PNAS

Supplementary Information 1 for

Trophic convergence of marine vertebrate communities worldwide.

Juan David González-Trujillo,[1,2,3] Jorge Assis,[4,5] Ester Serrão,[4] Mark John Costello[5], Eliza Fragkopoulou[4], Manuel Mendoza[3], Miguel B. Araújo[1,2,6]

[1] 'Rui Nabeiro' Biodiversity Chair, MED – Mediterranean Institute for Agriculture, Environment and Development & CHANGE – Global Change and Sustainability Institute, Universidade de Évora, Largo dos Colegiais, 7004-516 Évora, Portugal

[2] Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, Calle Jose Gutierrez Abascal, 2, 28006 Madrid, Spain

[3] Universidad Nacional de Colombia, Sede Bogotá, Facultad de Ciencias, Departamento de Biología, Cra 30 45 03, Ciudad universitaria, Bogotá, 111321, Colombia

[4] Centre of Marine Sciences (CCMAR/CIMAR LA), Universidade do Algarve, Faro, Portugal

[5] Faculty of Biosciences and Aquaculture, Nord University, Bodo, Norway

[6] Theoretical Sciences Visiting Program, Okinawa Institute of Science and Technology Graduate University, Onna, 904-0495, Japan.

*Corresponding authors: Juan David González-Trujillo, Miguel B. Araújo

Email: jdgonzalezt@gmail.com; maraujo@mncn.csic.es

This PDF file includes:

Supplementary text with an extended description of methods.
Figures S1 to S3.

Extended methods

The workflow to draw the global biogeography of marine trophic communities consists of four steps (Figure SM1.1):

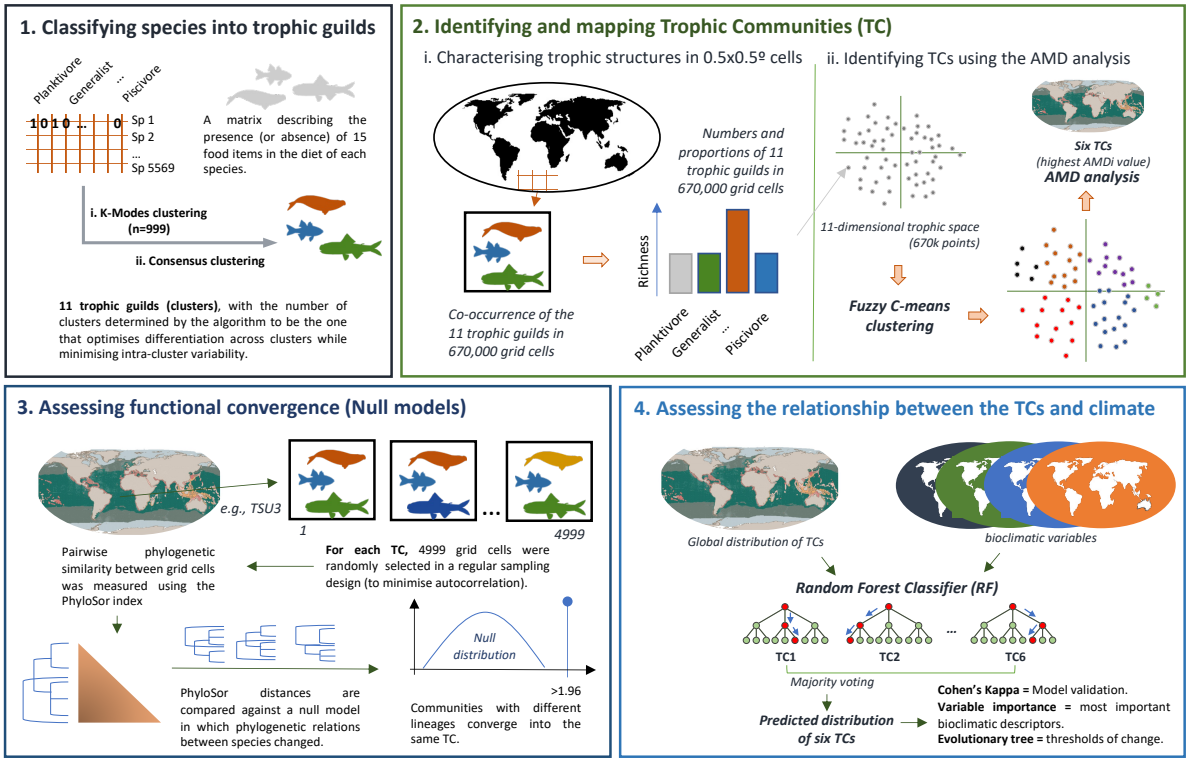

Figure SM1.1. Workflow describing the approach used for drawing the global distribution of marine trophic communities (TCs). The modelling approach consists of four stages: 1) classifying species into trophic guilds based on diet information; 2) identifying and mapping TCs globally on a grid cell surface based on the AMDi analysis; 3) assessing the convergence of communities into "functionally analogous" trophic communities irrespective of the evolutionary origins of their constituent species using Null models; and 4) training random forest classifiers and evolutionary trees to assess the relationship between climatic conditions and TC distribution across the globe.

1. Classifying species intro trophic guilds.

We used available information of species diets and unsupervised k-mode clustering to classify the 5,569 marine vertebrate species into trophic guilds. The diets of the 5569 species were characterized based on a comprehensive review of the available databases: SeaLifeBase (1) and Fishbase (2), which contain records of species consumed by marine mammals and fishes of the world based on observations, stomach contents and metabarcoding. We classified each of the consumed species into fifteen resource types: Detritus, Benthic Crustaceans, Cnidarians, Echinoderms, Sponges and Tunicates, Mollusks, Worms, non-crustacean benthic invertebrates, Benthic Algae and Weeds, Phytoplankton, Planktonic Crustaceans, Jellyfish and Hydroids, other planktonic invertebrates, Fish and non-Fish Vertebrates. Then, we assigned values of 1 to all types of resources consumed by every species, and zero otherwise. We decided to not give weights to diet items (i.e., assigning a value of '4' to the fish item given that the species is reported to consume four different species) given that sampling effort to characterize diets is not consistent across species. The resulting incidence matrix, with the 15 trophic resources as columns and the names of the species as rows, is included in the ZENODO (https://zenodo.org/doi/10.5281/zenodo.8375585) repository companion to this paper.

