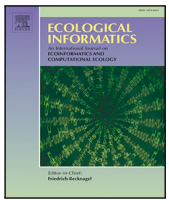
Ecological Informatics 95 (2026) 103704

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Machine learning-based identification of key indicator species in coral reef fish assemblages

Eberrhard Estephe Kana Djifack [a,b,c,e], Edi Prifti [a,d], Thomas Lamy [e], Florian Baletaud [c], Jean-Daniel Zucker [a,d], Norbert Tsopze [a,b], Laurent Vigliola [c], Eugeni Belda [a,d,*]

[a] IRD, Sorbonne University, UMMISCO, 4, place Jussieu, Paris, France
[b] Department of Computer Sciences, University of Yaoundé I, P.O. Box 812, Yaoundé, Cameroon
[c] ENTROPIE (IRD, UR, UNC, CNRS, IFREMER), Institut de Recherche pour le Développement (IRD), Centre IRD de Nouméa, 98000 Nouméa, New Caledonia, France
[d] INSERM, Nutrition et Obesities; systemic approaches, NutriOmique, AP-HP, Hôpital Pitié-Salpêtrière, 91 boulevard de l'Hôpital, 75013 Paris, France
[e] MARBEC, University of Montpellier, IRD, IFREMER, CNRS, Montpellier, France

ARTICLE INFO

Dataset link: here, here

Keywords:
Indicator species
Machine learning
Predomics
IndVal
TWINSPAN
Baited video
Marine tropical lagoon

ABSTRACT

Coral reef monitoring often relies on indicator species to reflect ecological conditions across habitats, yet existing identification methods such as IndVal and TWINSPAN vary widely in the number and identity of the taxa they select. Here, we compare these two traditional approaches with an interpretable machine learning method, Predomics, to identify parsimonious sets of indicator species using Baited Remote Underwater Video Stations (BRUVS) in a tropical lagoon of New Caledonia. TWINSPAN and Predomics consistently identified far fewer indicator species than IndVal, yet all methods achieved equivalent predictive accuracy in habitat classification. Notably, all approaches converged on the same core taxa for distinguishing inshore from offshore habitats, lending strong cross-method support to the ecological relevance of these species. Results were robust across both abundance and presence/absence data, and cross-validation confirmed the generalizability of the selected indicators to unseen samples. The identified indicator species, primarily wrasses, goatfishes, and threadfin breams, align with established habitat preferences. These results show that parsimonious, interpretable machine learning methods can match or complement classical approaches while delivering simpler, more actionable indicator sets for efficient and scalable reef health assessments and conservation planning.

1. Introduction

As human activities and global changes continue to intensify pressure on marine ecosystems (Halpern et al., 2019), improving the management and conservation of underwater wildlife has never been more crucial (Halliday et al., 2022). Regular and consistent monitoring of these ecosystems is essential for tracking changes over time, assessing their health and functionality (Emslie et al., 2020), and guiding effective conservation strategies (Danovaro et al., 2020). Traditional monitoring methods of marine biodiversity rely on direct visual observation techniques such as underwater visual censuses (UVC) operated by scuba divers (Caldwell et al., 2024) or baited remote underwater video stations (BRUVS) (Simpfendorfer et al., 2023). Regardless of the technique, marine biodiversity monitoring data include a list of species with associated abundance and size.

When assessing entire communities, a wide range of species can be identified and monitored, offering valuable insights into biodiversity patterns. However, for conservation purposes, it can be more useful and cost-effective to focus on a few key species whose abundances vary significantly between specific habitats, such as inside or outside marine protected areas (MPAs) or between impacted and control sites. Indicator species are deemed indicative of such habitats and sensitive to ecological shifts, making this concept popular in ecology for identifying species that are important to monitor environmental change (Siddig et al., 2016). Indicator species are usually chosen in the impact assessment depending on the target change, such as heavy metals in invertebrates for water pollution (Han and Han, 2024). On a broader ecological assessment spectrum, indicator species represent proxies of the environmental conditions (i.e. the habitat) by observing their general behaviour like presence/absence, abundance, size or other biological traits (Siddig et al., 2016). Fish can be a good general indicator species of habitat as they are visually identifiable and more easily observed or detected with adequate methods like cameras

* Corresponding author at: IRD, Sorbonne University, UMMISCO, 4, place Jussieu, Paris, France.
E-mail address: eugeni.belda@ird.fr (E. Belda).

https://doi.org/10.1016/j.ecoinf.2026.103704
Received 26 May 2025; Received in revised form 9 March 2026; Accepted 9 March 2026
Available online 10 March 2026

(Langlois et al., 2020; Wee et al., 2023). For example, the juvenile species of *Lutjanus argentimaculatus* is associated with mangrove nurseries habitat, which is a sensitive habitat (Russell and McDougall, 2005).

In ecology, one of the methods that allows to identify indicator species is the Indicator Value (IndVal) method (Dufrene and Legendre, 1997). IndVal is a statistical approach developed to assess the indicative value of a given species based on its association with a specific set of environmental characteristics. It combines species abundance and frequency data to calculate an association index between each species and habitat or environmental characteristics. Another classical method is Two-Way Indicator SPecies ANalysis (TWINSPAN) (Roleček et al., 2009), which is a hierarchical classification method that iteratively divides a dataset into two groups based on species composition to identify indicator species characteristic of each group. However, although very useful, these techniques identify important species based on their association with specific habitats, but they do not evaluate the generalizability of these associations to other datasets. In this context, Machine learning (ML) can be particularly useful to uncover patterns and relationships within complex datasets, while allowing to evaluate their generalizability in other datasets. ML is increasingly applied in ecology to extract meaningful ecological insights from complex datasets (Pichler and Hartig, 2023). For instance, Kumar et al. (2023) used ML techniques to monitor land-use and land-cover changes in the Roorkee region of India, providing robust tools for landscape-level management. Similarly, Morales and Villalobos (2023) demonstrated how ML can be applied in agriculture to improve crop yield predictions and soil management strategies. Other studies have shown how ML-based models can support monitoring of environmental changes and biodiversity across scales (e.g., Panigrahi et al. (2023)). Together, these studies illustrate the growing utility of ML in ecological and environmental sciences and emphasize the importance of adapting interpretable ML approaches to biodiversity monitoring in marine ecosystems. In particular, ML can identify indicator species by analysing feature rankings derived from predictive models using metrics such as Gini importance, SHAP values, or mean decrease accuracy, which quantify the contribution of individual features to model predictions (Nembrini et al., 2018; Lundberg et al., 2020).

By appropriately splitting census datasets into training and testing subsets, ML methods enable for robust evaluation of model performance on unseen data. Specifically, the cross-validation setting is a common way of evaluating the level of over-fitting of the modelling approach. Advanced ML algorithms, such as random forests (RF) and support vector machines (SVM), are often referred to as "black boxes" due to their complexity, making them challenging to interpret in ecological contexts. Interpretability is a crucial requirement for practical applications like ecological monitoring, forecasting, and even clinical practice (Camacho et al., 2018). In this context, the Predomics approach (Prifti et al., 2020) was proposed to create simplified ML prediction models grounded in ecological relationships between microbial species. Predomics implements algorithms capable of identifying indicator species in a community by evaluating their feature importance on binary classification tasks, which refers to the indicative value of each species. This approach is shown to achieve predictive performance comparable to complex state-of-the-art (SOTA) ML methods while utilizing a significantly smaller set of species, which makes it less prone to over-fitting. Besides practical benefits, a small number of indicator species also offers economic benefits while allowing for larger-scale applications.

In this study, we evaluate and compare the ability of IndVal, TWINSPAN and Predomics methods to identify indicator species in the field of Ecology. We rely on a marine case study consisting of a diverse set of tropical fish inhabiting three distinct coastal habitats. This study uses data from Baited Remote Underwater Video Stations (BRUVS) collected by our team between April and May 2016 in Nouméa, New Caledonia (Baletaud et al., 2022), a marine biodiversity hotspot in the South Pacific. The specific objectives of this research are (1) to use the Predomics machine learning algorithm to identify the most important fish species on binary classification tasks across the three habitats and also between zones (group of habitats); (2) to compare indicator species discovered by Predomics to those discovered using the traditional methods (IndVal and TWINSPAN); (3) and to evaluate the generalization and usefulness of key species identified by these methods, using sample subsets that were excluded during the indicator species selection process. By comparing all three approaches, our ultimate goal is to provide a baseline for ML algorithms to be used to identify indicator species in the field of ecology and conservation.

2. Materials and methods

2.1. Baited Remote Underwater Video Stations (BRUVS) dataset

We used the dataset collected by Baletaud et al. (2022), which involves the deployment of 60 Baited Remote Underwater Video Stations (BRUVS) in the Lagoon of Nouméa, New Caledonia, to evaluate differences in fish communities across different marine habitats (Fig. 1-A). The original study was designed to assess how fish species composition changes along an environmental gradient, from Inshore Bays to Offshore Barrier reefs. More specifically, the surveys covered three main marine habitats, namely a Bay, a Lagoon, and a Barrier reef, which are characterized by distinct sedimentary belts and benthic communities from coast to Offshore, respectively. Each habitat was surveyed at two replicate sites (transects, respectively Aboré and Mbéré). At each site, 10 BRUVS with GoPro Hero 2 cameras were deployed for one hour across three sites, following Juhel et al. (2018). In total, the 60 videos captured 148 distinct fish species. The abundance of each fish species was estimated manually using the MaxN metric, which provides the maximum number of individuals for each species observed simultaneously in a given video (Langlois et al., 2020). Fig. 1-A shows the distribution map of all 60 videos samples in the tropical Lagoon of New Caledonia. Fig. 1-B illustrates the distribution of species abundance (grouped by their prevalence across samples) and Fig. 1-C the species richness across each study site (number of observed fish species). This distribution shows that 77% of species in the community are rare or present in a limited number of samples, whereas 23% show at least 10% prevalence across different sites. Fig. 1-D, represents the overall difference between transects (replicates) and sites based on fish community composition, which shows a clear grouping of coastal habitats (*Bay* and *Lagoon*) that clearly separates from more distant *Barrier* habitats, justifying the pooling of *Bay* and *Lagoon* habitats in a *inshore* category in the original study, that was compared with the *offshore*, *Barrier* habitat, using the IndVal method to identify key species.

