

<https://doi.org/10.1038/s43247-026-03715-5>

The marine microbiome can accurately predict its chemical and biological environment



Emma Bell ^{1,2,4}, Karin Garefelt^{1,4}, Krzysztof T. Jurdzinski ^{1,4}, Luis F. Delgado¹, Fanny Lindrooth¹, Bengt Karlson ³ & Anders F. Andersson ¹ ✉

Microbial communities respond to physicochemical changes in the environment, making the microbiome a sensitive indicator of ecosystem status. Monitoring aquatic microbiomes is therefore essential for understanding ecosystem health and responses to change. While traditional monitoring relies on microscopy, DNA-based approaches that leverage advances in high-throughput sequencing are increasingly incorporated. Here, we evaluate the potential of using metabarcoding to predict abiotic and biotic parameters across spatiotemporal gradients of the Baltic Sea. The dataset comprises 397 seawater samples integrating prokaryotic and eukaryotic (16S and 18S rRNA gene) metabarcoding data with environmental measurements and plankton microscopy counts. Random Forest models based on metabarcoding data accurately predicted multiple physicochemical parameters and performed comparably to two other machine learning methods, XGBoost and TabPFN. Models using 16S rRNA gene data performed better than those using 18S rRNA gene data, with amplicon sequence variant-level yielding the most accurate results. Metabarcoding also exceeded plankton microscopy in predicting abiotic factors and effectively predicted the presence of phytoplankton and zooplankton genera from ≤ 1 L of water. Models trained on independent datasets accurately predicted several physicochemical parameters, though performance decreased for others highlighting challenges in transferability. Finally, metabarcoding-based predictions closely matched established eutrophication indicators of environmental status, demonstrating the utility of microbiome-based approaches for marine ecosystem monitoring and management.

Microorganisms and the ecological niches they inhabit have a reciprocal relationship. Microbial metabolism drives element and nutrient cycles and is constrained by environmental factors such as temperature, pH, salinity, redox potential, and available carbon and nutrient sources. At the same time, the physicochemical environment shapes the composition of the microbiome, since each microbial strain has a unique fitness distribution in a multidimensional niche space^{1,2}. The close link between microorganisms and their ecological niche means microbiome data can provide insight into the status of the ecosystem^{3–6}.

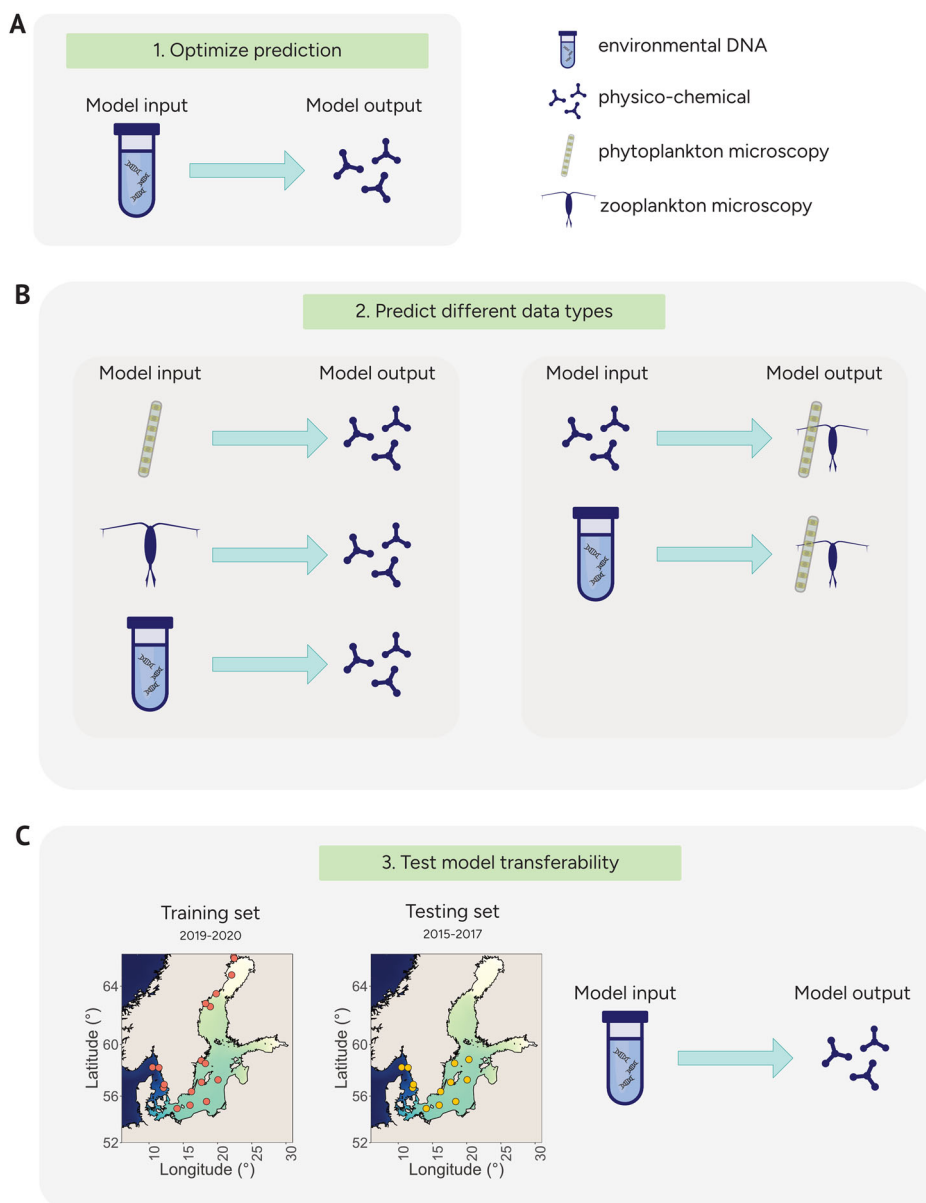
Marine environmental monitoring typically includes measurements of physical, chemical, and biological parameters. Microscopy is the traditional method for monitoring marine microorganisms and has been used for

decades in long-term monitoring programmes^{7–9}. The morpho-taxonomic data provide crucial information on the ecology of key plankton taxa and can serve as an indicator of ecosystem state¹⁰. However, the method is time-consuming, requires taxonomic expertise, and has limited taxonomic scope. Importantly, small microbes such as picoeukaryotes and bacterioplankton cannot typically be identified taxonomically despite carrying out important ecosystem functions. In contrast, metabarcoding and shotgun metagenomics make it possible to study entire microbial communities directly from environmental DNA^{11,12}. With the decreasing cost of high-throughput sequencing, these methods have become widely accessible, facilitating large-scale studies of marine microbial diversity¹³. Although these methods do not capture all the information obtained with microscopy, such as life stages, cell

¹Department of Gene Technology, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden. ²Baltic Sea Centre, Science for Life Laboratory, Stockholm University, Stockholm, Sweden. ³Department of Research and Development, Oceanography, Swedish Meteorological and Hydrological Institute, Västra Frölunda, Sweden. ⁴These authors contributed equally: Emma Bell, Karin Garefelt, Krzysztof T. Jurdzinski.

✉ e-mail: anders.andersson@scilifelab.se

Fig. 1 | Overview of study design. **A** Evaluation of different machine learning architectures and feature extraction techniques for physicochemical data prediction from metabarcoding data. **B** Comparison of physicochemical predictions based on metabarcoding or microscopy data, and of microscopy plankton presence/absence data using metabarcoding or physicochemical data. **C** Evaluation of model performance across datasets. Filled circles denote sampling points in the training set (red) and testing set (yellow), and the background colour denotes salinity.



sizes and potential colony-formation of the studied plankton, and the sequence data are not readily translated into cell counts or biomass, they do provide a more unbiased view of the microbial community and are increasingly argued to be included in routine monitoring programmes^{14,15}.