We established the trophic guilds with k-modes clustering (3) based on the simple-matching distance between the 5569 species in the 15-dimensional space defined by the presence of each type of resource in their diet (a vector of dimension 15) (Fig. SM1.1.1). Clustering analyses were performed using R statistical software with the package 'klaR' (4). As in the case of the k-means algorithm, the

results obtained using k-modes are sensitive to the location of the initial centroids. Thus, the same species might be assigned to a distinct trophic guild if the algorithm changes starting points in a second iteration. To give more consistency to the classification, we performed 999 classifications and used the 'diceR' package (5) to build a final consensus classification. The optimal number of trophic guilds (11 clusters) was identified by assessing the minimum number of clusters at which the total difference within clusters is minimized (Figure SM1.2). Trophic guilds were named after the resource or resources that were consistently consumed by all species in the same guild (see Fig. 1 in the main text).

All the lines of code required to reproduce this step are available at the Zenodo repository. Script name: "A. Trophic guilds"

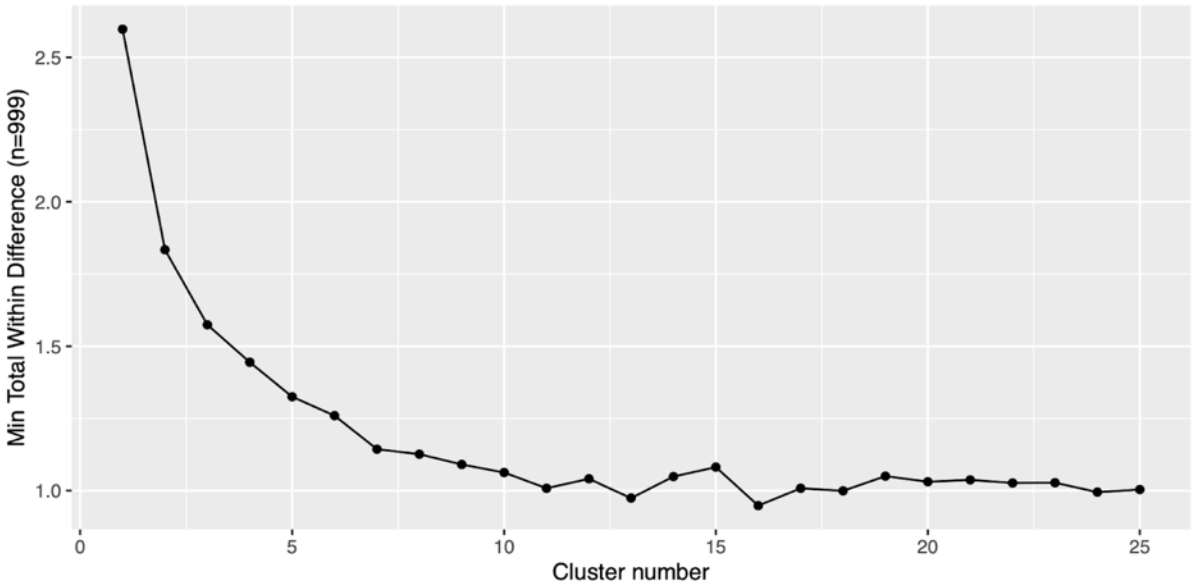

Figure SM1.2 Total within-cluster variance at different clustering solutions using the k-modes algorithm. The Optimum number of clusters (i.e., trophic guilds) was identified as the value (k =11) where the total within difference is minimized after 999 repetitions.

2. Identifying and mapping Trophic Communities (TC).

We followed the approach developed by Mendoza and Araújo (6) to identify similar types of trophic communities ('TC') across local communities in the global ocean. The approach of Mendoza and Araújo (6) consists of two major steps (Fig. SM1.1.2): (i) counting the number of species per trophic guild, hereafter called 'the trophic profile' of the community; and (2) estimating the optimal number of clusters (=TCs) in which trophic profiles can be classified, hereafter called 'trophic communities' (TC), by using a combination of C-means clustering and average mean decrease (AMD) analyses.'

In the first step, we characterized the trophic profile of vertebrate communities within 0.05 x 0.05º grid cells using species distribution data. This involved classifying species into trophic guilds and quantifying the number and proportions of trophic guilds for each grid cell. The resulting data matrix consisted of 668,602 grid cells as rows and 11 trophic guilds as columns, with every cell representing a point in an 11-dimensional 'trophic space' defined by the number and proportions of trophic guilds per grid cell.

In the second step, we identified the presence of groups of cells with similar trophic communities. We expected to find clusters of local communities (=grid cells) with higher similarity in terms of the number and proportions of co-occurring trophic guilds across the world. Such an expectation was assessed using AMD analysis (see (7)), in which the optimal number of clusters is determined by tracking the AMDi values through a series of C-means fuzzy clustering analysis runs with an increasing number of user-defined clusters (from 2 to 20). The AMDi, which quantifies cluster 'compactness,' is computed by averaging the highest membership degree across all samples obtained through the C-means fuzzy clustering algorithm. The algorithm assigns each data point a membership degree to each cluster based

on its Euclidean distance to the cluster center. The AMDi value ranges from 0 to 1, where 0 indicates a random sample space without discernible clusters, and 1 represents a space where all samples align perfectly with the centroid of a cluster.

After running 200 replicates for each user-defined number of clusters and tracking AMDi values, we found that AMDi peaked at the number 6, indicating that global grid cells can be classified into six clusters (=TCs) (Figure SM1.3). In this classification scheme, similarity between communities in terms of their trophic profiles is higher within each cluster than when comparing across clusters. In other words, communities within each TC are closely similar in terms of the number and proportions of co-occurring trophic guilds.

Finally, to assess the extent of divergence among cells regarding the mean number and proportion of trophic guilds that typify each trophic community, we employed the Bray-Curtis dissimilarity index. The index thus enabled us to quantify the extent to which the number and proportion of trophic guilds co-occurring in each cell diverged from the average number and proportion of trophic guilds (centroid) characterizing each of the six trophic communities identified with the AMD index (Figure SM1.4). A value approaching zero indicates that the guild numbers and proportions are in close alignment with the average numbers and proportions that typify a given trophic community. In contrast, a value approaching one suggests a divergence from the average numbers and proportions.

All the lines of code required to reproduce this step are available at the Zenodo repository. Script name: “B. Trophic structures.”