In Fig. 2, we summarized the analytical approach followed in the study. We utilized two types of input data derived from this dataset—fish MaxN abundance and presence/absence. We also analysed the differences in the indicator species obtained using the entire community (as in the original study) and using a subset of fish species presents in at least 10% of the sites in order to evaluate whether rare species may constitute a bias for the methods rather than refine results. These data abundance tables were analysed using IndVal, TWINSPAN and Predomics methods to identify indicator species. Fig. 2-A illustrates the approach focused on feature discovery with both the filtered and the complete datasets, while Fig. 2-B illustrates the generalization model evaluation procedure (the data is randomly split into training and testing subsets). Fig. 2-C shows a similar analysis as B, but instead of randomly splitting the data into train and test, training and testing datasets are defined by transects (train on Aboré and test on Mbéré, train on Mbéré and test on Aboré). We conducted these analyses for four binary classification tasks: *inshore (Bay and Lagoon)* vs. *offshore (Barrier)*, and the three pairwise comparisons among sites (*Bay*, *Lagoon*, and *Barrier*). In the next sections, we provide further details on the experimental approach.

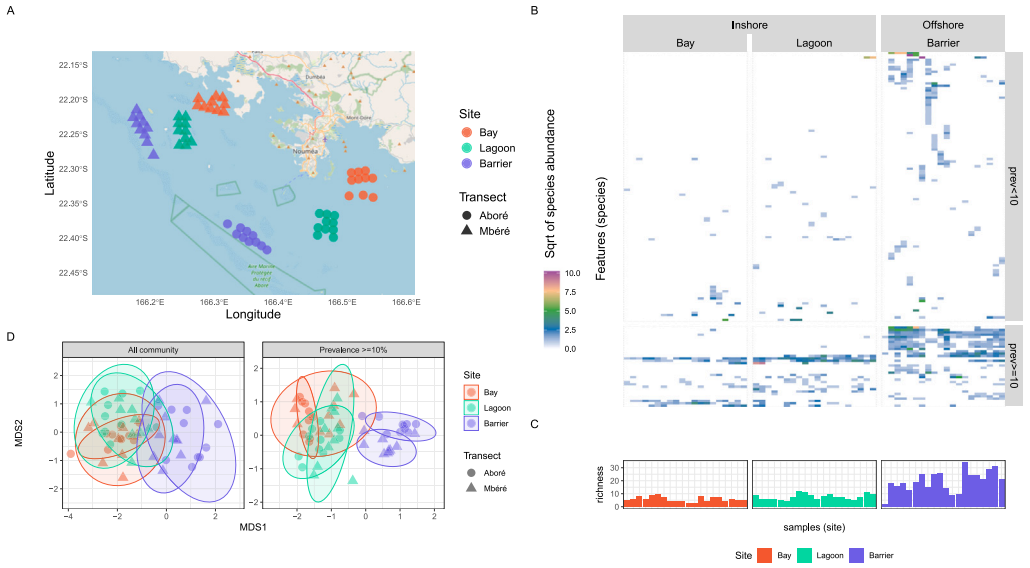


Fig. 1. **A.** A sampling map of 60 BRUVS data from the Inshore (Bay and Lagoon) and Offshore (Barrier) separated in 10 samples (videos) for each transect per site and collected by Baletaud et al. (2022) between April and May 2016 in New Caledonia. **B.** Heatmap of fish species' (y-axis) abundance (MaxN) grouped by `prev_rate` < 10% (with 114 species) and `prev_rate` ≥ 10% (with 34 species), based on the square root of species abundance on each sample across different sites (x-axis). **C.** Barplot of species' richness across samples in the Bay, Lagoon, and Barrier zones. **D.** Non-metric multidimensional distance scaling (NMDS) of the 60 videos samples between 3 sites: Bays, Lagoons and Barriers replicated in both transects. Ellipses group samples from the same transect of each site.

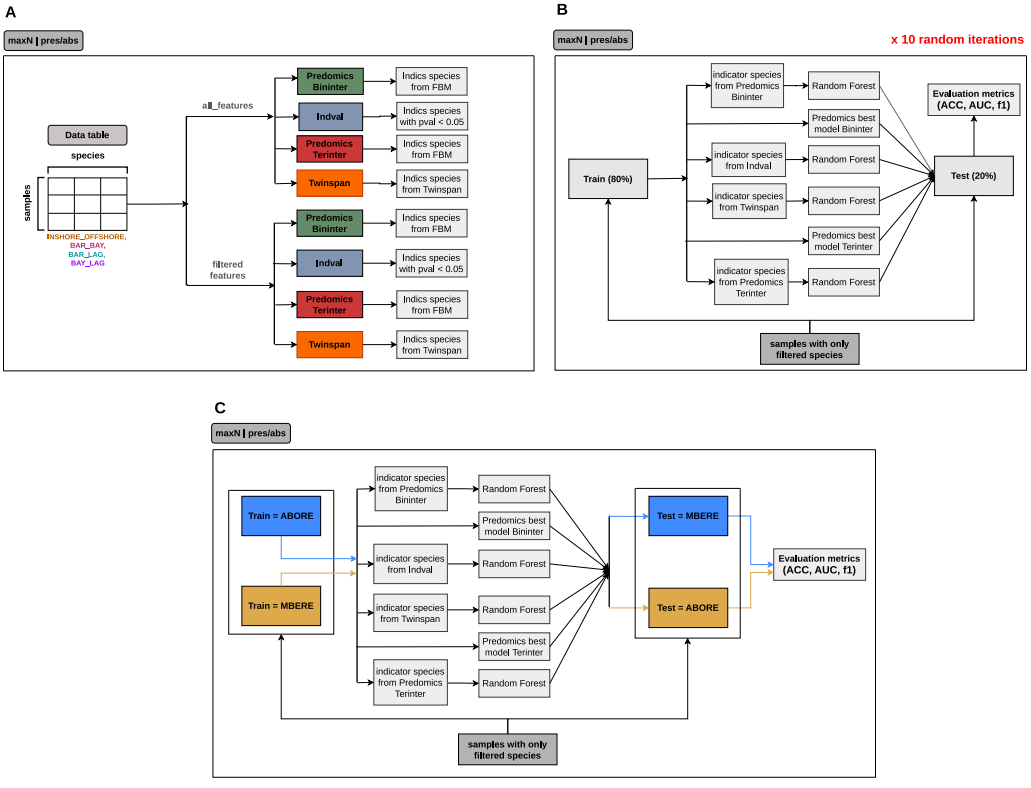


Fig. 2. Experimental approach pipeline. **A.** Indicator species identification using IndVal, TWINSPAN, Predomics terinter and bininter on abundance and presence/absence data tables using all the features (fish species) and those present at a minimum 10% prevalence. **B.** Diagram of generalization evaluation between Predomics' terinter and bininter best models and Random Forest models trained on indicator species identified by IndVal, TWINSPAN and Family of Best Models (FBM) of Predomics terinter and bininter. **C.** Diagram of generalization analysis with data split by transect (replicates) between FBM of Predomics' terinter and bininter models, Random Forest models (trained on indicator species identified by IndVal, TWINSPAN and FBM of Predomics terinter and bininter models).

2.2. Indicator species analysis

The IndVal method introduced by Dufrene and Legendre (1997) is a statistical method commonly used to identify indicator species that can be used in ecology and environmental monitoring as indicators of an ecosystem's quality, health, or changes. It calculates an indicator value based on species specificity (fidelity to a site) and relative abundance within that site. The significance of the indicator value is tested using permutations to ensure the association between species and site groups is not random. The indicator value is scaled from 0 to 100, with a high score indicating the species is both highly abundant and uniquely associated with a particular habitat. The method was originally tested on forest ecosystem data in Belgium (1997) and has been used across diverse ecosystems worldwide (Ricotta et al., 2015). Here, we used the IndVal method to identify key indicator species of either Inshore or Offshore zones of the tropical Lagoon of New Caledonia, as well as pairwise comparisons between Bay, Lagoon and Barrier sites. For each comparison, we run the IndVal method on complete species and species filtered at 10% of prevalence in abundance and presence/absence. The analysis was performed using the indval implementation from the R package labdsv version '2.1.0'.

2.3. TWINSPAN analysis

TWINSPAN (Two-Way INdicator SPecies ANalysis), introduced by Hill (1979), is a hierarchical classification technique widely used in ecology to analyse species composition and community structure. It proceeds by iteratively dividing a dataset into two groups based on species composition, producing a binary dendrogram that captures the relationships between samples and species. At each division step, the method identifies pseudospecies — transformed abundance classes — that best discriminate between the two resulting groups, which then serve as indicator species characteristic of each cluster. Originally developed for vegetation classification (Shimwell, 1971), TWINSPAN has since been applied across a broad range of ecological communities, including aquatic ecosystems. Here, we used the modified TWINSPAN algorithm proposed by Roleček et al. (2009), which improves upon the classical approach by incorporating an analysis of within-cluster heterogeneity prior to each division, thereby producing more ecologically coherent groupings. We applied this method to the BRUVS dataset to identify key fish species associated with each habitat (Bay, Lagoon, and Barrier), and compared the resulting indicator species with those obtained from the IndVal and Predomics approaches. The analyses were performed using the `twinspan` function from the `twinspanR` package version '0.22', which is currently restricted to Windows OS.