Recent research has adopted machine learning methods, such as Random Forests or Support Vector Machines, to harness the predictive power of microbiome data to categorise or forecast ecosystem conditions^{16,17}. Metabarcoding and metagenomics data have, for example, been used to predict soil health⁴, detect contamination in groundwater and sediment^{18–20}, predict physicochemical properties of lake water²¹, dissolved organic carbon in soil²², environmental status of freshwater reservoirs²³ and headwater streams²⁴. While most studies employing machine learning for this purpose train the models directly on sequence counts, sometimes aggregated at different taxonomic levels, recent studies have also employed deep learning approaches to obtain a smaller set of informative features to train the models on²⁵. A key challenge in generating accurate predictions is obtaining high-quality training datasets that capture the full range of an ecosystem's variability, including environmental gradients^{26,27}, seasonal changes, and response to disturbances²⁸, which are needed to reliably train machine learning models²⁹.

Machine learning using metabarcoding data has been used in a marine context, for example, for assessing the environmental impact of marine aquaculture^{30–32} or predicting bacterial production^{33,34}, but a systematic assessment of how effectively it can predict different environmental parameters of seawater is missing. Moreover, since many monitoring programs involve microscopy-based plankton identification, a comparison of the predictive power of the morpho-taxonomic data and metabarcoding data is relevant. Similarly, predicting microscopy counts would make metabarcoding studies more comparable with long-term microscopy-based time series, but has not been attempted yet. Finally, the transferability of microbiome-based predictive models when applied to a different dataset remains to be assessed.

Here, we use metabarcoding data to predict abiotic (physicochemical) and biotic (phytoplankton and zooplankton composition) parameters in the Baltic Sea (Fig. 1). We used a high-quality spatiotemporal dataset of prokaryotic and eukaryotic microbial communities that incorporates physicochemical and metabarcoding data from 397 samples sampled at 18 stations across the Baltic Sea, Kattegat, and Skagerrak in 2015–2017 and 2019–2020^{35,36}. Changes in time (seasonality) and space (geographical region and salinity gradient) are reflected in the datasets, making them ideal

for assessing the feasibility of using metabarcoding data as an environmental biosensor. We systematically evaluate the prediction performance of prokaryotic and eukaryotic marker genes (16S rRNA and 18S rRNA) at different taxonomic levels, feature extraction techniques, machine learning algorithms, and compare the performance of microscopy-based and metabarcoding-based predictions. Our dataset comprised two separate sampling and sequencing efforts, which allowed us to compare interannual predictions across different datasets with training and deploying on the same dataset. We further demonstrate this potential by predicting Baltic Marine Environment Protection Commission (HELCOM) indicators³⁷, which inform policy and management decisions to protect the ecological status of the Baltic Sea.

Results

Physicochemical properties of seawater can be accurately predicted from the microbiome with machine learning

Random Forest machine learning models were trained to predict physicochemical properties of seawater from metabarcoding data (Fig. 1A). Accuracy was assessed by comparing the true values to the predicted values for each model using the coefficient of determination (R^2 ; Root Mean Square Error also reported in Supplementary Table S1), using 10-fold cross validation. Metabarcoding-based models accurately predicted physicochemical properties of seawater (Fig. 2A, B). 16S rRNA gene data outperformed 18S rRNA gene data in predictive accuracy across all parameters, except chlorophyll-a (Chl-a), and for ASV level 16S performed significantly better than 18S ($P < 0.001$; Wilcoxon signed-rank test). For the models based on 16S rRNA gene, the median R^2 was 0.844, illustrated by the prediction of nitrate (Fig. 2C). Salinity and alkalinity were predicted with the greatest accuracy ($R^2 = 0.981$ and 0.979 , respectively), whereas ammonium (NH_4^+) and chlorophyll-a had the poorest ($R^2 = 0.491$ and 0.249). Overall, physicochemical parameters ranked similarly in terms of predictive performance across 16S and 18S rRNA gene-based models, although 16S rRNA gene-based models had higher overall accuracy.

We tested whether aggregation of data at higher taxonomic levels would improve prediction performance compared to using amplicon sequence variants (ASVs). Models trained with ASVs as input features had greater predictive accuracy than models trained with data aggregated at higher taxonomic levels for both 16S rRNA gene and 18S rRNA gene data (Fig. 2D). For models using 16S rRNA gene data, prediction performance decreased progressively with each step of taxonomic aggregation (ASV $\rightarrow$ phylum). In contrast, for models using 18S rRNA gene data, ASV-based predictions worked best followed by genus and family. As the ASV level performed best for both marker genes, ASV data were used as input for all subsequent analyses.

Using ASV-level metabarcoding data as input, we compared two alternative machine learning approaches to Random Forest; XGBoost and TabPFN. All three approaches demonstrated strong predictive capabilities, underscoring the efficacy of metabarcoding data for predicting physicochemical properties of seawater. XGBoost performed significantly better than Random Forest and TabPFN for both 16S and 18S rRNA gene data, whereas Random Forest and TabPFN were comparable (Fig. 2E). Random Forest models trained on ASVs completed in ~ 14 s wall-clock time per parameter with 32 GB of available RAM and without need for parallelisation. In contrast, XGBoost models trained on ASV data only with hyperparameter tuning ($\sim 1,632$ grid combinations) required 1–1.5 h per parameter, using parallel computation across 8 CPU cores. As Random Forest required substantially less computational resources than XGBoost and had comparable overall performance (mean R^2 0.786 for XGBoost versus 0.778 for Random Forest with 16S rRNA ASV data) we selected Random Forest for the remaining analyses.

Next, we transformed the metabarcoding data into lower-dimensional embeddings using deep representation learning techniques prior to training the Random Forests²⁵, to assess whether this would improve prediction performance. Out of the embedding methods, the convolutional autoencoder (CAE) yielded the best results (Fig. 2F). However, training on ASV

profiles outperformed all deep encoding embeddings (Fig. 2F), and therefore ASV profiles without additional embedding were used in the subsequent analyses. However, selecting only ASVs present in at least 10% of the samples outperformed using all the ASVs (Supplementary Fig. S1), and we used this feature selection approach in all the following analyses.

Metabarcoding data outperform microscopy-based monitoring data for predicting abiotic parameters

The monitoring of phytoplankton and zooplankton with microscopy is an integral component of many marine monitoring programmes, and is used to assess ecosystem status. We therefore compared the effectiveness of microscopy to metabarcoding data for predicting physicochemical parameters (Fig. 1B). Random Forest models trained with either 16S or 18S rRNA gene ASV profiles performed consistently better than models trained on microscopy counts of both phytoplankton and zooplankton (Fig. 3). The only exception was chlorophyll-a, the parameter most poorly predicted with metabarcoding, where predictions worked best, though still poorly, with phytoplankton microscopy data (Fig. 3A).