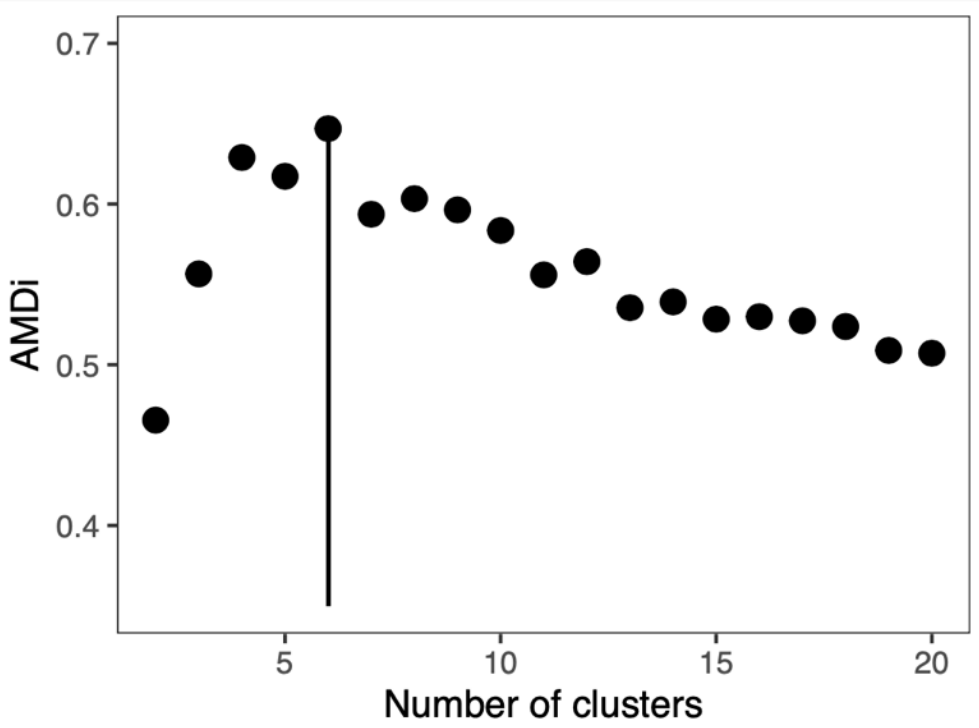

Figure SM1.3. AMD*i* curve obtained with 668 602 assemblages of 11 trophic guilds. Line point to the number of clusters i which the highest cluster definition is obtained after 999 repetitions.

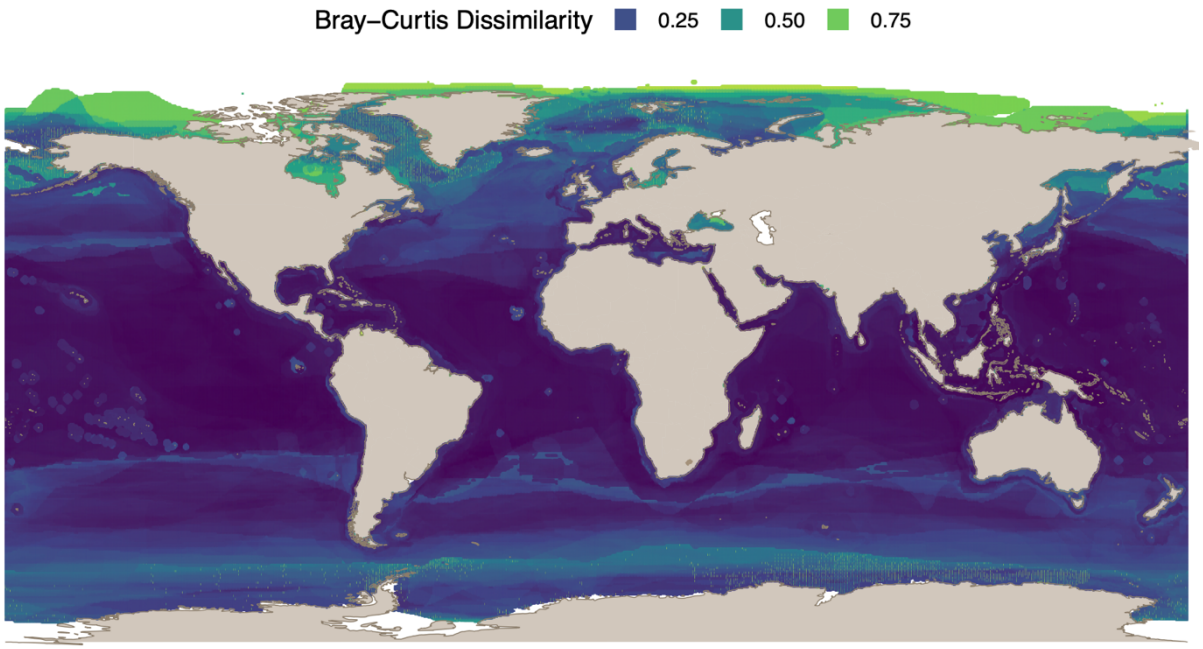

Figure SM1.4. The degree of divergence of each cell with respect to the average number and proportion of trophic guilds characterizing trophic communities is measured by means of the Bray-Curtis dissimilarity index. A value approaching zero indicates that the guild numbers and proportions are in close alignment with the average numbers and proportions that typify a given trophic community; conversely, a value approaching one suggests divergence from the average numbers and proportions.

3. Assessing functional convergence within TCs.

We built a null model to test whether communities of each tend to converge into analogous trophic configurations, irrespective of the evolutionary origins of their constituent species. In total, we built six models, one per TC.

Owing to the immense number of grid cells, and to avoid the effect of spatial autocorrelation, we implemented null models for a random, regularly spaced sample of 4,999 grid cells for each TC. For each sample, the proportion of shared lineages between local communities was measured using the PhyloSor pairwise dissimilarity index (8). This index measures the phylogenetic difference between two communities based on a comparison of the length of the tips of the phylogenetic tree, with a value of 1 if the communities share no tips (lineages) and 0 if they share all lineages. Then, following Leprieur and colleagues (9), for each pairwise comparison, a null distribution of PhyloSor values was generated by randomizing species along the vertebrate phylogeny 999 times, keeping species richness and species composition constant. Then, using the mean and standard deviation of the null distribution, a standardized effect size (SES) was calculated as follows:

$$SES = \frac{X_{obs} - \bar{X}_{null}}{SD(X_{null})}$$

where X_{obs} is the observed PhyloSor value, X_{null} is the mean of the null distribution, and SD (X_{null}) is the standard deviation of the null distribution. SES values greater than 1.96 indicate higher phylogenetic dissimilarity than expected from the species composition. This means that a pair of communities differ in the number and type of lineages that coexist, providing evidence that communities with different lineages could converge towards an analogous trophic configuration.