2.4. Predomics analysis

Predomics is an innovative machine learning approach developed by Prifti et al. (2020) and was initially tailored for microbiome datasets. Predomics builds predictive models for binary classification and regression tasks and is inspired by the interactions observed within microbial ecosystems. It aims to generate accurate predictive signatures while offering high interpretability. For this purpose, predictive accuracy and model sparsity are balanced, penalizing the addition of variables/features to the models if the improvement in model performance is not significant. Here, we used Predomics to perform four binary classification tasks on the BRUVS dataset. First, to identify key species of either *inshore* vs. *offshore* habitats. Second, on pairwise comparisons of individual sites (*Barrier–Bay*, *Barrier–Lagoon*, and *Bay–Lagoon*). Models were built with both the abundance (MaxN) and presence/absence data, for each classification task, using the entire set of fish species or with the fish species that were found in at least 10% of the video samples.

For each of the four classification tasks, two models were fitted based on Predomics Binary (bininter) and Ternary (terinter) languages (Prifti et al., 2020). The bininter models select a group of features (fish species) where the unweighted cumulative abundance of which allows to classify samples in a given class (site or zone), and the terinter models search for 2 distinct groups of species where the difference in their cumulative abundances allows to classify samples in a given class. During the training phase, Predomics was run with multiple instances of terbeam learners on Five-fold cross-validation setting. From the results of each model, we define indicator species as the species included in the Family of Best Models (FBM), which are defined as models whose accuracy is within a given window (non-significant difference) of the best model's accuracy. This window is defined by computing a significance threshold, assuming that accuracy follows a binomial distribution ($p<0.05$). No hyperparameter optimization was performed. Species included in the FBM were further explored in terms of prevalence across models and feature importance, described as the mean decrease accuracy (MDA) of the model after feature removal. The best Predomics model in each classification task were also compared with Random Forest models built on the same sets of indicator species.

In simpler terms, Predomics is a Machine Learning framework that can help ecologists identify key species that can reliably classify ecological samples, like different habitats or sites, using presence/absence or abundance data. It builds simple, interpretable models that balance accuracy and simplicity by including only species that truly improve predictions. The bininter model finds a single group of species whose combined presence or abundance indicates a particular class (e.g., Lagoon vs. Barrier), while the terinter model identifies two groups and uses the difference in their abundances to classify samples. Applied to BRUVS data, Predomics can extract small, meaningful sets of species that serve as ecological indicators, providing a powerful, rule-based approach to analyse complex communities without requiring advanced machine learning knowledge. The analyses were performed in R with the Predomics package version '1.1.0'.

2.5. Compositional analysis of fish species communities

We used the PERMANOVA test (from the adonis2 function of the vegan R package version '2.6.4'.) to evaluate statistically the impact of either the pairwise comparison (*Bay–Lagoon*, *Bay-Barrier* and *Barrier–Lagoon*) or the *inshore–offshore* classification on fish community composition defined from different sets of indicator species. The tests were performed on the distance metrics (using Bray-Curtis distances for abundance data and the Jaccard distance for presence/absence data), of the whole dataset (all fish species) as well as on the different subsets of key species identified by the IndVal, TWINSPAN and Predomics (FBM) models on the different classification tasks.

2.6. Random forest analysis

To evaluate the classification potential of the identified key species by Predomics, TWINSPAN and IndVal on different samples (i.e. not seen during the key species selection step), we conducted two types of analysis. First, the original dataset was randomly split into training (80% samples; train dataset) and testing (20% samples; test dataset) subsets. Random Forest (RF) models were trained on the indicator species identified in the train dataset (derived from IndVal, TWINSPAN, Predomics bininter and Predomics terinter), and evaluated on the corresponding test sets. Additionally, each classification task was evaluated using the best Predomics model identified on the training data. This process was repeated 10 times (different seeds) on the initial dataset. The results were compared in terms of accuracy, AUC and f1 score metrics (Fig. 2-B). Second, a similar analysis was conducted considering the original dataset design into two transects (replicates), using one transect for training and the other for testing, and vice versa (Fig. 2-C). The randomForest R package (Cutler et al., 2007) version '4.7-1.1' was used for model training and evaluation.

2.7. Statistical analyses

To assess the ecological coherence between the traditional methods (IndVal and TWINSPAN) and Predomics models (based on feature importance), Spearman correlations (Zar, 2005) were employed to compare the importance of indicator species throughout the classification tasks. Wilcoxon rank-sum tests (Ford, 2023) were used to compare the overall fish prevalence in the samples between those identified as indicator species and those that were non-indicator species. This was analysed for both abundance and presence/absence. We used pairwise Dunn's test to evaluate the generalization predicting performances of Predomics Best Models and the respective models built with Random Forest upon the different Indicator Species subsets. The dunn.test function from the dunn.test R package (version 1.3.6) was used.

3. Results

3.1. Indicator species sets derived from IndVal, TWINSPAN and predomics approaches

In the original paper of Baletaud et al. (2022) 44 Indicator Species of *inshore vs. offshore* sites were identified with the IndVal method using the MaxN abundance on the full list of observed fish in all video samples (60 samples of 148 species). In comparison, the FBM derived from bininter and terinter models on the same dataset included 35 and 7 species respectively (Supplementary Figure S1-A), with few overlaps in shared species (2 common species to the four methods, 4 species shared between IndVal and bininter FBM; Supplementary Figure S1-B). TWINSPAN identified the fewest indicator species with 5 species on MaxN, all of which were also identified by the IndVal method (Supplementary Figure S1-A, Supplementary Figure S1-B). Similar analysis on binary presence/absence data identified 45 Indicator Species with the IndVal method, 44 of which were shared with the results on MaxN (Supplementary Figure S1-A, Supplementary Figure S1-E). In contrast, only 2 and 6 species were included in the FBM of bininter and terinter Predomics models fitted on the same presence/absence data, all of which were shared with Indicator Species of the IndVal method (Supplementary Figure S1-A, Supplementary Figure S1-C), and also found in the FBM results derived from MaxN data (Supplementary Figure S1-D, Supplementary Figure S1-F). TWINSPAN similarly identified 7 indicator species on presence/absence data, all shared with the IndVal method (Supplementary Figure S1-A), where 5 of them were found by TWINSPAN in maxN (Supplementary Figure S1-G).

An important aspect of machine learning is how well the retrieved models perform on unseen data during the training/testing process, and in this context the feature prevalence is a variable that is commonly used to pre-filter the data to remove low-prevalence features that could decrease the performance of models on unseen data (Walsh et al., 2024; Asnicar et al., 2024). This was also observed in this dataset when we analysed the distributions of AUC (Area Under the Curve) values across the 5 folds of the cross-validation process of Predomics between the empirical data (samples used for model training) and the generalization data (samples held out during each fold). We observed that the AUC is significantly higher on both MaxN and presence/absence when the Predomics models were built using only the filtered subset of 34 species (23% of all observed fish species) present in at least 10% of the samples on both empirical and generalization data (pvalue < 0.05, Wilcoxon Rank-Sum test, Supplementary Figure S1-H).

With the 34 fish species present in at least 10% samples, the IndVal approach was still the one identifying the highest number of Indicator Species (25 and 26 species, respectively on MaxN and presence/absence data). Similarly, Predomics terinter identified 24 and 20 species, respectively on MaxN and presence/absence data (Fig. 3-A). In contrast, TWINSPAN identified the fewest indicator species, with 5 and 7 species on MaxN and presence/absence data respectively (Fig. 3-A). Most importantly, in comparison to the results with the full set of indicator species, a large overlap was observed in Indicator Species identified by the different approaches on both MaxN (Fig. 3-B) and presence/absence data (Fig. 3-C), even if we still observed 7 indicator species based on MaxN and 3 species based on presence/absence that were not identified by the IndVal method.

We also observed that the sets of Indicator Species detected on presence/absence data were highly overlapping with those from MaxN abundance data across the different approaches (Fig. 3-D, E, F, G). Furthermore, all the different sets of indicator species defined by the different methods and datasets seem to capture the most discriminant set of species to distinguish between Inshore and Offshore communities. This was supported by PERMANOVA tests, which showed that the variance explained by these indicator species sets was comparable to that obtained using the full species community (Fig. 3-H). Despite the initial filtering at 10% prevalence threshold, we still observed that fish species retained as Indicator Species by the different approaches showed higher prevalence on average than non-indicator species of the community (Fig. 3-I). When the feature importance metrics of the different methods (Indicator Value from IndVal, chi-square value from TWINSPAN, and Mean Decrease Accuracy from Predomics models) were compared pairwise, we observed a strong and significant positive correlation between these metrics when considering the common set of Indicator Species shared between IndVal and Predomics approaches (p-value < 0.05 on pairwise Spearman correlation in all common species sets except for common species between IndVal and Predomics bininter FBM on presence/absence data, Fig. 3-J); which contrasts with TWINSPAN vs. Predomics where none of the correlations between the metrics for the common set were significant (Fig. 3-J). Most importantly, we found that this common set of indicator species showed significantly higher importance values for the classification of Inshore and Offshore samples (Fig. 3-K–N).