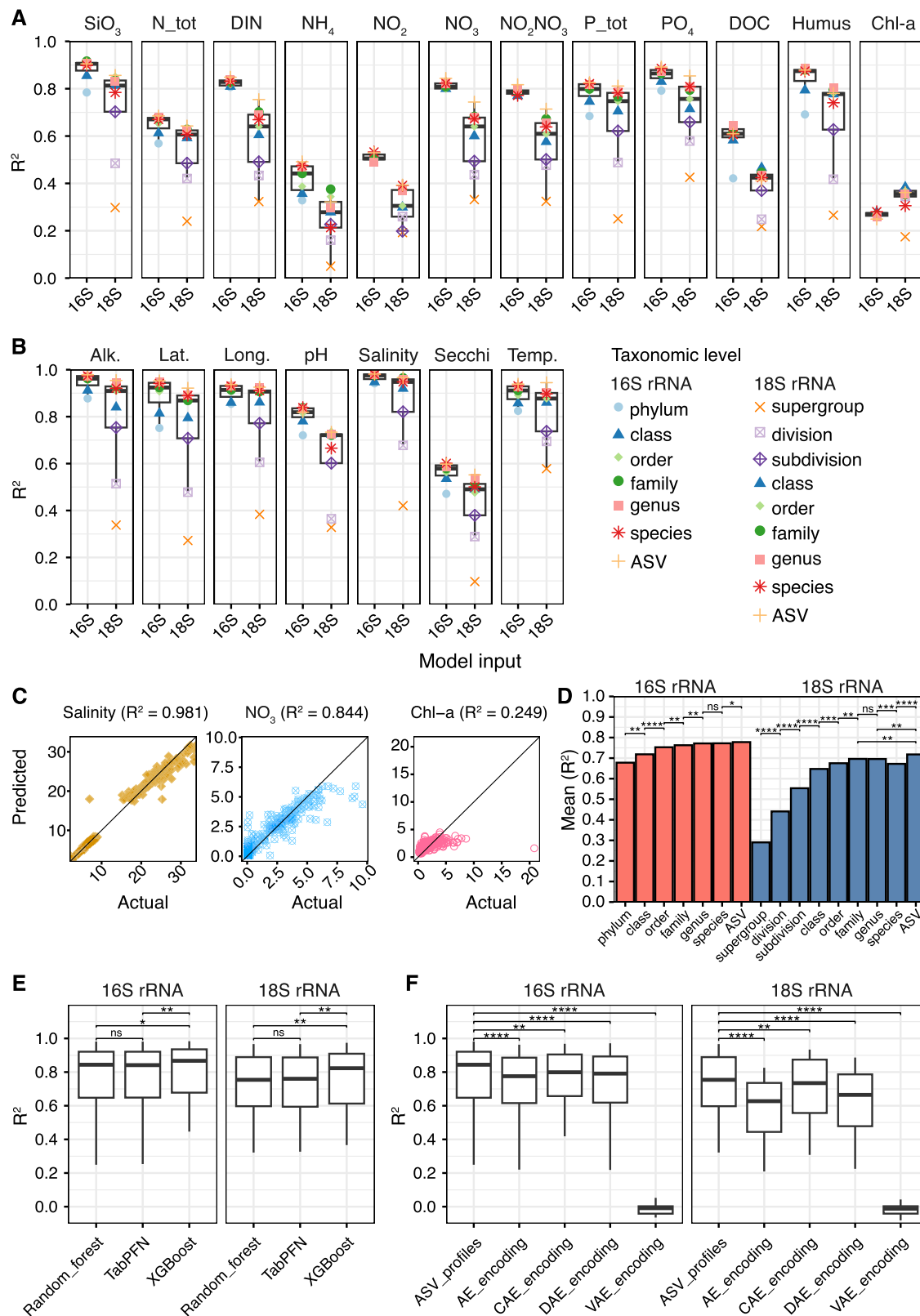
Metabarcoding can predict the presence of both phytoplankton and zooplankton taxa

Our results show that metabarcoding data accurately predicts physicochemical parameters of seawater (Figs. 2 and 3). We next assessed whether it can also predict biotic features (Fig. 1B). In this case, we focused on predicting the presence of specific phytoplankton and zooplankton genera, and compared training on either metabarcoding or physicochemical data since the latter are often the basis for generating species distribution models of plankton^{38–40}. For each eukaryotic plankton genus detected with both microscopy and metabarcoding (according to the taxonomic annotation of the sequences) we predicted its presence or absence in the microscopy data with Random Forest models trained on (1) physicochemical data, (2) 16S rRNA ASV data, (3) 18S rRNA ASV data, and compared these predictions to (4) using the presence/absence of the genus in the 18S rRNA gene data (Fig. 4). For both phytoplankton and zooplankton, machine learning predictions with ASV profiles had accuracies on par with those based on physicochemical data, and 16S and 18S ASV data performed equally well (Fig. 4C, I). As expected, the accuracy of the predictions increased with the number of samples in which the genus was identified with microscopy (Fig. 4A, B, D, G, H, J). Interestingly, machine learning using ASV profiles consistently outperformed using the presence/absence of the genus in the 18S rRNA gene data (Fig. 4F, L). This was most pronounced for zooplankton, where median F1-score with machine learning were ~ 0.96 compared to ~ 0.56 using the presence/absence of the corresponding genus in the metabarcoding data (Fig. 4I). In addition to predicting presence/absence, we predicted the counts obtained with microscopy of each taxon. Performance was considerably worse than predicting presence/absence, but the overall trends were the same (Supplementary Fig. S2).

Interannual predictions across datasets

We evaluated the temporal transferability of our machine learning models by training them on data from one time period and evaluating them on data from a different, non-overlapping time period (Fig. 1C). Specifically, we compared three strategies for predicting physicochemical data: (1) training on a different dataset (2019–2020) to the test data (2015–2017), (2) training with the same dataset as the test data (2015–2017), and (3) training using both datasets (2015–2017 and 2019–2020) for predicting the test data (2015–2017). We chose to use the 2015–2017 dataset as the test data, due to its narrower geographic coverage relative to the 2019–2020 data, thus avoiding extrapolation outside the geographic environmental gradients of the training data (Fig. 1C).

The performance of the interannual predictions (“different dataset”) compared to using the “same dataset” or “both datasets” models varied between parameters (Fig. 5A). In general, using the same dataset worked better than using a different dataset, and using both datasets worked best (Fig. 5B). However, for the six parameters best predicted with the “same



dataset” model ($R^2 > 0.8$), the interannual predictions performed comparably (Fig. 5C; $P = 0.56$).

We further assessed whether the poorer performance of some “different dataset” models reflect noisy or biased predictions. To this end, for all the predicted parameters, we obtained R^2 values from linear regression between predicted and actual values, as a measure of variance ignorant of bias. The

linear regression-based R^2 values did not significantly differ between the “different dataset” and “same dataset” models ($P = 0.28$), while the “both datasets” model was more precise than both other models ($P < 0.01$). This indicates that the poorer performance of the “different dataset” model for some parameters (e.g., total phosphorus; Supplementary Fig. S3B) is due to bias, i.e. systematic over or under estimation, rather than noise.

Fig. 2 | Performance of machine learning models for predicting physicochemical parameters from metabarcoding data. In all plots, performance is presented as coefficient of determination (R^2) between observed and predicted values.

A Predictive performance of Random Forest models for nutrient and biological parameters using metabarcoding data aggregated at different taxonomic levels. **B** Predictive performance of Random Forest models for physical parameters using metabarcoding data aggregated at different taxonomic levels. **C** Scatter plots of actual versus predicted data for best (Salinity), median (NO_3), and worst (Chl-*a*) performing Random Forest models using 16S rRNA gene ASV data. **D** Mean model performance at different taxonomic levels across all environmental parameters for

models trained on 16S (red bars) and 18S (blue bars) rRNA gene metabarcoding. **E** Predictive performance of three machine learning algorithms: Random Forest, XGboost, and TabPFN, for predicting all physicochemical parameters from ASV data. **F** Predictive performance of Random Forest models with deep encoding architectures: shallow autoencoder (AE), convolutional autoencoder (CAE), deep autoencoder (DAE), variational autoencoder (VAE), for predicting all physicochemical parameters from ASV profiles. For each type of autoencoder, different architectures were tested, and only the best performing is shown. Asterisks indicate significance of Wilcoxon signed-rank tests: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Symbols in (A, B) denote taxonomic rank.

Predicting indicators for good environmental status

To support ecological status assessments and inform management decisions in the Baltic Sea, the Baltic Marine Environment Protection Commission (HELCOM) uses a set of indicators across the themes: Biodiversity, Eutrophication, Maritime Activities, and Hazardous Substances⁴¹. For eutrophication, the Ecological Quality Ratio (EQR) values are used to assess deviation from reference conditions, where values close to one indicate good environmental status and values close to zero indicate poor ecological status⁴².