All the lines of code required to reproduce this step are available at the Zenodo repository. Script name: “D. Null models.”

4. Assessing the relationship between climate and TCs' distribution.

We trained two machine learning algorithms (Random Forests (10) and Evolutionary classification trees) to relate the distribution of TC to a set of ecologically meaningful environmental predictors.

We trained 99 RF classifiers using subsets of the data with a balanced number of samples per trophic community given that the number of grid cells is highly unbalanced across TCs (e.g., the prevalence of TC1 to 3 is above 0.3 and of TC4 to 6 is below 0.05). Subsets were created by randomly sampling 2,000 grid cells per TC without replacement. Moreover, as the distribution of trophic communities is expected to be spatially structured, there is a risk that model evaluation measures are inflated due to spatial autocorrelation (11). To limit such a risk, we evaluated the performance of RF classifiers using a spatial block cross-validation with geographically independent test data (12). Specifically, we split the globe into four blocks covering a similar number of grid cells per TC and fitting models that included the cells from all bins except the one used for testing (Figure SM1.4).

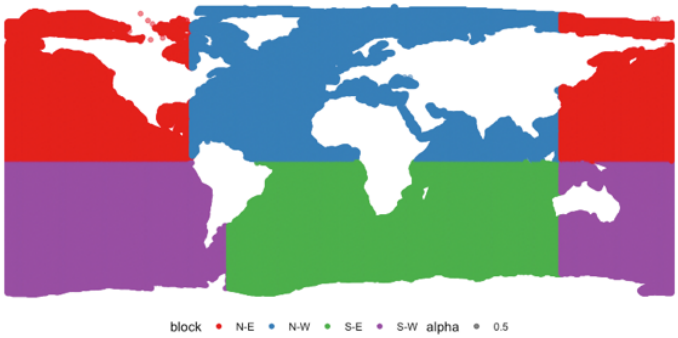

Figure SM1.5. Extent and location of the spatial blocks used in the cross-validation process.

At the end of this sequential process, we estimated the Cohen's Kappa coefficient to determine the overall model performance. This index compares observed (Fig. SM1.5a) and predicted (Fig. DM1.5b) grid cell distributions to determine the percentage of correct classification. All RF classifiers were trained with the algorithm available in the 'RandomForest' package (13).

We assessed the importance of environmental predictors to discriminate between types of trophic communities using the mean decrease in accuracy and the Gini index. The most important variables (i.e., the ones with the highest mean decrease inaccuracy and the Gini index, Fig. 4a) were used to train evolutionary learning of globally optimal classification trees ('ECT', (14)) with the 'evtree' package (15). The advantage of ECT over ensemble methods such as RF is their comprehensibility, as they provide a unique decision tree in which environmental thresholds can be identified.

All the lines of code required to reproduce this step are available at the Zenodo repository. Script name: "C. RF and Classification trees."

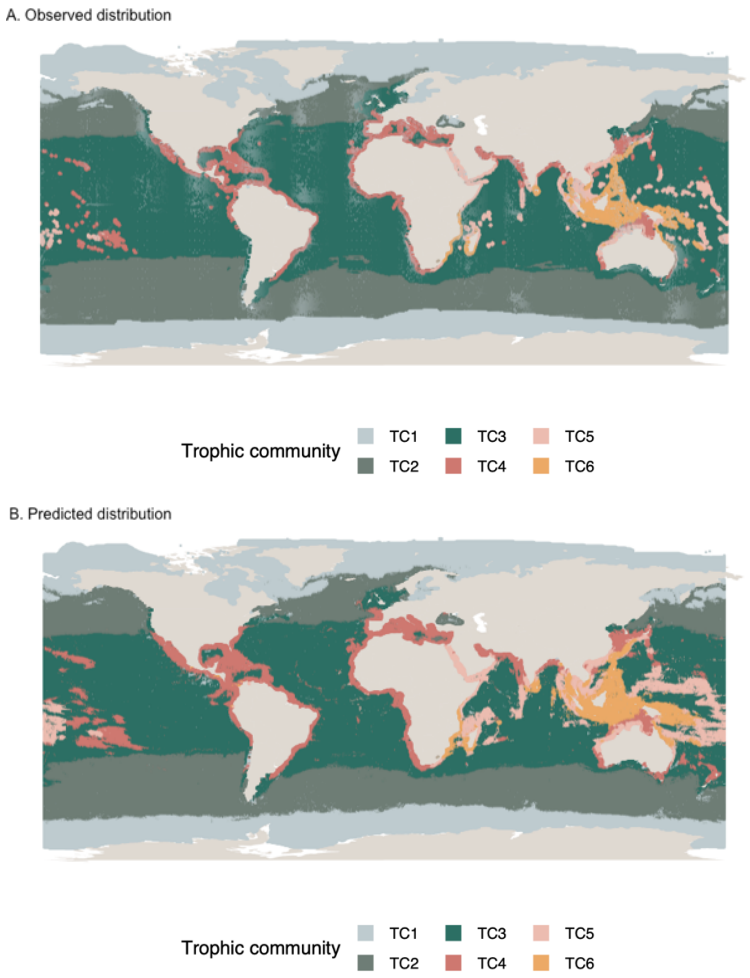

Figure SM1.6. Observed and predicted distributions of marine trophic communities (TCs) using 668 602 assemblages of 11 trophic guilds. A. Observed distribution of the six TCs in the present-day. B. Distribution of the six TCs in the present-day as predicted by the Random Forest Classifier with the highest Cohen's Kappa value (0.56).

References

1. M. L. D. Palomares, D. Pauly, SeaLifeBase. Deposited 2023.

2. R. Froese, D. Pauly, FishBase. Deposited 2023.

3. A. Chaturvedi, P. E. Green, J. D. Caroll, K-modes Clustering. *J. of Classification* **18**, 35–55 (2001).

4. C. Roever, *et al.*, Package 'klaR.' *ftp. rediris. org/mirror/CRAN/web/packages/klaR/klaR. pdf (Last viewed June 10, 2020)* (2020).

5. D. S. Chiu, A. Talhouk, diceR: an R package for class discovery using an ensemble driven approach. *BMC Bioinformatics* **19**, 11 (2018).