When we reproduced the same analyses on all pairwise combinations of individual sites (Bay, Lagoon, Barrier), we observed a similar pattern as the results on Inshore vs. Offshore data (Supplementary Figure S2). First, IndVal still rendered a higher number of Indicator Species than other methods in comparisons of sites close to the coast (Bay, Lagoon) vs. the coral reef (Barrier; Supplementary Figure S2-A). The exception seems to be when Indicator Species are identified between the two sites closer to the coast (Bay vs. Lagoon), where all four methods yielded less than 10 Indicator Species on both MaxN and presence/absence data (Supplementary Figure S2-A). Second, we observed that the different subsets of Indicator Species defined significantly different fish community compositions across compared sites with similar variance explained as that of the entire fish community, except for the analyses of Bay vs. Lagoon on maxN abundances (pvalue <0.05; PERMANOVA test; Supplementary Figure S2-B), with a prevalence that tends to be on average higher than other fish species of the community, particularly for Indicator Species derived from the IndVal method (pvalue <0.05, Wilcoxon Rank-Sum tests; Supplementary Figure S2-D). And third, we consistently observed a strong and significant positive correlation between the feature importance metrics (indicator value for IndVal method, chi-square value for TWINSPAN, and MDA value for Predomics methods) in common species shared between IndVal and Predomics approaches with the exception of Bay vs. Lagoon comparison (Supplementary Figure S2-C), which also contrasts for TWINSPAN vs. Predomics. The common sets of indicator species shared between IndVal and Predomics, and between TWINSPAN and Predomics approaches were largely the most important ones in terms of binary classification of samples across individual sites (Supplementary Figure S2-E, F, G, H).

In summary, TWINSPAN and Predomics models consistently identified a reduced number of potential indicator species across different datasets than the IndVal approach, with significant overlap between approaches (IndVal, TWINSPAN and Predomics). We also observed that subcommunities defined by these indicator species explained a similar fraction of compositional variation across pairwise habitats as the entire fish community, except for the Bay vs. Lagoon comparison, where a low number of Indicator Species were retrieved.

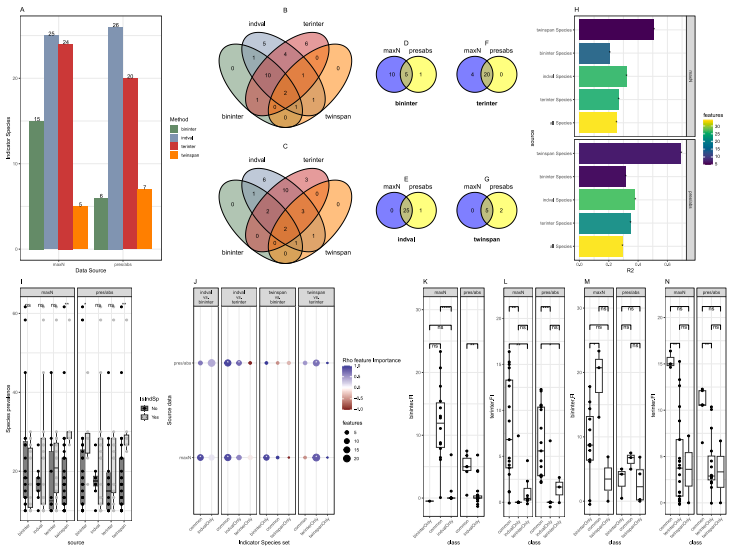


Fig. 3. Indicator species selected from species filtered at 10% prevalence threshold by IndVal, TWINSPAN, Predomics bininter and terinter models on *inshore–offshore* comparison. **A.** Number of indicator species found per method with abundance (MaxN) and presence/absence (pres/abs) data. **B–C.** Venn diagrams representing overlaps in indicator species shared by the four approaches in abundance and presence/absence data respectively. **D–G.** Venn diagrams representing overlaps in indicator species shared between MaxN and presence/absence data, respectively from Predomics bininter, IndVal, Predomics terinter and TWINSPAN methods. **H.** Variance explained in fish community composition between *inshore–offshore* communities derived from indicator species from the four approaches and based on the whole fish communities (all species; n = 34 species) based on PERMANOVA analyses on beta diversity matrices (Bray-Curtis for MaxN; Jaccard distances for presence/absence data). Barplots are coloured by the number of fish species in each community. * = pvalue < 0.05 on PERMANOVA tests. **I.** Prevalence distribution of fish species retained as indicator species (IsIdnSp = Yes) vs. non-indicator species of the community (IsIndSp = No) across the four approaches (x-axis) on MaxN and presence/absence data. * = pvalue < 0.05, ns = pvalue>0.05, Wilcoxon Rank-Sum test. **J.** Dotplots representing Spearman rho values from pairwise correlation of feature importance metrics between traditional methods (Indicator Value from IndVal, chi-square value from TWINSPAN) and Predomics methods (Mean Decrease Accuracy) for common and method-specific Indicator Species (x-axis) on MaxN and presence/absence data (y-axis). Dots are coloured by Spearman Rho with size proportional to the number of fish species in each comparison. * = pvalue < 0.05 on Spearman correlation. **K-L.** Boxplots of Predomics feature importance values (Mean Decrease Accuracy) on Indicator Species shared with IndVal method (common) vs. specific of Predomics bininter (K) and terinter (L) models on maxN and presence/absence data. **M-N.** Same as (K-L), but between TWINSPAN and Predomics models. ns: $p > 0.05$, *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001, Wilcoxon Rank-Sum tests.

3.2. Overview of the most relevant fish species on pairwise classification of sampling sites

A detailed analysis of the Indicator Species retained by the Predomics approaches shows that bininter models, which search for groups of species whose cumulative sum classifies a sample as Inshore or Offshore, predominantly retain species strongly associated with the Offshore site. Notably, species like *Halichoeres trimaculatus*, *Parupeneus barberinoides* and *Lethrinus variegatus*, stand out, appearing in 70% of models in FBM and showing the highest MDA values in terms of feature importance. These species are present in 76.66% of Offshore samples compared to only 0.83% of Inshore samples, and they are retained using both MaxN abundance and presence/absence data (Fig. 4-A, B). In comparison, with Predomics terinter models, which search for groups of species where the difference in their cumulative abundance classifies a sample as Inshore or Offshore, highlight fish species strongly associated with Inshore sites. Notably, species like *Nemipterus peronii*, *Lethrinus genivittatus* and *Atule mate* emerge as significant indicators (Fig. 4-A, B).

These species also emerge as Indicator Species in the pairwise analyses between individual habitats (Supplementary Figure S3). For Barrier–Bay and Barrier–Lagoon comparisons, the species *Nemipterus peronii* exhibits the highest feature importance and prevalence scores (Supplementary Figure S3,**A–D**, Feature importance, prevalence) associated with Bay and Lagoon, attesting to its strongly association with Inshore zone (Fig. 4). Similarly, the species *Lethrinus rubrioperculatus*, *Lethrinus variegatus* and *Halichoeres trimaculatus* are strongly associated with Barrier site (Offshore). Most importantly, Predomics methods also allow us to identify key Indicator Species in the comparison of the two coastal habitats, like the Bay–Lagoon comparison, where we observed that *Leiognathus fasciatus* is strongly associated with the Bay site, while *Pentapodus nagasakiensis*, *Pseudalutarius nasicornis*, *Arothron stellatus*, and *Upeneus moluccensis* are linked to the Lagoon site (Supplementary Figure S3,**E–F**).

3.3. Evaluation of generalization performances of models

Next, we explored the predictive power of indicator species defined by the different approaches (IndVal, TWINSPAN, Predomics bininter and terinter methods) in binary habitat classification tasks on samples excluded from the indicator species definition step. For this purpose, we repeated the identification of Indicator Species 10 times with random splits of the data, selecting 80% of the samples to define the Indicator Species (training samples) and 20% of the samples (holdout samples) to evaluate their predictive power on habitat classification with a Random Forest approach. We also included the best Predomics models on each classification task in the comparison (Fig. 2-B). On the Random Forest framework, models fitted with the different sets of indicator species showed mean AUC values across the 10 experiments ranging from 0.87 to 1, indicating a good assignment of samples to the corresponding habitat not only on Inshore vs. Offshore classification but also on all pairwise habitat combinations and data sources (maxN and presence/absence data; Fig. 5-A). We only observed a slight decrease in performance on models trained with IndVal and terinter indicator species on Bay vs. Lagoon classification, with a more pronounced decrease for models trained with TWINSPAN indicator species, which causes a statistically significant difference with other methods (pvalue <0.05; Pairwise Dunn tests Supplementary Figure S5). The performance of the best Predomics models on each classification task was slightly lower in terms of mean AUC, ranging from 0.81 to 0.99, with lower