We evaluated the potential of metabarcoding data to predict HELCOM eutrophication indicators using the “different dataset” model, and calculated the EQR based on both predicted and measured values of dissolved inorganic phosphate, water transparency (Secchi depth), total nitrogen, or total phosphorus. EQR estimates based on predicted physicochemical parameters matched those based on measured concentrations ($R^2 = 0.67$, mean absolute error MAE = 0.058) (Fig. 6A). The assessment of eutrophication status based on measured data (Fig. 6B) is closely replicated by the predicted data (Fig. 6C).

Discussion

Using metabarcoding data, we show that the taxonomic structure of microbial communities can be used to predict a range of both biotic and abiotic parameters in seawater. Using different machine learning approaches, physicochemical parameter predictions based on ASV-level 16S rRNA gene metabarcoding consistently outperformed those based on other feature types, showing that bacterial community structure holds a strong predictive power. Among machine learning models tested, the gradient boosting framework XGBoost achieved the highest overall accuracy but required substantially more computational resources than Random Forest. Random Forest models trained directly on ASV data outperformed those built using features extracted with autoencoders. Other studies have also found that feature extraction does not necessarily improve the performance of ensemble tree methods for environmental prediction⁴³. Considering both predictive performance and training time, we selected the Random Forest without feature extraction for predicting abiotic and biotic parameters from microbiome data. However, selecting only prevalent features (ASVs present in at least 10% of the samples) both increased the accuracy and decreased the memory usage and training time in our analyses. The generally high performance of all machine learning architectures, however, demonstrates the overall robustness of metabarcoding data for predicting physicochemical parameters. These results highlight the potential of microbial community profiles to serve as biosensors for marine monitoring and to be integrated into predictive frameworks that enhance our understanding of ecosystem state and function.

Overall, model performance varied more between the predicted parameters than between the machine learning approaches used, indicating that data characteristics have a greater impact on model performance. The physicochemical parameters that were most accurately predicted – temperature, salinity, and alkalinity – are known to be important for structuring the marine microbiome. Both salinity^{26,36,44} and temperature^{13,45} are key drivers of marine microbiome composition. The predictive power of these physical parameters likely reflects both their biological relevance and the consistency with which they are being measured. These parameters are easy to measure accurately with no difference in methodology between different

sampling teams, instruments, or years, thus reducing potential analytical variability due to measurement limitations. Additionally, temperature and salinity would undergo rapid changes only with the inflow of another water body mass. Meanwhile, other factors can be rapidly transformed by the microorganisms themselves, originating from or resulting in physiological changes. It takes time for these changes to influence community composition through altered growth rates. In such cases, the machine learning models could output false values reflecting the previous state of the system. This is unlikely to be the case for salinity or temperature. In contrast, the physicochemical parameters that had the weakest predictive performance were chlorophyll-*a* and ammonia. The poor prediction of chlorophyll-*a* is intriguing given its wide usage in environmental monitoring. A potential explanation could be that fluorometric chlorophyll-*a* measurements, as used here, have been affected by hour-to-hour changes connected to non-photochemical quenching^{46,47}. Since this reflects short-term physiological adaptation mechanisms in the phytoplankton, it will not be mirrored in the community composition, and thus cannot be predicted from metabarcoding. Other unidentified phenomena might have also comprised the chlorophyll predictions. Meanwhile, ammonia is considered one of the most challenging nutrients to measure accurately in seawater analysis⁴⁸. Furthermore, in our data, the majority of ammonia data values are low with a few outliers (Supplementary Fig. S4). Thus, both analytical variability and spread of data likely influence the model performance.

Model performance was greater when using ASV profiles compared to higher taxonomic levels or microscopy data, highlighting the value of high-resolution microbial community information. This aligns with a previous study on lakes where Random Forest models based on operational taxonomic units (OTUs) outperformed those based on higher taxonomic levels²¹. Another study focusing on headwater streams reported better performance using order-level classification than OTUs, but this study used a different algorithmic approach (Discriminant Analysis of Principal Components)²⁴, unassessed here. In the case of 16S rRNA gene data, prediction accuracy monotonically increased with finer taxonomic levels. For 18S rRNA gene data, however, the best performance after ASVs was at the family level. This likely reflects lower annotation rates at finer taxonomic levels for 18S, resulting in the exclusion of a substantial fraction of the data. As reported by Jurdzinski et al.³⁶, 78% and 44% of 16S rRNA gene reads in this dataset are annotated to genus and species levels, respectively, compared to 44% and 11% for 18S rRNA gene reads³⁶. This points to a limitation of 18S rRNA gene reference databases and/or the taxonomic resolution of the amplified region (V4). Future studies using long-read sequencing that enables the recovery of full-length 16S rRNA or 18S rRNA ASVs may have the potential to further improve prediction performance.

Microscopy-based identifications of phytoplankton and zooplankton genera were better predicted using Random Forest models based on full ASV profiles than with the direct annotation of amplicon sequences with taxonomic databases. As the difference was strongest for zooplankton, we interpret that this reflects that the sample volume used for metabarcoding (<1 L) is too small to accurately quantify mesozooplankton taxa that are present in low concentrations. In comparison, cubic metres of water are sampled with zooplankton nets for microscopy counting. Given this, it is remarkable that machine learning using microbiome data from the small sample volumes can so accurately predict the presence/absence of zooplankton taxa.