6. M. Mendoza, M. B. Araújo, Climate shapes mammal community trophic structures and humans simplify them. *Nat Commun* **10**, 5197 (2019).

7. M. Mendoza, M. B. Araujo, Biogeography of bird and mammal trophic structures. *Ecography* **2022**, e06289 (2022).

8. J. A. Bryant, *et al.*, Microbes on mountainsides: Contrasting elevational patterns of bacterial and plant diversity. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 11505–11511 (2008).

9. F. Leprieur, *et al.*, Quantifying Phylogenetic Beta Diversity: Distinguishing between 'True' Turnover of Lineages and Phylogenetic Diversity Gradients. *PLoS ONE* **7**, e42760 (2012).

10. L. Breiman, Random forests. *Machine learning* **45**, 5–32 (2001).

11. P. Segurado, M. B. Araujo, W. E. Kunin, Consequences of spatial autocorrelation for niche-based models. *J Appl Ecology* **43**, 433–444 (2006).

12. D. R. Roberts, *et al.*, Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017).

13. A. Liaw, M. Wiener, Classification and regression by randomForest. *R news* **2**, 18–22 (2002).

14. M. Krętowski, M. Grześ, "Global learning of decision trees by an evolutionary algorithm" in *Information Processing and Security Systems*, K. Saeed, J. Pejaś, Eds. (Springer-Verlag, 2005), pp. 401–410.

15. T. Grubinger, A. Zeileis, K.-P. Pfeiffer, **evtree** : Evolutionary Learning of Globally Optimal Classification and Regression Trees in *R*. *J. Stat. Soft.* **61** (2014).

PNAS

Supplementary Information 2 for

Trophic convergence of marine vertebrate communities worldwide.

Juan David González-Trujillo,[1,2,3] Jorge Assis,[4,5] Ester Serrão,[4] Mark John Costello[5], Eliza Fragkopoulou[4], Manuel Mendoza[3], Miguel B. Araújo[1,2,6]

[1] 'Rui Nabeiro' Biodiversity Chair, MED – Mediterranean Institute for Agriculture, Environment and Development & CHANGE – Global Change and Sustainability Institute, Universidade de Évora, Largo dos Colegiais, 7004-516 Évora, Portugal

[2] Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, Calle Jose Gutierrez Abascal, 2, 28006 Madrid, Spain

[3] Universidad Nacional de Colombia, Sede Bogotá, Facultad de Ciencias, Departamento de Biología, Cra 30 45 03, Ciudad universitaria, Bogotá, 111321, Colombia

[4] Centre of Marine Sciences (CCMAR/CIMAR LA), Universidade do Algarve, Faro, Portugal

[5] Faculty of Biosciences and Aquaculture, Nord University, Bodo, Norway

[6] Theoretical Sciences Visiting Program, Okinawa Institute of Science and Technology Graduate University, Onna, 904-0495, Japan.

*Corresponding authors: Juan David González-Trujillo, Miguel B. Araújo

Email: jdgonzalezt@gmail.com; maraujo@mncn.csic.es

This PDF file includes:

Supplementary text with an extended description of the sensitivity analysis and the correspondence with other biogeographical classification systems.
Table SM2.1.
Figures SM2.1 to SM2.2.

SENSITIVITY ANALYSIS

The information on species diet is far from comprehensive and accurate. Despite using the most up-to-date and complete data available and homogenizing it into the coarsest possible dietary categories to avoid potential biases, we acknowledge that additional future data may reveal that one or more species have more generalist diets than currently reported.

In our analysis, several species were classified within specialist trophic guilds, such as pelagic or benthic specialist invertivores, based on the presence of one or two items in their guts. Therefore, the classification of these species may change as additional stomach contents are analyzed. To evaluate the impact of such a reclassification, a sensitivity analysis was conducted using a dataset in which species with the fewest food items were excluded from the analysis. Specifically, we excluded those species with fewer than three, six, 10, 12, 15, 18 or 21 food items recorded in the database and re-ran the entire pipeline to obtain the classification of trophic communities.

Given that some species might also be erroneously classified, and that the total number of species would be significantly reduced (by almost 50% when considering those with 21 or more food items), we proceeded to create four additional datasets. Two datasets were created in which 25 and 50% of the specialist species with 3, 6, 10, 12, 15, 18 or 21 food items recorded in the database were reclassified as generalists. Two further datasets were created in which 25 and 50% of the specialist species with 3, 6, 10, 12, 15, 18 or 21 food items recorded in the database were reclassified as "specialist-generalists" (see Table SM2.1).

Table SM2.1. Datasets and scenarios used in the sensitivity analysis.

Dataset	Description
Data removal: Less accurate species	This dataset represents a situation in which specialists, such as pelagic invertivores, are assumed to be misclassified due to a lack of accurate gut information. It includes seven additional subsets of data in which specialist species with 3, 6, 10, 12, 15, 18 or 21 food items recorded in the database were removed.
Reclassification: 25% as generalists	This dataset represents a situation in which 25% of specialists (such as pelagic invertivores) are assumed to have been misclassified due to the absence of accurate gut information. The dataset comprises seven additional subsets of data, in which specialist species with 3, 6, 10, 12, 15, 18, or 21 food items recorded in the database have been reclassified as 'generalists'.
Reclassification: 50% as generalists	This dataset represents a situation in which 50% of specialists (such as pelagic invertivores) are assumed to have been misclassified due to the absence of accurate gut information. The dataset comprises seven additional subsets of data, in which specialist species with 3, 6, 10, 12, 15, 18, or 21 food items recorded in the database have been reclassified as 'generalists'.
Reclassification: 25% as specialist-generalists	This dataset represents a situation in which 50% of specialists (such as Pelagic invertivore) are presumed to have been misclassified due to an absence of accurate gut information. The dataset comprises seven additional subsets of data, in which specialist species with 3, 6, 10, 12, 15, 18 or 21 food items recorded in the database have been reclassified as “specialist-generalists.” To illustrate, a pelagic invertivore specialist or a pelagic piscivore-invertivore species has undergone reclassification, becoming instead a pelagic generalist invertivore.
Reclassification: 50% as specialist-generalists	This dataset represents a situation in which 50% of specialists (such as Pelagic invertivore) are presumed to have been misclassified due to an absence of accurate gut information. The dataset comprises seven additional subsets of data, in which specialist species with 3, 6, 10, 12, 15, 18 or 21 food items recorded in the database have been reclassified as “specialist-generalists.” To illustrate, a pelagic invertivore specialist or a pelagic piscivore-invertivore species has undergone reclassification, becoming instead a pelagic generalist invertivore.