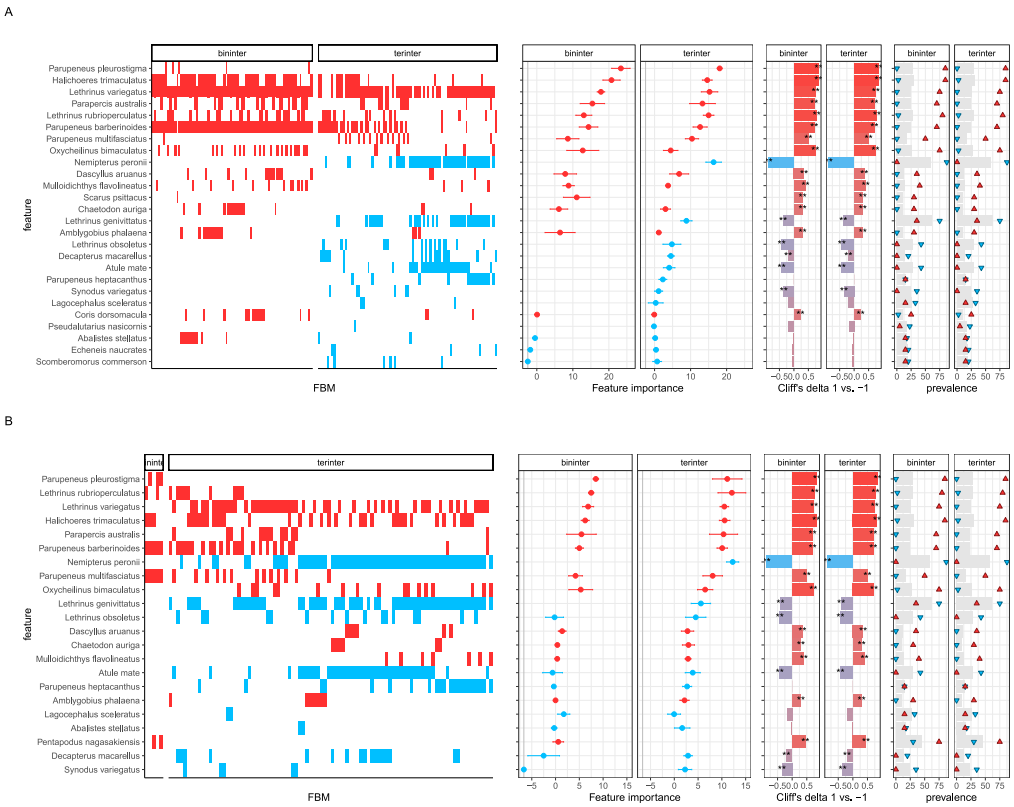


Fig. 4. Summary of retained fish species with Predomics approach on Inshore–Offshore classification task. **A.** and **B.** show indicator species identified by the FBM of Predomics models (72 and 5 models out of 434 for bininter; 80 and 90 out of 400 for terinter, in terms of abundance and presence/absence, respectively). *First plot:* Presence of fish species (y-axis) across the FBM from Predomics bininter and terinter models (indicator species). *Second plot.* Feature importance of key species identified by the models in Inshore (blue) and Offshore (red). *Third plot.* Cliff's delta effect sizes of changes in MaxN abundances of key species between Inshore and Offshore zones. *Fourth plot.* Prevalence of species in *inshore* (blue) and *offshore* (red).

values obtained with the terinter models and particularly on the Bay vs. Lagoon classification (Fig. 5-A). However, these Predomics models were significantly smaller in terms of included fish species, ranging on average between 3 and 5 fish species vs. 3.5 and 24 species for random forest models (Fig. 5-B). When we visualize the maxN abundance profiles of the Indicator Species across train and test samples of the different experiments, we observed that fish species commonly retrieved by the different approaches captured species strongly associated with a given habitat on train and test samples, which explains the high performance of predictive models observed above. This is illustrated in Fig. 5-C for the set of 7 fish species commonly retained by IndVal, TWINSPAN and the FBM of Predomics bininter and terinter approaches on the Inshore vs. Offshore classification, which corresponds mainly to species present in the Offshore sites (coral reef) and nearly absent in Inshore sites (Fig. 5-C). Similar results are observed in the three additional habitats comparison (Supplementary Figure S4). In summary, these different approaches allow to identify species strongly associated with different habitats, which supports their usage as signatures of fish habitat associations useful for ecological monitoring approaches.

We also performed a similar analysis, defining the indicator species and prediction models in one of the transects and applying the models to the other. This approach aimed to assess the predictive performance of the different sets of indicator species on ecological replicates of the same sampling sites. In addition, we trained Random Forest models using the indicator species identified by the Best Models of Predomics bininter and terinter in order to evaluate the effect of the numerical thresholds defined on Predomics model balances (values above/below which a sample is assigned to a given class) on the accuracy of classification. We observed an overall better performance in terms of AUC when indicator species were defined on the Mberé transect (AUC values ranging from 0.795 to 1) vs. when indicator species were defined on the Aboré transect (AUC values from 0.27 to 1; Fig. 6-A). Notably, when training on Aboré and testing on Mbéré, we observed a major drop in AUC performance of the best Predomics terinter models that was not explained by the threshold used by the models for classification (Random Forest models trained on the same fish species as Predomics terinter models showed equivalent drops in AUC, Fig. 6-A). This drop was particularly pronounced for the Barrier–Bay and Barrier–Lagoon comparisons using both abundance and presence/absence data, and for the Inshore–Offshore comparison using abundance data, where AUC values ranged from 0.27 to 0.72 (Fig. 6-A). When focusing on indicator species identification, we observed similar patterns of key species as for analysis done with the different approaches in Fig. 2-A with filtered features. For Inshore vs. Offshore comparison, species like *Nemipterus peronii*, *Lethrinus genivittatus* and *Atule mate* were associated with Inshore while *Halichoeres trimaculatus*, *Parupeneus barberinoides* and *Lethrinus variegatus* were linked to Offshore. Similarly, for Barrier–Bay and Barrier–Lagoon comparisons, we also found *Lethrinus rubrioperculatus*, *Lethrinus variegatus* and *Halichoeres trimaculatus* strongly associated with Barrier site; then *Nemipterus peronii* and *Atule mate* to Bays and Lagoons. Notably, species such as *Dascyllus aruanus* and *Chromis viridis* were strongly associated with the Aboré transect of the Barrier site. Here, we also discriminated Bay from Lagoon with indicator species such as *Leiognathus fasciatus* associated with Bay; and *Pentapodus nagasakiensis*, *Pseudalutarius nasicornis*, *Arothron stellatus* and *Upeneus moluccensis* to Lagoon (Fig. 4, Supplementary Figure S3, Fig. 6, B–E).

4. Discussion

In the present study, we evaluated the performance of the machine learning approach Predomics to identify Indicator Species in three

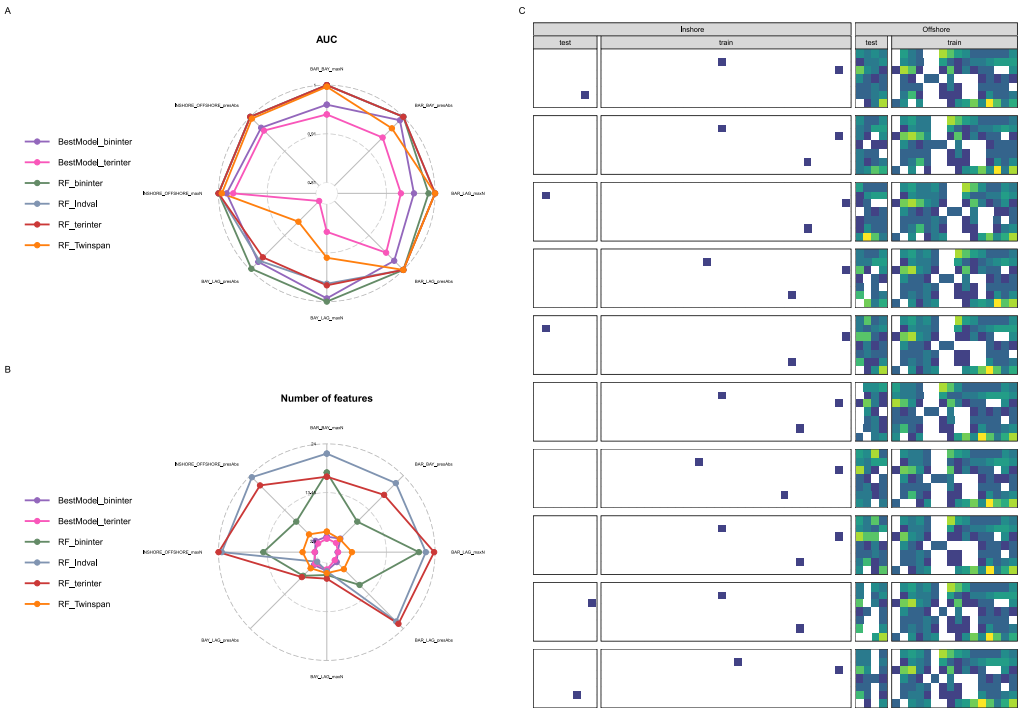


Fig. 5. Performance of Indicator Species on binary sample classification across pairwise habitats. **A.** Radar plot of average AUC values for Random Forest models and the best Predomics models across four classification tasks, each corresponding to a pairwise comparison of habitats: BAR (coral reef habitat), Bay (Bay habitat), LAG (Lagoon habitat); and Inshore (Bay + Lagoon habitats) - Offshore (Barrier habitat). Models were evaluated on MaxN and presence/absence data. For each comparison and data source, Indicator Species were selected using four approaches—the IndVal method, TWINSPAN, and Predomics bininter and terinter models—on 80% of randomly selected samples. These selected species were then used to train Random Forest models, which were evaluated on the remaining 20% of the samples for the same classification task. Additionally, the best Predomics models (binary and ternary) identified during training were also tested on the remaining 20% of the samples. This procedure was repeated 10 times. **B.** Same as **A.** but comparing the number of fish species (features) included in each model. **C.** Heatmap of square root-transformed MaxN abundances of indicator species shared between the IndVal, TWINSPAN, Predomics bininter and terinter approaches (y-axis) across samples (x-axis) over 10 experimental replicates (horizontal facets) for Inshore vs. Offshore comparison. Results for the other 3 pairwise habitat comparisons are included in Supplementary Figure S4.

different habitats in New Caledonia and we compared its results with IndVal and TWINSPAN approaches commonly used in ecology (Baletaud et al., 2022; Roleček et al., 2009). Predomics models were initially developed to work with microbial abundance profiles derived from metagenomics dataset, with the aim of providing more accurate classification and regression models than other machine learning methods, while using a smaller number of variables (Prifti et al., 2020). This approach has been primarily applied in clinical contexts such as the classification of healthy individuals vs. individuals with type II diabetes (Dash et al., 2023), or patients living with HIV (Belda et al., 2024).