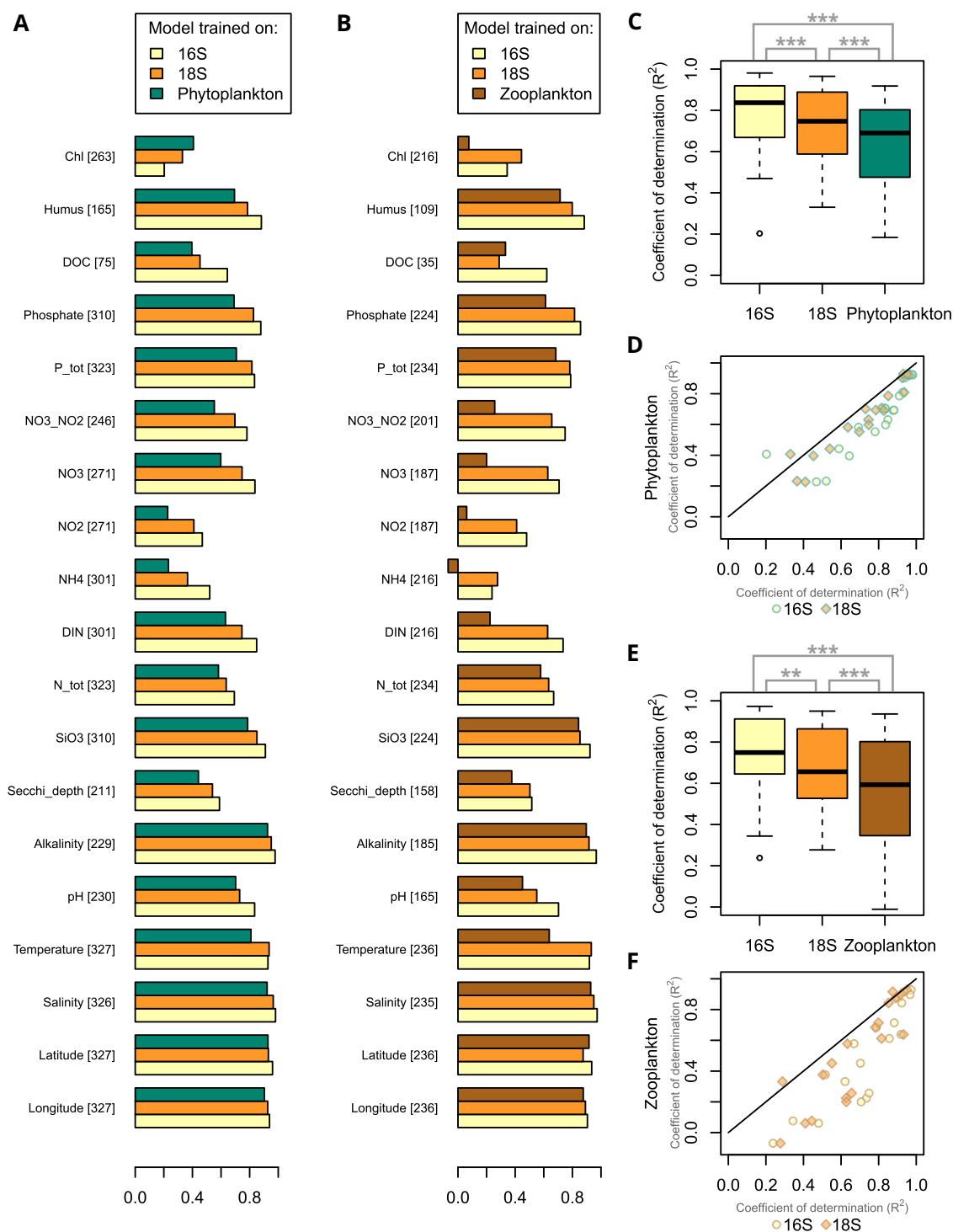
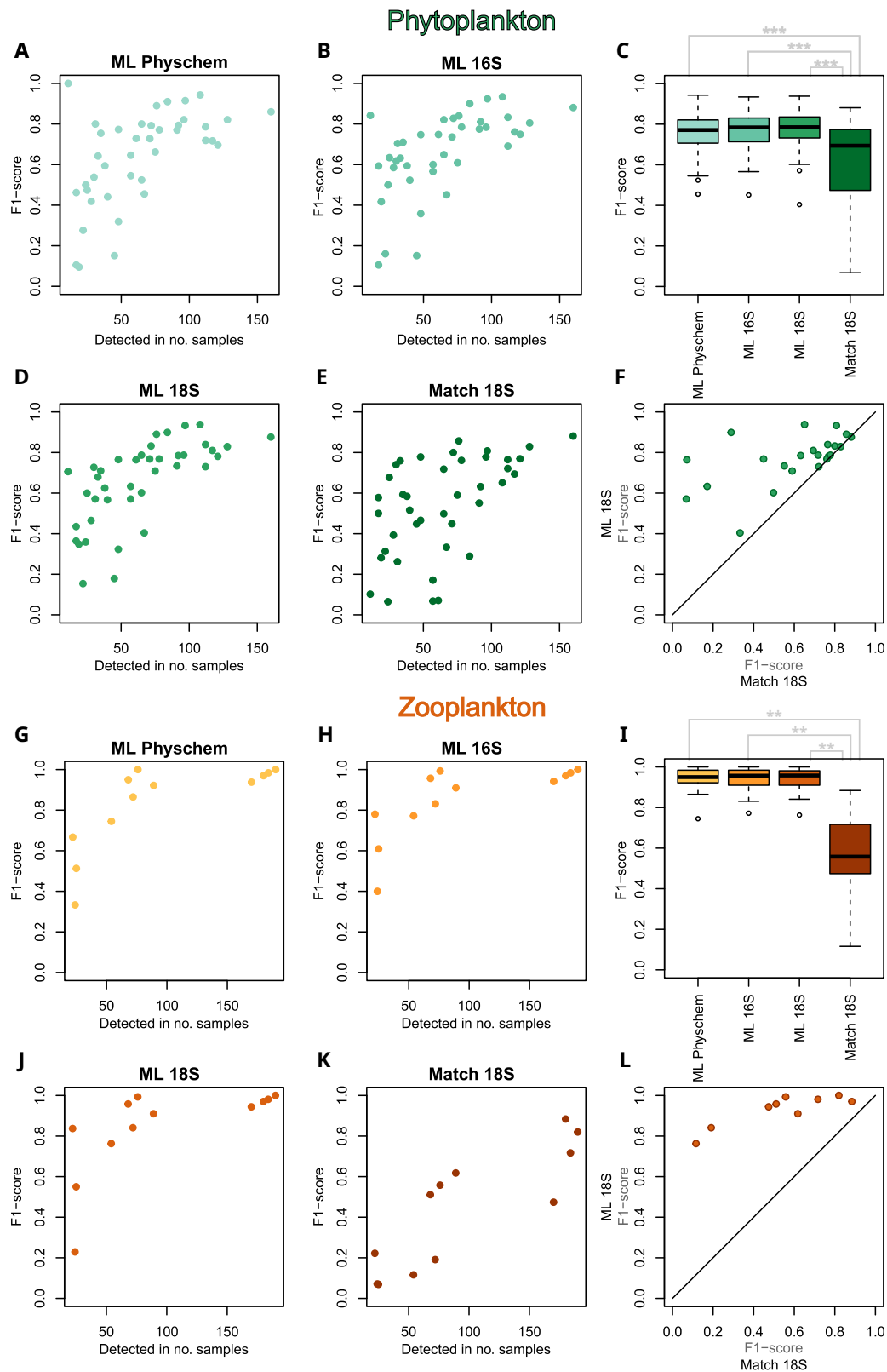


Fig. 3 | Performance of physicochemical parameter predictions using metabarcoding versus microscopy data. Machine learning (Random Forest) trained on either 16S or 18S rRNA gene ASV profiles, or microscopy counts of phytoplankton (A, C, D) or zooplankton (B, E, F). The microscopy-counted taxa were identified to either the genus or species level. The same physicochemical parameters are evaluated in all the panels of the figure. Numbers in brackets in (A) and (B)

indicate the number of samples included; only samples with 16S and 18S rRNA gene metabarcoding and microscopy data of either phytoplankton or zooplankton were included. Asterisks indicate significance of Wilcoxon signed-rank tests: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Bar and symbol colours indicate the type of input data used by the models: phytoplankton microscopy (green), zooplankton microscopy (brown), 18S rRNA gene profiles (orange) and 16S rRNA gene profiles (yellow).

Our results show that models trained on data from different years (“different dataset” models) can achieve similar accuracy to models trained on the same dataset for parameters that are well predicted by the latter. However, for other parameters, “different dataset” predictions were systematically worse. In general, these models were comparably precise as the “same dataset” models but less accurate due to biased predictions. For a few

parameters, the “different dataset” models performed unusually worse than the other models. In these cases, the “different dataset” models gave similar predictions irrespective of actual values (Secchi depth and nitrite, Supplementary Fig. S3A, C) or underestimated the actual values (total phosphorus, Supplementary Fig. S3B). The poorer “different dataset” predictions may arise from differences in physicochemical measurements between the



sampling years. Models may also be less transferable between datasets due to contextual differences between sampling years, as microbial communities and their associated environmental conditions are expected to change over time. The validity of models is therefore inherently dynamic. Our data suggest models are robust and transferable across 4–5 years separation, however, longer-term application would require periodic retraining and

validation to ensure coverage of the prevailing range of environmental conditions. Importantly, the “both dataset” model was both more accurate and precise than the “same dataset” model, showing that increasing the volume and diversity of training data improves model performance.