To assess the robustness of the trophic biogeography scheme (Figure 2 in the main manuscript), an element-centric comparison (1) was conducted to determine the extent to which the final classification would change if the aforementioned five scenarios were to occur. Element-centric comparison focuses

on common memberships between data elements (e.g., grid cells) induced by the cluster structure rather than overlaps between clusters induced by elements. The agreement and consistency scores were employed to quantify the robustness of the algorithm. The agreement score is used to assess the degree of regular grouping of elements across iterations with respect to the reference clustering (i.e., the final clustering solution). Values closer to 1 indicate that the final classification scheme is similar to that obtained in each scenario. The consistency score within a set of clusters reflects the consistency across the seven subsets of each data set. Values closer to 1 indicate that a given grid cell is consistently classified as belonging to the same cluster across the seven scenarios of each dataset.

The sensitivity analysis indicated that the overall classification of trophic communities is robust to potential misclassification by specialists due to the lack of precision of the data. The median values of 'agreement' for the five datasets are greater than 0.75 (Fig. SM2.1A), indicating that the same biogeographic classification scheme can be obtained by using a smaller number of species or a dataset in which generalist species are misclassified as specialists. Notably, coastal zones exhibited higher agreement values, whereas zones of lower agreement were observed in the transitions between pelagic TCs (compare Figs. 2 and SM2.2).

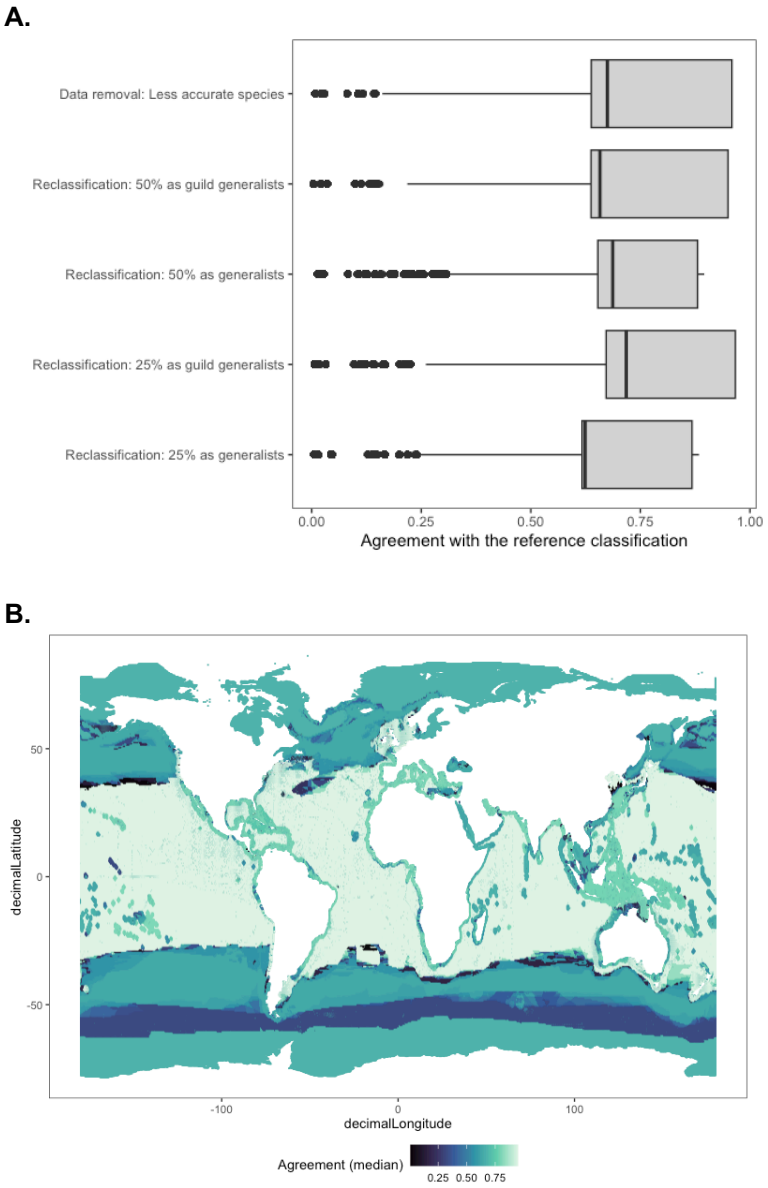

Figure SM2.1. The sensitivity of the biogeographic classification to changes in species numbers and trophic guild identity. A. Boxplot summarizing the degree of agreement between the final classification (Figure 2 in the main text) and classifications emerging from datasets in which the number of species

was reduced with either insufficient diet information or a reclassification in guild identity (Details of the datasets can be found in table SM2.1). B: The median agreement value for each grid cell, calculated as the median agreement obtained for each of the subsets of data created in the sensitivity analysis.

The sensitivity analysis also indicates that the classification scheme remains largely consistent irrespective of whether species with a low (n=3) or medium (n=21) level of completeness in their diet contents information are excluded or reclassified. The median values for the seven scenarios were all above a value of 0.8 (Figure SM2.2A), indicating a high degree of consistency across all the produced datasets.