By prioritizing sparsity, Predomics models allow the extraction of the most relevant features (e.g., microbial species in a metagenomics dataset; fish species in a BRUVS dataset) for a given classification task. As such, species included in Predomics FBM are conceptually equivalent to the notion of indicator species, defined as species that bear meaningful ecological information, such as the community or habitat types they tend to live in, or the environmental conditions they tend to be associated with De Cáceres et al. (2010). We observed that the number of potential indicator species obtained from the FBM of Predomics is smaller than that obtained by the IndVal method, while remaining comparable to those from TWINSPAN; retaining the same power to discriminate between ecological habitats in the PERMANOVA test as the whole fish community. Most importantly, we observed that these sets of Indicator Species remain the same whether dealing with species abundance (MaxN) or species presence/absence. Accurate quantitative profiling of natural communities from different data sources is a well known issue in ecology, particularly in microbial ecology, where different approaches have been proposed to address problems like the compositional nature of relative abundance profiles (Gloor and Reid, 2016) or the normalization strategies to account for confusion factors before association studies (Lin and Peddada, 2020). Here, even if the context is different, our results show that Indicator Species derived from abundance and presence/absence data are largely overlapping for the four approaches, which reinforces their ecological significance across different data types and analytical methods.

Having common sets of indicator species with both abundance and presence/absence data is particularly valuable, as it enhances the reliability of ecological assessments and allows for more consistent comparisons across studies, ecosystems, and methodologies. It also facilitates integrative analyses, where data from diverse sampling techniques and ecological contexts can be combined to improve our understanding of biodiversity patterns and environmental changes. Such consistency is critical for informing conservation strategies, monitoring ecosystem health, and guiding policy decisions based on ecological indicators that remain stable across different measurement approaches. We cannot discard though that in other ecological contexts each data source could be informative about different aspects of species communities, like large-scale biogeographical patterns that could be more evident with presence/absence data vs. local population dynamics that could be more evident from abundance data.

We also observed that we can further reduce the set of indicator species by looking at the intersection of species sets retrieved by the different approaches, which contain the ones with the highest predictive value. This is consistently observed not only for the initial classification of Inshore vs. Offshore sampling sites but for all other pairwise comparisons of Barrier, Bay, and Lagoon sites. The variability of outcomes across different computational methods for a

Fig. 6. Performance of Indicator Species on binary classification across pairwise habitats, split by replicas. **A.** AUC values for Random Forest models and the best Predomics models (x-axis) across four classification tasks, each corresponding to a pairwise comparison of habitats: BAR (coral reef habitat), BAY (Bay habitat), LAG (Lagoon habitat); Inshore (Bay + Lagoon habitats) and Offshore (Barrier habitat). Models were evaluated on MaxN and presence/absence data. For each comparison and data source, Indicator Species were selected using four approaches—IndVal method, TWINSPAN method, Predomics bininter and terinter models—on one of the two replicas (Aboré, Mbéré). These selected species were then used to train Random Forest models, which were evaluated on the unseen samples of the other replicas for the same classification task. Moreover, the best Predomics models (binary and ternary) identified during training were also tested on the samples of the other replicas. **B–E** Heatmap of square root-transformed MaxN abundances of the full set of key (indicator) fish species included in prediction models shown in panel A for each classification task (y-axis) across samples (x-axis) over two replicates (train Aboré/test on Mbéré, and train on Mbéré/test on Aboré) (vertical facets) respectively for Inshore vs. Offshore, BAR vs. BAY, BAR vs. LAG and BAY vs. LAG comparison.

given task is something common in biology that could be perceived as overwhelming. In microbial ecology, the fact that by combining different abundance profile methods and mathematical approaches for differential abundance analyses (equivalent somehow to the concept of indicator species) it is possible to have at least one significant association for all variables in a community dataset (Yang and Chen, 2022) has led some authors to propose ensemble approaches like omnibus testing to combine the results of different differential abundance methods into a single assessment of the association of a given species to a given condition (Nearing et al., 2022). Here, the intersection of different methods for Indicator Species definition (IndVal, TWINSPAN, Predomics bininter and terinter methods) allows us to define the most relevant species sets in terms of association with respect to different habitats, again a strategy that could be applied if a more reduced set of indicator species is needed for environmental monitoring of natural ecosystems. The role of the identified species as relevant markers of fish habitats is reflected in the high performance of machine learning models learned on Indicator Species defined on random splits of the full dataset (80% of the samples for identifying indicator species and train classification models that are evaluated on the remaining 20% of samples) as well as with models trained on indicator species defined on the Aboré transect and applied to Mberé transect and vice-versa. In this context, a proper split between training and test data is crucial to prevent overfitting, a common pitfall in machine learning where models memorize patterns specific to the training data rather than learning generalizable relationships (Asnicar et al., 2024; Walsh et al., 2024). Overfitting can lead to misleadingly high accuracy during training while failing to make reliable predictions on unseen data. By rigorously maintaining separate training and test sets, we can better assess the true predictive power of the identified species, ensuring that the models capture meaningful ecological associations rather than noise or dataset-specific artefacts. This careful approach strengthens the robustness and applicability of machine learning in ecological monitoring, reinforcing the reliability of indicator species as diagnostic tools for habitat assessment.

A closer look at the most prevalent indicator fish species derived from Predomics methods indicates a strong association with habitats. This is the case for *Halichoeres trimaculatus* (threespot wrasse), *Parupeneus barberinoides* (bicolour goatfish), *Lethrinus rubrioperculatus* (spotcheek emperor), and *Parupeneus multifasciatus* (manybar goatfish) that are often found in Offshore Barrier reefs (Fig. 4). These species are known for living in coral reefs, usually near sandy areas, rocky zones, or places with a mix of sand and coral. They are found in both shallow waters and deeper areas up to 100 m. These fish mainly eat small sea creatures like crabs, snails, worms, and tiny fish. Bicolour goatfish use special feelers (called barbels) to find food in the sand, while wrasses, like *Halichoeres trimaculatus*, follow other fish that stir up the sand to catch escaping prey. Because they need coral reefs and sandy areas for food and shelter, these fish are closely connected to

Offshore Barrier reefs (Froese and Pauly, 2024). In contrast, *Nemipterus peronii* (notchedfin threadfin bream) is a species of demersal fish that typically inhabits marine and brackish waters, living at depths between 17 and 100 m on sandy and muddy substrates. This species is often associated with coastal environments, such as Bays and Lagoons, which provide suitable habitats rich in food resources. Its diet mainly consists of benthic organisms, particularly crustaceans (about 70% of its diet), along with polychaete worms and small fishes. These Inshore habitats offer abundant benthic prey, supporting *Nemipterus peronii*'s nutritional needs, which explains its strong link to these environments (Froese and Pauly, 2024).

We observed that *Nemipterus peronii* was associated with the Inshore zone (Bay and Lagoon) but cannot be used as a discriminative species between Bay and Lagoon sites. In contrast, the species *Leiognathus fasciatus*, *Pentapodus nagasakiensis*, *Pseudalutarius nasicornis*, *Arothron stellatus*, and *Upeneus moluccensis* could be used for this purpose (Supplementary Figure S3-**E,F**). In this context, *Leiognathus fasciatus* (striped ponyfish) is known for inhabiting shallow coastal waters and estuarine environments, where it feeds on benthic invertebrates and zooplankton. Its body structure is adapted for navigating sandy or muddy substrates, and its diet includes small crustaceans and fish larvae, abundant in these habitats. Also, the species *Pentapodus nagasakiensis*, *Pseudalutarius nasicornis*, *Arothron stellatus*, and *Upeneus moluccensis* are ecologically linked to Lagoon habitats, which provide essential food resources, shelter, and breeding grounds. *Pentapodus nagasakiensis* prefers sandy or rubble substrates near reefs, feeding on benthic invertebrates. *Pseudalutarius nasicornis* inhabits seagrass beds and sandy areas, consuming small invertebrates and algae. *Arothron stellatus* thrives in Lagoons with coral and seagrass, feeding on corals, sponges, and crustaceans. *Upeneus moluccensis* forages in sandy or muddy substrates, preying on benthic invertebrates (Froese and Pauly, 2024).

Overall, our study demonstrates that machine learning models, particularly those implemented through the Predomics framework, can effectively identify indicator fish species that are ecologically informative and strongly associated with specific habitat types, highlighting the potential of data-driven approaches to enhance our understanding of habitat–species relationships in complex marine ecosystems and support the development of targeted conservation and monitoring strategies using functionally relevant indicator taxa.

Our approach is not limited to the New Caledonian lagoon context. Because Predomics identifies stable, interpretable, and parsimonious indicators, it can be readily adapted to other reef systems by retraining the model with local assemblage data. The same workflow could be applied to coral reefs in the Caribbean or Indian Ocean, where similar habitat gradients and fish functional guilds occur (Pichler and Hartig, 2023; Stuart-Smith et al., 2013). Such adaptability supports the use of interpretable ML frameworks for global coral reef monitoring and for developing standardized, transferable sets of ecological indicators across regions.

Implications and Perspectives

The broader significance of our findings is that reef fish assemblages can be effectively characterized using a small subset of taxa, thereby reducing both sampling and computational demands without compromising ecological insight. This efficiency is particularly valuable for biodiversity monitoring in resource-limited contexts, a common challenge in tropical reef management.

Looking ahead, three key directions emerge from this study: (i) extend the framework to additional reef and marine systems to evaluate its cross-ecosystem applicability; (ii) explore temporal dynamics to track the stability of indicator species under environmental change; and (iii) integrate these minimal indicator sets into strategic conservation planning, adaptive management, and policy frameworks, particularly for marine protected area monitoring and evaluation.

