that of *Truncorotalia truncatulinoides* (d’Orbigny, 1839) [20]) but has four rather than six chambers in the outer whorl.

Partition 2—Aggregated collections excluding rare morphotypes (Figure 3).

All specimens within $\pm 1.645\sigma$ of the centroid (nominally 90% of a multivariate normal population) are included (here, 79%). A subset of randomly drawn specimens from this model shows well-developed normalforms, several with an angular periphery and one in which it is curved, but the presence of kummerform specimens leads to much greater diversity in shape compared with Partition 1.

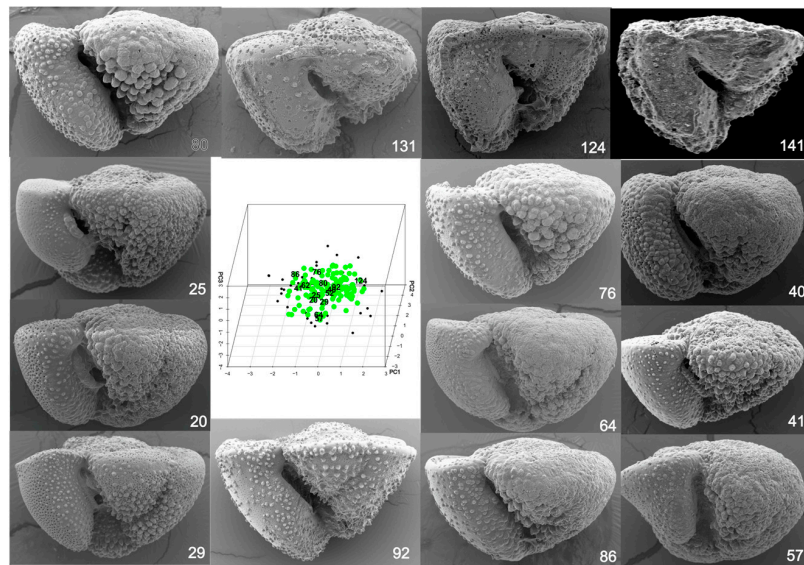


Figure 3. PC1-3 projection of aggregated data on axial outlines of specimens processed and normalized by the procrustes algorithm. Randomly selected specimens from the $\pm 1.645\sigma$ partition from the centroid are shown in green.

Partition 3—Specimens at the boundary of the aggregated collections (Figure 4).

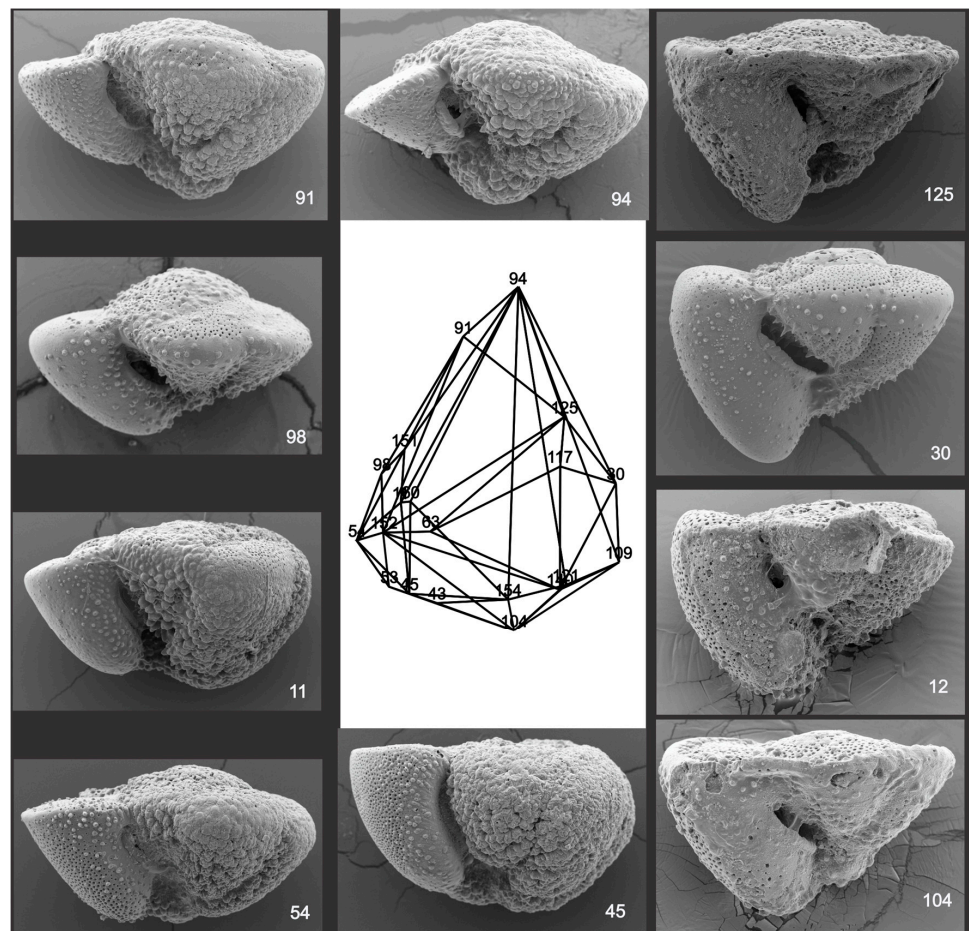


Figure 4. Selected specimens around the convex hull of the PC1-3 projection of aggregated data on axial outlines of specimens processed and normalized by the procrustes algorithm.