Most of the physicochemical parameters predicted here are routinely measured in the Swedish pelagic marine monitoring program and in

Fig. 4 | Performance of methods to predict the presence of eukaryotic phytoplankton and zooplankton genera. The presence or absence of each eukaryotic phytoplankton (A–F) and zooplankton (G–L) genus in individual samples was predicted using Random Forest models trained on physicochemical variables (ML Physchem), 16S rRNA gene ASV profiles (ML 16S), or 18S rRNA gene ASV profiles (ML 18S). In addition, genus presence/absence was inferred directly from 18S rRNA gene data (Match 18S) based on if sequences annotated to the genus were present. Only genera detected with microscopy and 18S rRNA gene metabarcoding (according to the taxonomic annotation of the sequences) in at least one sample each were included. In C, F, I and L, only genera detected with microscopy in >50 samples were included. F and L present comparisons of the accuracy of predictions obtained

with Random Forest using 18S rRNA gene ASV profiles (y-axis) with inference directly from the 18S rRNA gene data (x-axis). “Detected in no. of samples” refers to the number of samples in which a given genus was identified by microscopy. Only physicochemical parameters measured in at least 80% of the samples were used to build the models (see Methods for details), and only samples where these parameters were measured and that had both 16S and 18S metabarcoding data were included. For phytoplankton, 206 samples and for zooplankton, 190 samples were analysed. Asterisks indicate significance of Wilcoxon signed-rank tests: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Throughout the figure, green represents phytoplankton and brown represents zooplankton, with different shades indicating the different prediction methods.

environmental monitoring more broadly. However, as exemplified by the data used in this study, direct measurements are missing for many samples and parameters (Supplementary Fig. S4). If metabarcoding data are collected for microbiome monitoring, they can be used to help fill these gaps. Furthermore, microbiome-based models trained on comprehensive monitoring datasets, as done here, can provide valuable contextual information for smaller studies in the region by estimating environmental conditions at sites where only a subset of measurements is available. By predicting missing parameters, these models can support comparative analyses and situate local observations within a broader regional framework. If research questions arising require physicochemical measurements that were not collected, these cannot be done post factum, but they can be estimated using the methods presented here. Importantly, microbiome-based predictions can be extended to additional factors with minimal extra cost. Using the microbiome as a biosensor enables a multitude of information to be retrieved from a single sample, using one preparation and analysis pipeline. Our primary aim was to evaluate multiple modelling approaches and to determine whether microbial community composition can reliably predict environmental factors. Given its strong predictive power for a range of physicochemical and biological variables, the approach used here could likely be extended to infer new targets such as pollutants or toxins⁴².

In the context of the Baltic Sea, we explored microbiome-based prediction potential as an indicator of ecosystem health. Given that EQR ranges from 0 to 1, the level of predictive accuracy achieved (MAE = 0.058) suggests sufficient resolution to support ecological classification or to flag potential eutrophication hotspots. Such predictions could be used to screen for locations with potential changes in environmental status, which could then be confirmed with more direct measurements. Our findings are in line with earlier work demonstrating that microbiome-based machine learning models can reliably predict biotic indices to assess environmental impacts, for example, in marine aquaculture³¹. This supports the use of microbial profiles as indicators of ecosystem health.

In conclusion, this study highlights the ability of 16S and 18S rRNA gene metabarcoding to serve as biosensors for marine environmental monitoring. Using machine learning, we accurately predicted a range of biotic and abiotic variables, including key HELCOM indicators. Bacterial ASV data consistently yielded the highest performance across a broad range of parameters. By comparing the predictability of different variables, we map the strengths and challenges of metabarcoding-based machine learning. Our findings show that microbial community data, combined with robust models and diverse training datasets that comprehensively represent environmental variability, can inform ecological assessments across time and space, offering a scalable tool for environmental monitoring.

Methods

Datasets

Environmental data from the Swedish pelagic marine monitoring program include a range of physical, chemical, and biological properties of seawater and are collected according to the guidelines of the HELCOM-COMBINE manual⁴⁹. The datasets are freely available, and we downloaded them from the SHARK database (<https://shark.smhi.se/hamta-data/>). We selected physicochemical parameters measured for at least 10% of the samples in the dataset: salinity, temperature, pH, alkalinity, Secchi depth, SiO₃, total

nitrogen (N_{tot}), dissolved inorganic nitrogen (DIN), NH₄⁺, NO₂[−], NO₃[−], combined measurements of NO₃[−] and NO₂[−], total phosphorous (P_{tot}), PO₄, dissolved organic carbon (DOC), humus, and chlorophyll-a. In addition, sample date and time, longitude, and latitude were available for all samples. In total, we used 397 samples to create models from 16S rRNA or 18S rRNA gene amplicon data, with one sample from the original dataset³⁶ missing physicochemical data. The average data input for each parameter was 311 data points ranging from $n = 75$ (DOC) to $n = 378$ (temperature) (Supplementary Fig. S4).

We also retrieved microscopy-based counts of phytoplankton and zooplankton from the Swedish pelagic marine monitoring program from the SHARK database. Phytoplankton were sampled using an integrated hose sample, typically from 0 to 10 m depth, preserved in Lugol's solution, settled in sedimentation chambers, and identified and counted with inverted light microscopy⁵⁰ and reported as individuals per L. Zooplankton were collected with vertically towed nets, with depth ranges varying between stations (from 0–6 m to 0–193 m depth). The zooplankton were identified and counted with light microscopy, and abundances were reported as individuals per m³. For each taxon, we summed counts across size classes (phytoplankton) and growth stages and sexes (zooplankton) when applicable. Only taxa identified to at least the genus level, and for zooplankton belonging to Arthropoda and Rotifera, were included in the analyses. Samples with both metabarcoding and plankton data were available from 18 stations for phytoplankton and 16 stations for zooplankton. When predicting physicochemical parameters using microscopy data, we used plankton counts (individuals per volume unit). When we predicted microscopy identifications of plankton genera using metabarcoding or physicochemical data, we first aggregated microscopy counts at the genus level, and then either individuals per volume unit (Supplementary Fig. 2) or presence/absence (main results, Fig. 4) were predicted. For the latter, we converted the microscopy counts to binary data (presence/absence).

Biomass for microbial DNA has been collected and sequenced using DNA metabarcoding in parallel with the pelagic marine monitoring program in multiple sampling efforts, from distinct integrated hose samples, typically from 0 to 10 m depth. We used previously published amplicon sequencing data with ENA accession numbers PRJEB55296³⁵ and PRJEB84926³⁶. These amplicon datasets contain both 16S rRNA (V3–V4 region) and 18S rRNA (V4 region) gene metabarcoding that capture prokaryotic (with good coverage of bacteria and poor coverage of archaea⁵¹) and eukaryotic communities, respectively. They were created with the primer pairs 341F CCTACGGGNGGCWGCAG and 805R GACTACHVGGTA TCTAATCC⁴⁴ (for the 16S rRNA gene) and V4F CCAGCASCYCGCGG-TAATTCC and V4RB ACTTTCGTTCTTGATYRR⁵² (for the 18S rRNA gene). Amplicon datasets were processed with the nf-core/ampliseq pipeline⁵³, which implements DADA2⁵⁴ to produce amplicon sequence variants (ASVs). ASVs were taxonomically annotated within DADA2 using a curated version⁵⁵ of the Genome Taxonomy Database (GTDB; version R06-RS202)⁵⁶ for 16S rRNA gene and the Protist Ribosomal Reference database (PR²; version 5.0.0)⁵⁷ for 18S rRNA gene. For 16S rRNA gene, ASVs were also annotated with SILVA⁵⁸ (version 138.1) in order to identify chloroplast and mitochondrial sequences. Further details on the amplicon data generation and processing are available in Latz et al.³⁵ and Jurdzinski et al.³⁶.