5. Conclusion

This study demonstrates that an interpretable machine learning framework (Predomics) can reliably identify indicator species in reef fish assemblages, performing comparably to IndVal and TWINSPAN while selecting fewer, more parsimonious species sets. The approach achieves high predictive performance using only a small, ecologically meaningful subset of species — a parsimony shared by TWINSPAN, which identified the fewest indicators across all methods. This efficiency reduces sampling and computational effort while preserving interpretability and ecological relevance. Notably, the convergence of all four approaches on a common core of indicator species for Inshore–Offshore classification reinforces the robustness of these indicators across fundamentally different analytical frameworks. The main contribution of this work is methodological: it bridges classical ecological methods (IndVal and TWINSPAN) and modern interpretable machine learning, showing that data-driven models can yield biologically grounded and actionable insights. It also highlights that distinguishing closely related coastal habitats (e.g., Bay vs. Lagoon) remains challenging across all approaches. More broadly, the study underlines the value of parsimony for biodiversity monitoring under limited resources and demonstrates the practical applicability of interpretable machine learning for reef management. Future work should assess the transferability of these indicator sets across reef systems, explore their temporal stability under environmental change, and integrate them into automated, real-time monitoring frameworks for conservation planning.

CRediT authorship contribution statement

Eberrhard Estephe Kana Djifack: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Edi Prifti:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Thomas Lamy:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Florian Baletaud:** Writing – review & editing, Writing – original draft, Supervision, Methodology. **Jean-Daniel Zucker:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Norbert Tsopze:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Laurent Vigliola:** Writing – review & editing, Writing – original draft, Resources, Data curation, Supervision, Conceptualization. **Eugeni Belda:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization.

Funding

This work was supported by the AIME (Artificial Intelligence For Marine Ecosystems) project co-funded by the French National Research Agency (ANR) and the French Development Agency (AFD).

Declaration of competing interest

The authors declared that, this work was supported by the AIME (Artificial Intelligence For Marine Ecosystems) project co-funded by the French National Research Agency (ANR) and the French Development Agency (AFD).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ecoinf.2026.103704.

Data availability

All processed BRUVS data used for this article are archived in the public repository Zenodo here. The New Caledonian legislation regarding sensitive environmental data does not permit unrestricted public access. Accordingly, access to the data will require a Data Use Agreement (DUA), which will be systematically granted for reproducibility purposes. All the R scripts for IndVal, TWINSPAN and Predomics analyses and the evaluation notebooks are accessible on a public GitHub repository here.

References

Asnicar, F., Thomas, A.M., Passerini, A., Waldron, L., Segata, N., 2024. Machine learning for microbiologists. Nat. Rev. Microbiol. 22 (4), 191–205. http://dx.doi.org/10.1038/s41579-023-00984-1.

Baletaud, F., Gilbert, A., Mouillot, D., Come, J.M., Vigliola, L., 2022. Baited video reveal fish diversity in the vast inter-reef habitats of a marine tropical lagoon. Mar. Biodivers. 52 (2), 16. http://dx.doi.org/10.1007/s12526-021-01251-3, URL: https://link.springer.com/10.1007/s12526-021-01251-3.

Belda, E., Capeau, J., Zucker, J.D., Chatelier, E.L., Pons, N., Oñate, F.P., Quinquis, B., Alili, R., Fellahi, S., Katlama, C., Clément, K., Fève, B., Jaureguiberry, S., Goujard, C., Lambotte, O., Doré, J., Prifti, E., Bastard, J.P., 2024. Major depletion of insulin sensitivity-associated taxa in the gut microbiome of persons living with HIV controlled by antiretroviral drugs. BMC Med. Genom. 17 (1), 209. http://dx.doi.org/10.1186/s12920-024-01978-5.

Caldwell, I., McClanahan, T., Oddenyo, R., Graham, N., Beger, M., Vigliola, L., Sandin, S., Friedlander, A., Randriamanantsoa, B., Wantiez, L., Green, A., Humphries, A., Hardt, M., Caselle, J., Feary, D., Karkarey, R., Jadot, C., Hoey, A., Eurich, J., Wilson, S., Crane, N., Tupper, M., Ferse, S., Maire, E., Mouillot, D., Cinner, J., 2024. Protection efforts have resulted in ĩ0% of existing fish biomass on coral reefs. Proc. Natl. Acad. Sci. U. S. America 121 (42), http://dx.doi.org/10.1073/pnas.2308605121.

Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., Collins, J.J., 2018. Next-generation machine learning for biological networks. Cell 173 (7), 1581–1592. http://dx.doi.org/10.1016/j.cell.2018.05.015, Publisher: Elsevier. URL: https://www.cell.com/cell/abstract/S0092-8674(18)30592-0.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792. http://dx.doi.org/10.1890/07-0539.1, URL: http://doi.wiley.com/10.1890/07-0539.1.

Danovaro, R., Fanelli, E., Aguzzi, J., Billett, D., Carugati, L., Corinaldesi, C., Dell'Anno, A., Gjerde, K., Jamieson, A.J., Kark, S., McClain, C., Levin, L., Levin, N., Ramirez-Llodra, E., Ruhl, H., Smith, C.R., Snelgrove, P.V.R., Thomsen, L., Van Dover, C.L., Yasuhara, M., 2020. Ecological variables for developing a global deep-ocean monitoring and conservation strategy. Nat. Ecol. Evol. 4 (2), 181–192. http://dx.doi.org/10.1038/s41559-019-1091-z.

Dash, N.R., Al Bataineh, M.T., Alili, R., Al Safar, H., Alkhayyal, N., Prifti, E., Zucker, J.D., Belda, E., Clément, K., 2023. Functional alterations and predictive capacity of gut microbiome in type 2 diabetes. Sci. Rep. 13 (1), 22386. http://dx.doi.org/10.1038/s41598-023-49679-w.

De Cáceres, M., Legendre, P., Moretti, M., 2010. Improving indicator species analysis by combining groups of sites. Oikos 119 (10), 1674–1684. http://dx.doi.org/10.1111/j.1600-0706.2010.18334.x, URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0706.2010.18334.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1600-0706.2010.18334.x.

Dufrene, M., Legendre, P., 1997. Species assemblages and indicator species: The need for a flexible asymmetrical approach. Ecol. Monograph. 67 (3), 345–366. http://dx.doi.org/10.1890/0012-9615(1997)067[0345:SAAIST]2.0.CO;2, URL: http://doi.wiley.com/10.1890/0012-9615(1997)067[0345:SAAIST]2.0.CO;2.

Emslie, M.J., Bray, P., Cheal, A.J., Johns, K.A., Osborne, K., Sinclair-Taylor, T., Thompson, C.A., 2020. Decades of monitoring have informed the stewardship and ecological understanding of Australia's great barrier reef. Biol. Cons. 252, 108854. http://dx.doi.org/10.1016/j.biocon.2020.108854, URL: https://www.sciencedirect.com/science/article/pii/S0006320720309125.

Ford, C., 2023. The wilcoxon rank sum test | UVA library. URL: https://library.virginia.edu/data/articles/the-wilcoxon-rank-sum-test.

Froese, R., Pauly, D., 2024. FishBase. URL: https://fishbase.de/.

Gloor, G.B., Reid, G., 2016. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Can. J. Microbiol. 62 (8), 692–703. http://dx.doi.org/10.1139/cjm-2015-0821, URL: http://www.nrcresearchpress.com/doi/10.1139/cjm-2015-0821.

Halliday, W.D., Brittain, S.A., Niemi, A., Majewski, A.R., Mouy, X., Insley, S.J., 2022. The underwater soundscape of minto inlet, northwest territories, Canada. ARCTIC 75 (4), 462–479. http://dx.doi.org/10.14430/arctic76400, Number: 4, URL: https://journalhosting.ucalgary.ca/index.php/arctic/article/view/76400.

Halpern, B.S., Frazier, M., Afflerbach, J., Lowndes, J.S., Micheli, F., O'Hara, C., Scarborough, C., Selkoe, K.A., 2019. Recent pace of change in human impact on the world's ocean. Sci. Rep. 9 (1), 11609. http://dx.doi.org/10.1038/s41598-019-47201-9, Publisher: Nature Publishing Group, URL: https://www.nature.com/articles/s41598-019-47201-9.

Han, W., Han, Q., 2024. Macrobenthic indicator species: From concept to practical applications in marine ecology. Glob. Ecol. Conserv. 55, e03262. http://dx.doi.org/10.1016/j.gecco.2024.e03262, URL: https://www.sciencedirect.com/science/article/pii/S2351989424004669.

Hill, M., 1979. TWINSPAN—A fortran program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. In: Cornell's Ecology Program Series. Journal Abbreviation: Cornell's Ecology Program Series.

Juhel, J.B., Vigliola, L., Mouillot, D., Kulbicki, M., Letessier, T.B., Meeuwig, J.J., Wantiez, L., 2018. Reef accessibility impairs the protection of sharks. J. Appl. Ecol. 55 (2), 673–683. http://dx.doi.org/10.1111/1365-2664.13007, URL: https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13007.

Kumar, A., Garg, R.D., Singh, P., Shankar, A., Nayak, S.R., Diwakar, M., 2023. Monitoring the land use, land cover changes of roorkee region (uttarakhand, India) using machine learning techniques. Int. J. Soc. Ecol. Sustain. Dev. (IJSESD) 14 (1), 1–16. http://dx.doi.org/10.4018/IJSESD.316883, Publisher: IGI Global Scientific Publishing, URL: https://www.igi-global.com/article/monitoring-the-land-use-land-cover-changes-of-roorkee-region-uttarakhand-india-using-machine-learning-techniques/www.igi-global.com/article/monitoring-the-land-use-land-cover-changes-of-roorkee-region-uttarakhand-india-using-machine-learning-techniques/316883?utm_source=chatgpt.com.