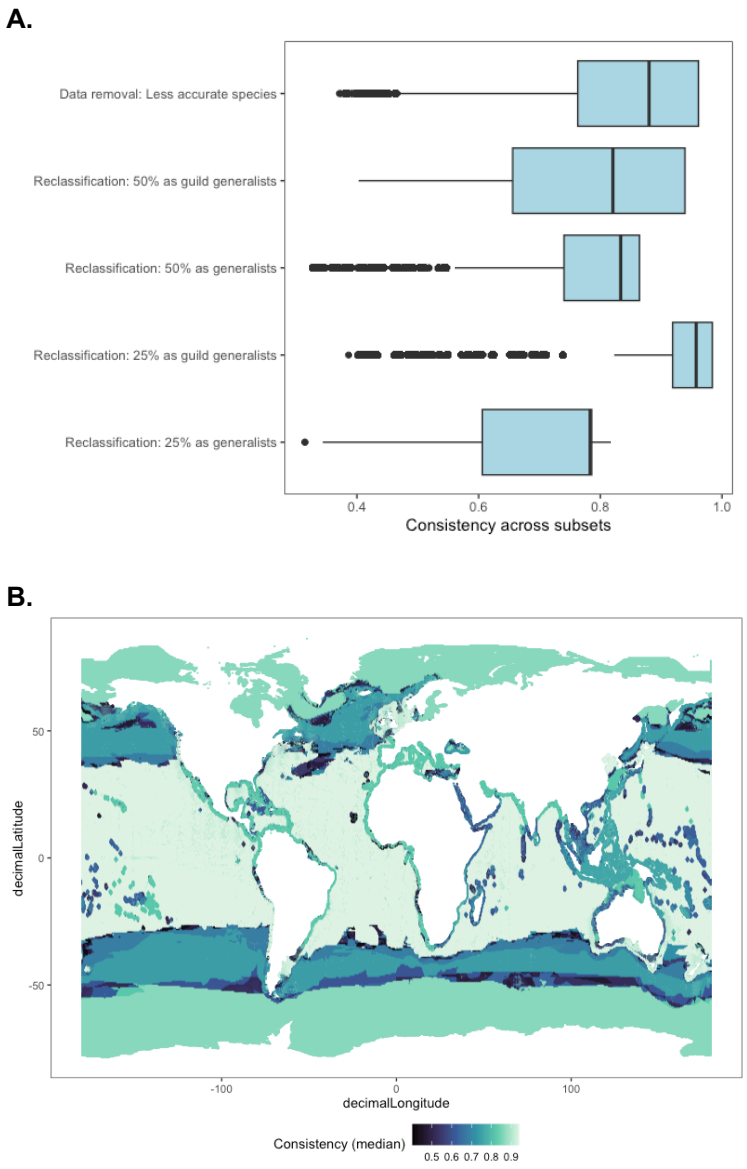

Figure SM2.2. The sensitivity of the biogeographic classification to changes in species numbers and trophic guild identity. A. Boxplot summarizing the degree of consistency across classifications emerging from datasets in which the number of species was reduced with either insufficient diet information or a reclassification in guild identity (Details of the datasets can be found in table SM2.1). B: The median

consistency value for each grid cell, calculated as the median agreement obtained for each of the subsets of data created in the sensitivity analysis.

All the lines of code required to reproduce this step are available at the Zenodo repository (https://zenodo.org/doi/10.5281/zenodo.8375585). Script name: "E. Sensitivity analysis."

Influence of large migrant species on the classification scheme

We also evaluated the influence of large migrant species on the trophic regionalization scheme we have generated. To this end, we excluded the 109 species identified by Bentley et al. (2) as large migrants from the original dataset. With the aforementioned dataset, which excluded large migrants, we conducted an additional AMDi analysis (see Methods and Extended Methods in Supplementary Material 1) to generate clusters with comparable numbers and proportions of trophic guilds. The results were found to be qualitatively and quantitatively similar to those obtained with the complete dataset. Six community trophic communities with a similar global distribution were obtained (fig. SM2.3), which, according to the Kappa index close to 0.9, corresponds to the one produced from the original dataset.

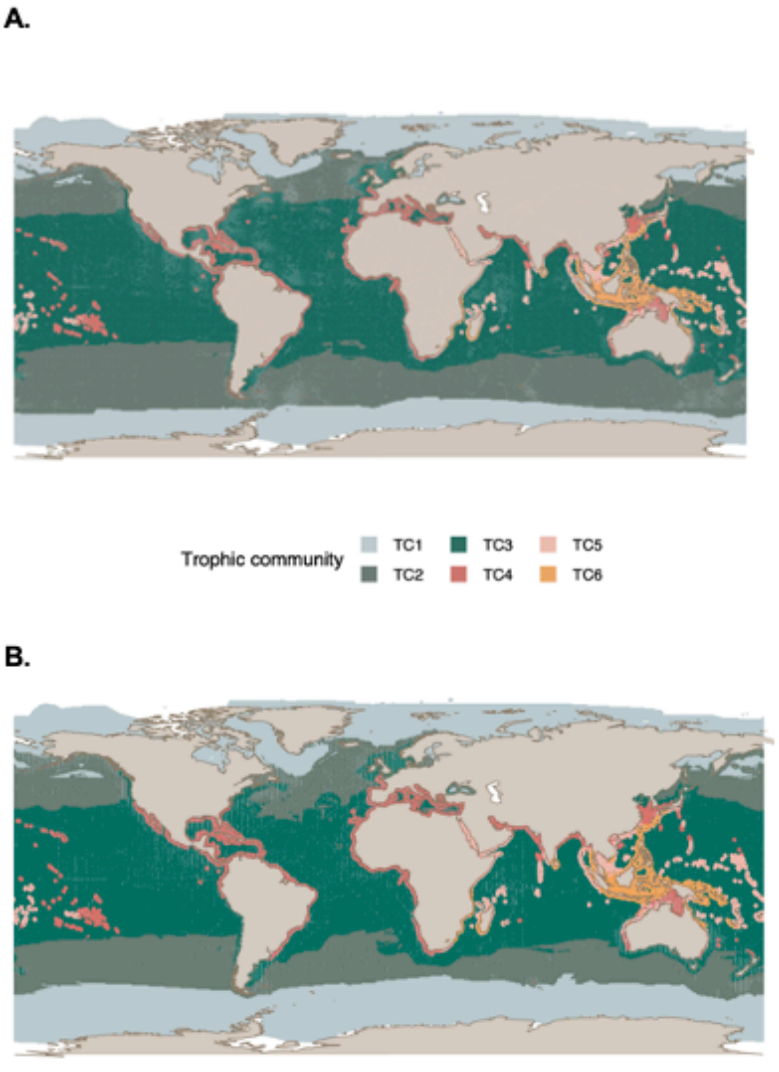

Figure SM2.3. Global distribution of marine trophic communities including (A) and excluding (B) large migrant species.

All the lines of code required to reproduce this step are available at the Zenodo repository (https://zenodo.org/doi/10.5281/zenodo.8375585). Script name: "F. Large Migrant Species."

CORRESPONDENCE WITH OTHER BIOGEOGRAPHICAL CLASSIFICATION SCHEMES

To assess alignment between our regionalization framework and conventional biogeographic schemes centered on species identity, we juxtaposed spatial distribution of with three established biogeographic frameworks: Marine Ecoregions of the World (MEOW) (3), Pelagic Provinces of the World (PPOW) (3), and Marine Biogeographic Realms based on species endemicity (BMRE) (4). These biogeographic classifications were selected because they are among the most widely used frameworks in marine biogeography, are based on species endemicity and diversity, and provide open-access data, facilitating reproducibility and ensuring compatibility with our analyses. Furthermore, because these frameworks are based on species endemicity and diversity, they also allow us to indirectly test whether communities with different species composition are functionally analogous.