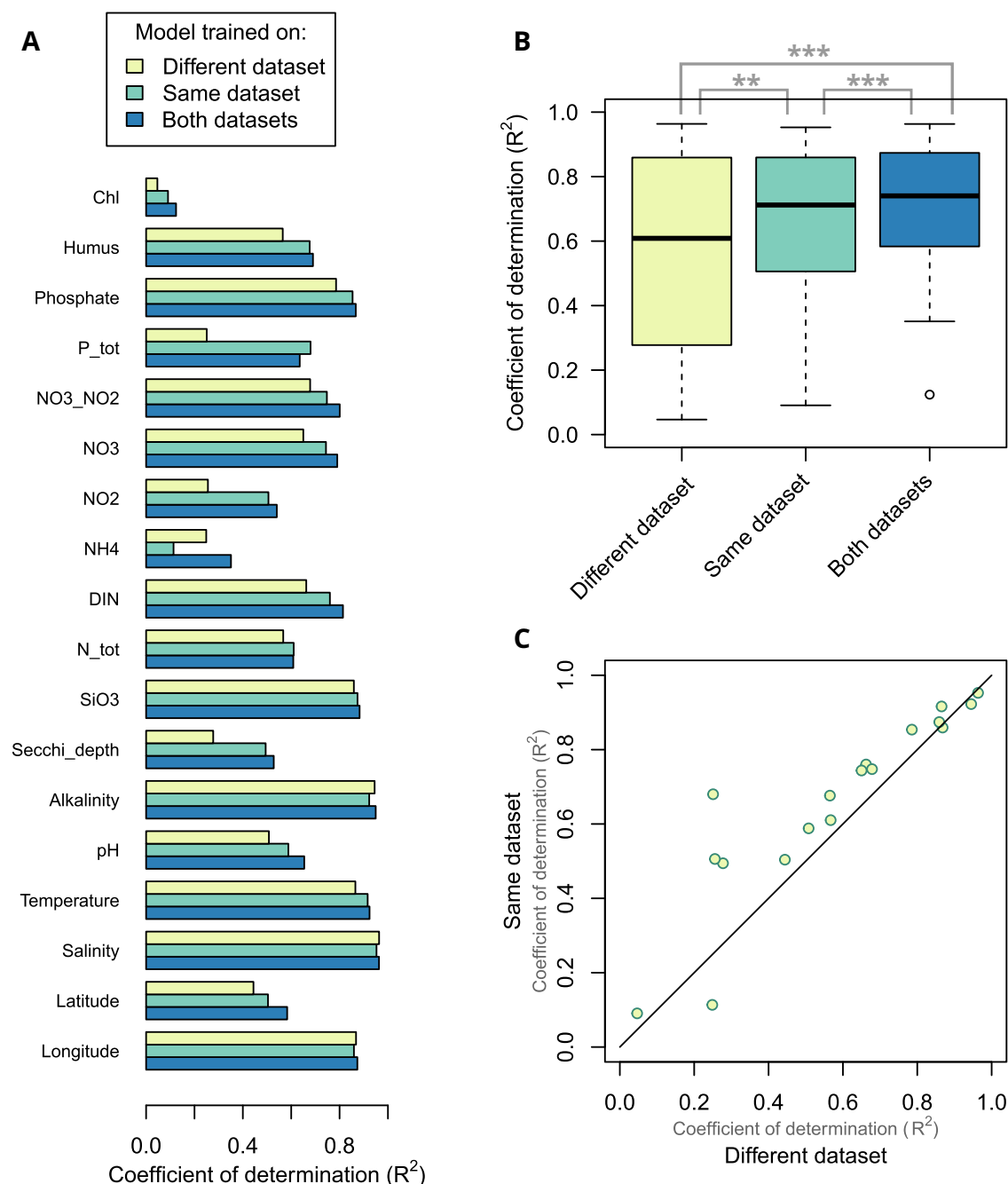


Fig. 5 | Transferability of machine learning models. Performance of Random Forest models based on 16S rRNA gene ASV profiles for predicting physicochemical parameters from metabarcoding data using different datasets. Parameters were predicted for the 2015–2017 dataset with models trained on (1) the 2019–2020 dataset (“different dataset”; yellow), (2) the 2015–2017 dataset (“same dataset”; green), or (3) both 2015–2017 and 2019–2020 datasets (“both datasets”; blue). Note that the “same dataset” and “both datasets” models used 10-fold cross-validation to

ensure that the same samples were not used for predictions and training, as was also done in the other analyses of this study. **A** coefficient of determination (R^2) for each predicted parameter and training dataset. **B** Boxplots of R^2 values for all parameters for the different training datasets (as shown in **A**). **C** Comparison of R^2 values obtained from training on the “same dataset” or “different dataset”. Asterisks indicate significance of Wilcoxon signed-rank tests: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Preparing the input data (features)

We removed ASVs annotated as metazoa (18S rRNA) and chloroplast and mitochondria (16S rRNA) from the dataset. Low prevalence ASVs that were present in $\leq 10\%$ of all samples were also excluded to limit the number of features. Low prevalence ASV removal was done separately for each type of analysis. As an example, for the ASV-level predictions of physicochemical parameters of Fig. 2, this resulted in 1770 out of 117,296 ASVs remaining for 16S and 891 out of 44,750 remaining for 18S rRNA. Feature tables at higher taxonomic classification were created from aggregated ASV counts at the

taxonomic levels: phylum, class, order, family, genus, and species for the 16S rRNA gene, and supergroup, division, subdivision, class, order, family, genus, and species for the 18S rRNA gene. We normalised the count tables by dividing ASV counts by the total ASV counts (excluding removed features) of each sample. Although there are methods intended to correct for the proportional nature of ASV data (e.g., centred logratio, isometric log-ratio), recent benchmarking studies indicate that proportion-based methods often perform comparably or even better for predictive modelling tasks^{59,60}.