The convex hull locates the smallest convex enclosing space represented by the aggregated PC1-3 data; specimens lie at the vertices of the triangulated axial shape volume and the sides of the triangles are Euclidean inter-point distances. Because it defines the edges of the shape volume, highly diverse shapes are identified. The most remote specimen (#94), with all values for PC1-3 $> 2.8\sigma$, is a uniquely shaped kummerform. In contrast are individuals like #12 and #125 in which the conical form is similar to some in Partition 2 that are close to the convex hull.

4. Discussion

Axial shape data indicate that Quaternary–Recent *Truncorotalia crassaformis* likely comprises regional, water mass-related populations of highly variable morphology. Variation is complex and related to growth in latest ontogeny: similar architectures occur in each collection but no morphotype is dominant. Water mass isolation may contribute to inter-population differences in shape.

No particular claim is made about the suitability of values used for Partitions 1-2; other settings can be trialled. Partition 1 identifies specimens that experts might say represent the axial shape of *T. crassaformis* without qualification: its weak conical form is recognizable in all. Such specimens might be used in taxonomic atlases, as they are unlikely to confuse readers. However, as variation is greatly under-represented in the partition, its use as a training set is constrained. Partition 2 captures about 80% of the specimens. It could serve as a source of representatives for ML applications, as it filters out potentially controversial specimens at the margins of the shape space. Interestingly, it also excludes a few strongly conical normalforms that might be regarded as good examples of *T. crassaformis*. Partition 3 includes specimens that require interpretation by experts. Some, like #91 and #94, can be accommodated in the species, as they arise from the distortion of shape at the final growth stage. Others, like #12, #30, #104, and #125, have strongly developed conical form due to the axial extension of late-formed chambers. More problematic is #98, which is a normalform with an elliptical profile generated by a high spiral formed by early whorls and weak axial extension of the 6th chamber. It resembles *Globorotalia crassula* Cushman, Stewart, and Stewart (1930) [21], whose status is controversial, being regarded as a junior synonym of *T. crassaformis* by [22].

That training sets selected by experts serve as the basis for machine recognition of planktonic foraminiferal species draws attention to their subjective and often partial diagnoses. Recourse to selection of representatives by experts obfuscates the problem because their decisions are formed by visual judgement about individuals. It is a barren strategy for increasing knowledge of *T. crassaformis* as a morphospecies. This conclusion is likely applicable to other testate zooplankton. The alternative advocated here is to sample the morphology of the source populations. This has information on species architecture which is easily captured, analyzed morphometrically, and partitioned into sets suitable for the random selection of representatives. Useful diagnostic information should emerge when the sampling strategy is widened to allow for the inclusion of specimens that resemble the target species and unsupervised learning algorithms are applied.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/geosciences14060169/s1>, Scott geosciences zooplankters.RData contains outline coordinates for the 157 specimens in binary format. Scott geosciences zooplankters.R is a script with information on the structure and executable code for manipulating the outline data. It may be read as a text file.

Funding: This research received no external funding.

Data Availability Statement: data are contained within the article

Acknowledgments: I am most grateful to James Ingle, Stanford University, for contributing his collections from Lomita Quarry and nearby localities.