Langlois, T., Goetze, J.S., Bond, T., Monk, J., Abesamis, R.A., Asher, J., Barrett, N., Bernard, A.T.F., Bouchet, P.J., Birt, M.J., Cappo, M., Currey-Randall, L.M., Driessen, D., Fairclough, D.V., Fullwood, L.A.F., Gibbons, B.A., Harasti, D., Heupel, M.R., Hicks, J., Holmes, T.H., Huveneers, C., Ierodiaconou, D., Jordan, A., Knott, N.A., Lindfield, S., Malcolm, H.A., McLean, D., Meekan, M., Miller, D., Mitchell, P.J., Newman, S.J., Radford, B., Rolim, F.A., Saunders, B.J., Stowar, M., Smith, A.N.H., Travers, M.J., Wakefield, C.B., Whitmarsh, S.K., Williams, J., Harvey, E.S., 2020. A field and video annotation guide for baited remote underwater stereo-video surveys of demersal fish assemblages. [GOOS ENDORSED PRACTICE]. http://dx.doi.org/10.25607/OBP-963, Accepted: 2020-11-26T15:58:34Z, URL: https://repository.oceanbestpractices.org/handle/11329/1461.

Lin, H., Peddada, S.D., 2020. Analysis of microbial compositions: a review of normalization and differential abundance analysis. NPJ Biofilms Microbiomes 6 (1), 60. http://dx.doi.org/10.1038/s41522-020-00160-w.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2 (1), 56–67. http://dx.doi.org/10.1038/s42256-019-0138-9.

Morales, A., Villalobos, F.J., 2023. Using machine learning for crop yield prediction in the past or the future. Front. Plant Sci. 14, http://dx.doi.org/10.3389/fpls.2023.1128388, Publisher: Frontiers, URL: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2023.1128388/full.

Nearing, J.T., Douglas, G.M., Hayes, M.G., MacDonald, J., Desai, D.K., Allward, N., Jones, C.M.A., Wright, R.J., Dhanani, A.S., Comeau, A.M., Langille, M.G.I., 2022. Microbiome differential abundance methods produce different results across 38 datasets. Nat. Commun. 13 (1), 342. http://dx.doi.org/10.1038/s41467-022-28034-z.

Nembrini, S., König, I.R., Wright, M.N., 2018. The revival of the gini importance? Bioinform. (Oxford, England) 34 (21), 3711–3718. http://dx.doi.org/10.1093/bioinformatics/bty373.

Panigrahi, S., Maski, P., Thondiyath, A., 2023. Real-time biodiversity analysis using deep-learning algorithms on mobile robotic platforms. PeerJ Comput. Sci. 9, e1502. http://dx.doi.org/10.7717/peerj-cs.1502, URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10495972/.

Pichler, M., Hartig, F., 2023. Machine learning and deep learning—A review for ecologists. Methods Ecol. Evol. 14 (4), 994–1016. http://dx.doi.org/10.1111/2041-210X.14061, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.14061, URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14061.

Prifti, E., Chevaleyre, Y., Hanczar, B., Belda, E., Danchin, A., Clément, K., Zucker, J.D., 2020. Interpretable and accurate prediction models for metagenomics data. GigaScience 9 (3), giaa010. http://dx.doi.org/10.1093/gigascience/giaa010, URL: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giaa010/5801229.

Ricotta, C., Carboni, M., Acosta, A.T., 2015. Let the concept of indicator species be functional!. J. Veg. Sci. 26 (5), 839–847, Publisher: Wiley, URL: https://www.jstor.org/stable/43912906.

Roleček, J., Tichý, L., Zelený, D., Chytrý, M., 2009. Modified TWINSPAN classification in which the hierarchy respects cluster heterogeneity. J. Veg. Sci. 20 (4), 596–602. http://dx.doi.org/10.1111/j.1654-1103.2009.01062.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1654-1103.2009.01062.x, URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1654-1103.2009.01062.x.

Russell, D.J., McDougall, A.J., 2005. Movement and juvenile recruitment of mangrove jack, Lutjanus argentimaculatus (Forsskål), in northern Australia. Mar. Freshwater Res. 56 (4), 465–475. http://dx.doi.org/10.1071/MF04222, Publisher: CSIRO PUBLISHING, URL: https://www.publish.csiro.au/mf/MF04222.

Shimwell, D.W., 1971. The Description and Classification of Vegetation. London, Sidgwick & Jackson, URL: http://archive.org/details/descriptionclass0000shim.

Siddig, A.A.H., Ellison, A.M., Ochs, A., Villar-Leeman, C., Lau, M.K., 2016. How do ecologists select and use indicator species to monitor ecological change? Insights from 14 years of publication in *ecological indicators*. Ecol. Indic. 60, 223–230. http://dx.doi.org/10.1016/j.ecolind.2015.06.036, URL: https://www.sciencedirect.com/science/article/pii/S1470160X15003696.

Simpfendorfer, C.A., Heithaus, M.R., Heupel, M.R., MacNeil, M.A., Meekan, M., Harvey, E., Sherman, C.S., Currey-Randall, L.M., Goetze, J.S., Kiszka, J.J., Rees, M.J., Speed, C.W., Udyawer, V., Bond, M.E., Flowers, K.I., Clementi, G.M., Valentin-Albanese, J., Adam, M.S., Ali, K., Asher, J., Aylagas, E., Beaufort, O., Benjamin, C., Bernard, A.T.F., Berumen, M.L., Bierwagen, S., Birrell, C., Bonnema, E., Bown, R.M.K., Brooks, E.J., Brown, J.J., Buddo, D., Burke, P.J., Cáceres, C., Cambra, M., Cardeñosa, D., Carrier, J.C., Casareto, S., Caselle, J.E., Charloo, V., Cinner, J.E., Claverie, T., Clua, E.E.G., Cochran, J.E.M., Cook, N., Cramp, J.E., D'Alberto, B.M., de Graaf, M., Dornhege, M.C., Espinoza, M., Estep, A., Fanovich, L., Farabaugh, N.F., Fernando, D., Ferreira, C.E.L., Fields, C.Y.A., Flam, A.L., Floros, C., Fourqurean, V., Gajdzik, L., Barcia, L.G., Garla, R., Gastrich, K., George, L., Giarrizzo, T., Graham, R., Guttridge, T.L., Hagan, V., Hardenstine, R.S., Heck, S.M., Henderson, A.C., Heithaus, P., Hertler, H., Padilla, M.H., Hueter, R.E., Jabado, R.W., Joyeux, J.-C., Jaiteh, V., Johnson, M., Jupiter, S.D., Kaimuddin, M., Kasana, D., Kelley, M., Kessel, S.T., Kiilu, B., Kirata, T., Kuguru, B., Kyne, F., Langlois, T., Lara, F., Lawe, J., Lédée, E.J.I., Lindfield, S., Luna-Acosta, A., Maggs, J.Q., Manjaji-Matsumoto, B.M., Marshall, A., Martin, L., Mateos-Molina, D., Matich, P., McCombs, E., McIvor, A., McLean, D., Meggs, L., Moore, S., Mukherji, S., Murray, R., Newman, S.J., Nogués, J., Obota, C., Ochavillo, D., O'Shea, O., Osuka, K.E., Papastamatiou, Y.P., Perera, N., Peterson, B., Pimentel, C.R., Pina-Amargós, F., Pinheiro, H.T., Ponzo, A., Prasetyo, A., Quamar, L.M.S., Quinlan, J.R., Reis-Filho, J.A., Ruiz, H., Ruiz-Abierno, A., Sala, E., de León, P.S., Samoilys, M.A., Sample, W.R., Schärer-Umpierre, M., Schlaff, A.M., Schmid, K., Schoen, S.N., Simpson, N., Smith, A.N.H., Spaet, J.L.Y., Sparks, L., Stoffers, T., Tanna, A., Torres, R., Travers, M.J., van Zinnicq Bergmann, M., Vigliola, L., Ward, J., Warren, J.D., Watts, A.M., Wen, C.K., Whitman, E.R., Wirsing, A.J., Wothke, A., Zarza-González, E., Chapman, D.D., 2023. Widespread diversity deficits of coral reef sharks and rays. Sci. (New York, N.Y.) 380 (6650), 1155–1160. http://dx.doi.org/10.1126/science.ade4884.

Stuart-Smith, R.D., Bates, A.E., Lefcheck, J.S., Duffy, J.E., Baker, S.C., Thomson, R.J., Stuart-Smith, J.F., Hill, N.A., Kininmonth, S.J., Airoldi, L., Becerro, M.A., Campbell, S.J., Dawson, T.P., Navarrete, S.A., Soler, G.A., Strain, E.M.A., Willis, T.J., Edgar, G.J., 2013. Integrating abundance and functional traits reveals new global hotspots of fish diversity. Nature 501 (7468), 539–542. http://dx.doi.org/10.1038/nature12529, Publisher: Nature Publishing Group, URL: https://www.nature.com/articles/nature12529.

Walsh, C., Stallard-Olivera, E., Fierer, N., 2024. Nine (not so simple) steps: a practical guide to using machine learning in microbial ecology. MBio 15 (2), e0205023. http://dx.doi.org/10.1128/mbio.02050-23.

Wee, A.K.S., Salmo III, S.G., Sivakumar, K., Then, A.Y.H., Basyuni, M., Fall, J., Habib, K.A., Isowa, Y., Leopardas, V., Peer, N., Artigas-Ramirez, M.D., Ranawana, K., Sivaipram, I., Suleiman, M., Kajita, T., 2023. Prospects and challenges of environmental DNA (eDNA) metabarcoding in mangrove restoration in southeast Asia. Front. Mar. Sci. 10, http://dx.doi.org/10.3389/fmars.2023.1033258, Publisher: Frontiers, URL: https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2023.1033258/full.

Yang, L., Chen, J., 2022. A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. Microbiome 10 (1), 130. http://dx.doi.org/10.1186/s40168-022-01320-0.

Zar, J.H., 2005. Spearman rank correlation. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of Biostatistics, first ed. Wiley, http://dx.doi.org/10.1002/0470011815.b2a15150, URL: https://onlinelibrary.wiley.com/doi/10.1002/0470011815.b2a15150.