We used the largest hierarchical level (realms) for spatial overlap analysis, as this represents very large regions where biota at higher taxonomic levels are internally coherent because of a shared evolutionary history. MEOW, the first comprehensive classification system, was formulated to categorize global coastlines and shelves using a nested arrangement of realms, provinces, and ecoregions informed by Briggs' contributions *(5)* and expert criteria. PPOW expanded upon MEOW, including 37 pelagic provinces across four overarching realms, to encompass global surface pelagic waters. The combined scheme (MEOW + PPOW) consisted of 16 realms covering continental shelves and offshore waters. Finally, BMRE introduced a data-centric methodology, delineating 30 marine realms based on the distribution patterns of 65,000 marine species, with 18 realms on continental shelves and 12 in offshore waters, acknowledging broader species distribution ranges in pelagic and deep-sea ecosystems compared to coastal domains.

The realms of these three regionalization schemes are completely or partially nested within the distribution ranges of the six types of trophic communities (‘TCs’) (Figure SM2.4). Our trophic-based regionalization includes half of the coastal and shelf areas in the 12 Marine Ecoregions of the World (MEOW) scheme; approximately one-third of the 16 coastal, shelf, and pelagic areas in the updated scheme that includes the Pelagic Provinces of the World (PPOW); and a fifth of the 30 Marine Biogeographic Realms based on the Species Endemicity (BMRE) scheme.

Regarding the MEOW and PPOW schemes, TCs 1 and 2 enclosed the coastal Arctic, Southern Ocean, and Temperate Northern Pacific and Atlantic realms (Fig. SM2.4a), as well as the pelagic Arctic, Southern and Northern Cold-Water realms from higher latitudes (Fig. SM2.4b). Likewise, TC3 covered the pelagic and offshore provinces in equatorial waters, namely the Indo-Pacific and Atlantic Warm Water realms (Fig. SM2.4a), coastal Tropical Atlantic, and Central, Eastern and Western Indo-Pacific realms (Fig. SM2.4b). Notably, realms in coastal and shelf regions are nested within the extension of TCs 4 to 6, even if they are far apart in space (e.g., the Tropical Atlantic, and Temperate Northern Pacific realms are partially nested within TC4).

A similar pattern was observed when assessing the concordance between our trophic-based approach and BMRE scheme (Fig. SM2.3c). Realms in pelagic areas at higher latitudes (e.g. 1-8 and 30) were partially or fully nested within TCs 1 and 2, and those closer to the equator (e.g., 9, 10, 17, 19, 21, and 22) were partially nested within TCs 2 and 3 (Fig. SM2.4c). Similar to MEOW and PPOW, coastal and shelf areas were nested within TCs 4 to 6. It is worth noting that the classification scheme based on BMRE data uses a coarse resolution (5°), which makes the correspondence between areas and coastal and shelf TCs less sharp compared with MEOW and PPOW (see the greater number of overlapping lines in Fig. SM2.4c).

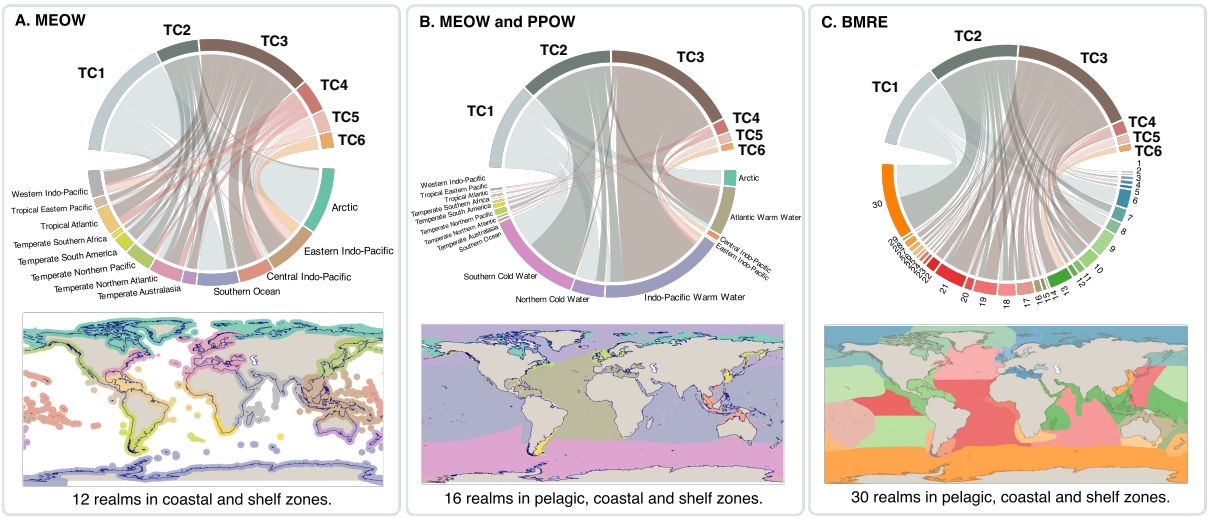

Figure SM2.4. Correspondence between the observed distribution of marine trophic communities and previous bioregionalization proposals. Thickness of the flow lines indicates the number of grid cells that correspond between our scheme of trophic communities ('TC') (up) and previous biogeographic regionalization schemes (down): Marine Ecoregions of the World (MEOW), Pelagic Provinces of the World (PPOW), and Marine Biogeographic Realms based on species endemicity (BMRE).

References

1. A. J. Gates, I. B. Wood, W. P. Hetrick, Y.-Y. Ahn, Element-centric clustering comparison unifies overlaps and hierarchy. *Sci Rep* **9**, 8574 (2019).

2. L. Bentley, *et al.*, Marine megavertebrate migrations connect the global oceans. [Preprint] (2024). Available at: https://www.researchsquare.com/article/rs-4457815/v1 [Accessed 2 October 2024].

3. M. D. Spalding, *et al.*, Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas. *BioScience* **57**, 573–583 (2007).

4. M. J. Costello, *et al.*, Marine biogeographic realms and species endemicity. *Nat Commun* **8**, 1057 (2017).

5. J. C. Briggs, B. W. Bowen, A realignment of marine biogeographic provinces with particular reference to fish distributions: Marine biogeographic provinces. *Journal of Biogeography* **39**, 12–30 (2012).