Conflicts of Interest: The author declares no conflict of interest

References

1. Deng, C.; Xunbi, J.; Rainey, C.; Zang, J.; Lu, W. Integrating machine learning with human knowledge. *iScience* **2020**, *23*, 101656. [[CrossRef](#)] [[PubMed](#)]
2. Ranaweera, K.; Bains, S.; Joseph, D. Analysis of image-based classification of foraminiferal tests. *Mar. Micropaleontol.* **2009**, *72*, 60–65. [[CrossRef](#)]
3. Al-Sabouni, N.; Fenton, I.S.; Telford, R.J.; Kučera, M. Reproducibility of species recognition in modern planktonic foraminifera and its implications for analyses of community structure. *J. Micropalaeontol.* **2018**, *37*, 519–534. [[CrossRef](#)]
4. Mitra, R.; Marchitto, T.M.; Ge, Q.; Zhong, B.; Kanakiya, B.; Cook, M.S.; Fehrenbacher, J.S.; Ortiz, J.D.; Tripathi, A.; Lobaton, E. Automated species-level identification of planktic foraminifera using convolutional neural networks, with comparison to human performance. *Mar. Micropaleontol.* **2019**, *147*, 16–24. [[CrossRef](#)]
5. Schiebel, R. Planktic foraminiferal sedimentation and the marine calcite budget. *Glob. Biogeochem. Cycles* **2002**, *16*, 1065. [[CrossRef](#)]
6. Hsiang, A.Y.; Brombacher, A.; Rillo, M.C.; Mleneck-Vautravers, M.J.; Conn, S.; Lordsmith, S.; Hull, P.M. Endless forams: >34,000 modern planktonic foraminiferal images for taxonomic training and automated species recognition using convolutional neural networks. *Paleoceanogr. Paleoclimatol.* **2019**, *34*, 1157–1177. [[CrossRef](#)] [[PubMed](#)]
7. Fenton, I.S.; Baranowski, U. and 22 others. Factors affecting consistency and accuracy in identifying modern macroperforate planktonic foraminifera. *J. Micropalaeontol.* **2018**, *37*, 431–443. [[CrossRef](#)]
8. Galloway, J.J.; Wissler, S.G. Pleistocene foraminifera from the Lomita Quarry, Palos Verde Hills, California. *J. Paleontol.* **1927**, *1*, 35–87.
9. Lessa, D.; Morard, R.; Jonkers, L.; Venancio, I.M.; Reuter, R.; Baumeister, A.M.; Albuquerque, A.; Kučera, M. Distribution of planktonic foraminifera in the subtropical South Atlantic: Depth hierarchy of controlling factors. *Biogeosciences* **2020**, *17*, 4313–4342. [[CrossRef](#)]
10. Ridgway, K.R.; Dunn, J.R.; Wilkin, J.L. Ocean interpolation by four-dimensional least squares—application to waters around Australia. *J. Atmos. Ocean. Technol.* **2002**, *19*, 1357–1372. [[CrossRef](#)]
11. Siccha, M.; Kučera, M. ForCenS, a curated database of planktonic foraminiferal species. *Sci. Data* **2017**, *4*, 170109. [[CrossRef](#)] [[PubMed](#)]
12. Ferraro, M.B.; Giordani, P.; Serafini, A. Fclust: An R package for fuzzy clustering. *R J.* **2019**, *11*, 198–210. [[CrossRef](#)]
13. Lidz, B. *Globorotalia crassaformis* morphotype variations in Atlantic and Caribbean deep-sea cores. *Micropaleontology* **1972**, *18*, 194–211. [[CrossRef](#)]
14. Lazarus, D.B.; Spencer-Cervato, C.; Pika-Biolzi, M.; Beckmann, J.-P.; von Salis, K.H.; Hilbrecht, H.; Thierstein, H.R. *Revised Chronology of Neogene DSDP Holes from the World Ocean*; Ocean Drilling Program, Texas A&M University, Technical Note 24, 1–301; International Ocean Discovery Program (IODP): College Station, TX, USA, 1995.
15. Richey, J.N.; Reynolds, C.E.; Tappa, E.; Thunell, R. Weekly resolution particulate flux from a sediment trap in the northern Gulf of Mexico, 2008–2012. In *United States Geological Survey Open-File Report 2014-1035*; US Geological Survey: Reston, VA, USA, 2014; Volume 9, pp. 1–12.
16. Nelson, C.S.; Hendy, C.H.; Cuthbertson, A.M. *Compendium of Stable Oxygen and Carbon Isotope Data for the Late Quaternary Interval of Deep-Sea Cores from the New Zealand Sector of the Tasman Sea and Southwest Pacific Ocean*; Occasional Report 16; University of Waikato, Department of Earth Sciences: Hamilton, New Zealand, 1993; pp. 1–87.
17. Scott, G.H.; Ingle, J.C., Jr.; McCane, M.; Powell, C., II; Thunell, R.C. *Truncorotalia crassaformis* from its type locality: Comparison with Caribbean plankton and Pliocene relatives. *Mar. Micropaleontol.* **2015**, *112*, 1–12. [[CrossRef](#)]
18. Spröte, P.; Schmidt, F.; Fleming, R.F. Visual perception of shape altered by causal history. *Sci. Rep.* **2016**, *6*, 36245. [[CrossRef](#)] [[PubMed](#)]
19. Elder, J.H. Shape from contour: Computation and representation. *The Annual Review of Vision Science* **2018**, *4*, 423–450. [[CrossRef](#)] [[PubMed](#)]
20. d’Orbigny, A. Foraminifères des Iles Canaries. In *Histoire naturelle des Iles Canaries*; Barker-Webb, P., Berthelot, S., Eds.; Bethune: Paris, France, 1839; pp. 120–146.
21. Cushman, J.A.; Stewart, R.E.; Stewart, K.C. Tertiary foraminifera from Humboldt County, California. *Trans. San Diego Soc. Nat. Hist.* **1930**, *6*, 41–94.
22. Schiebel, R.; Hemleben, C. *Planktic Foraminifers in the Modern Ocean*; Springer-Verlag: Berlin/Heidelberg, Germany, 2017; pp. 1–358. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.