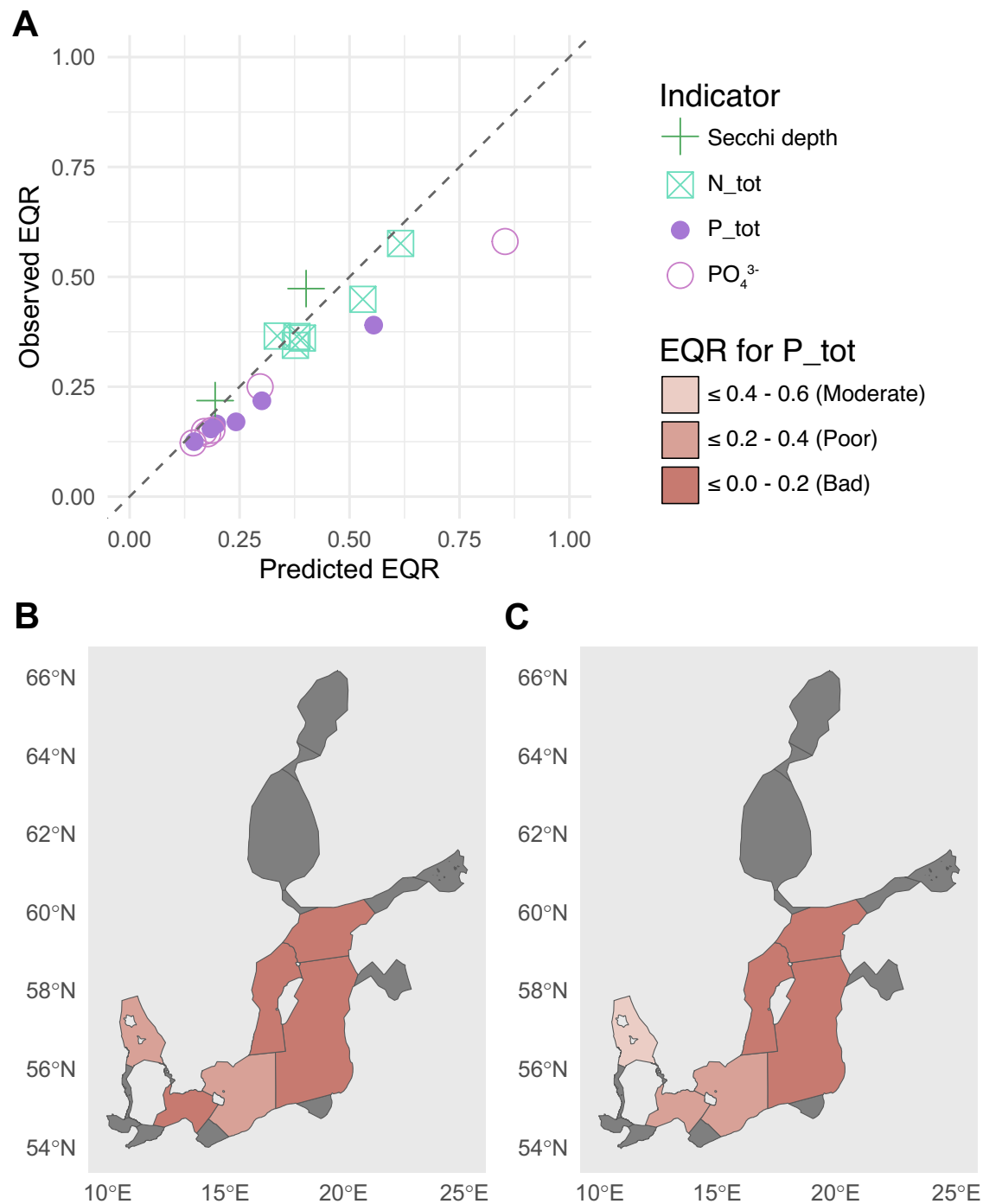


Fig. 6 | Calculating Ecological Quality Ratio (EQR) based on measured and predicted physicochemical parameters. Physicochemical parameters were predicted from 16S rRNA gene ASV profiles for the 2015–2017 dataset using the Random Forest algorithm trained with the 2019–2020 dataset (i.e., “different dataset”). **A** Observed and predicted EQR for water transparency (Secchi depth;

green cross), total nitrogen (N_tot; green square), total phosphorus (P_tot; purple filled circle), and dissolved inorganic phosphate (PO₄³⁻; purple open circle), in different basins (Kattegat, Arkona Basin, Bornholm Basin, Eastern Gotland Basin, Western Gotland Basin, Northern Baltic Proper). EQR based on **B** measured values of P_tot, and **C** predicted values of P_tot, with the colour indicating the EQR.

With autoencoders from the deep representation learning framework DeepMicro²⁵, we created deep encodings of the relative abundance data²⁵. Multiple encoders from the DeepMicro framework were used: Shallow Autoencoders (SAE), Deep Autoencoders (DAE), Variational Autoencoders (VAE), and Convolutional Autoencoders (CAE). All configurations listed in Supplementary Table S1 of the DeepMicro study²⁵ were used as dimensions for the networks.

Microscopy counts of zooplankton and phytoplankton were also used as features. We limited the features to taxa present in >10% of the samples

for training the machine learning models (just as for the metabarcoding counts). Taxonomic resolution was generally at the genus level, and in some cases species-level annotations were available.

Predicting physicochemical properties

We trained three model architectures, Random Forest, XGBoost, and TabPFN, to predict physicochemical parameters of seawater. These architectures were chosen for evaluation because they represent complementary paradigms for tabular regression. Random Forest provides a robust, low-

assumption baseline, XGBoost (gradient boosting) captures complex patterns through sequential boosting and often achieves state-of-the-art performance⁶¹, and TabPFN introduces a transformer-based approach that can learn from tabular data without extensive hyperparameter tuning⁶². Data from all years (2015–2017, 2019–2020) were included using 10-fold cross-validation. The folds for cross-validation were determined using ‘vfold_cv’ in the package rsample (v. 1.2.1)⁶³ with splitting based on the response variable.

Random Forest models were trained with the R package ranger (v. 0.16)⁶⁴ using default settings except for the number of trees, which was set to 2000. XGBoost models were trained using nested cross-validation with the R package caret (v7.0.1)⁶⁵. For each parameter (response variable), the data were split into 10 outer folds. Within each training outer fold, hyperparameter tuning was performed using 10-fold inner cross-validation. A grid search was conducted across a set of hyperparameters with varying nrounds (200–1000 in steps of 50), eta (0.005, 0.010, 0.100, 1.000), max_depth (1, 2, 4, 8, 16, 30), and subsample (0.4, 0.5, 0.6, 0.7). Parameters gamma, colsample_bytree, and min_child_weight were set to 0, 0.8, and 1, respectively. The best-performing hyperparameter combination identified through grid search was then used to train the final model. TabPFN was performed using the TabPFNRegressor class from the TabPFN Python library (v2.0.3)⁶².

Evaluating the predictions

To evaluate the models, we chose the coefficient of determination (R^2) as the metric. Other studies that utilised microbiome data to predict environmental parameters have also used $R^{21,43,66,67}$, which makes our results comparable to other efforts. The R^2 was calculated using the formula:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where SS_{res} and SS_{tot} are the residual and total sums of squares, respectively. y_i is the actual value of the parameter in sample i , $\hat{y}_i$ is the predicted value of the parameter in sample i , and $\bar{y}$ is the mean of the actual values for the parameter.

When comparing machine learning to correlating the relative abundances of ASVs with microscopy, for the latter approach, we used the R^2 for linear regression between the microscopy and metabarcoding data. In other words, for the correlation, $\hat{y}_i$ corresponds to the value predicted by the linear fit found with linear regression.

The classifiers, i.e., the models predicting presence/absence of phytoplankton and zooplankton taxa, were evaluated using F1-score. F1-score is a harmonic mean of precision and recall, and was calculated using the MLmetrics (v. 1.1.3) R package⁶⁸. The formula for F1-score is:

$$F1 = \frac{2TP}{2TP + FP + FN}$$

Where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. We counted presence as a positive result and absence as a negative result. All the significance tests comparing predictions with different methods or based on different data were based on the (two-sided paired) Wilcoxon signed-rank test.

Predicting microscopy identifications of plankton taxa

To predict the presence/absence of phytoplankton and zooplankton genera as determined by microscopy analysis, we tried three different approaches. First, we used physicochemical data as features to predict the microscopy-determined presence or absence of each genus in each sample. To address the tradeoff between the number of samples and parameters, we used only the parameters with data available for at least 80% of the samples with microscopy and metabarcoding measurements available. Thus, the excluded parameters were DOC, humus, Secchi depth, and alkalinity. For phytoplankton presence/absence predictions combined measurements of NO_3^- and NO_2^- were also excluded. Second, we used 16S rRNA gene or 18S rRNA gene ASVs to make the same predictions. Third, we used the presence/

absence of the same taxa obtained from 18S rRNA gene metabarcoding to predict the microscopy-determined presence/absence. For training, we converted the microscopy count table into binary data, with all counts >0 changed to 1. The Random Forest trained on this data returned values between zero and one. For evaluation, we converted the predicted values ≥ 0.5 to 1, and < 0.5 to 0, thus treating the random forest predictor as a binary classifier (while keeping the uncertain predictions in the original output). In addition to predicting presence/absence, the same setup was used to predict microscopy counts (individuals per volume unit) of the plankton genera (presented in Supplementary Fig. S2), running Random Forest in regression mode and using coefficient of determination (R^2) for the evaluations.

Training and predicting on different datasets

To test the transferability of the machine learning models, we trained and evaluated the models using data from different sampling years. Physicochemical parameters were predicted using 16S rRNA gene ASV relative abundances for the 2015–2017 dataset with models trained on either the 2019–2020 dataset (“different dataset”), the 2015–2017 dataset (“same dataset”), or both the 2015–2017 and the 2019–2020 datasets (“both datasets”). For evaluating the “same dataset” and “both datasets” models, 10-fold cross-validation was used. The same parameters were used to train the Random Forest models as those used for the prediction of physicochemical data.

Calculating environmental status indicators from predicted and measured values

We used the model trained with the “different dataset” approach to predict physicochemical parameters for the years 2015–2017, using the corresponding measured values as a reference. Ecological Quality Ratios (EQRs) were then calculated for the sampled basins, by computing the ratio of the actual condition to reference values⁴². We calculated EQRs for both predicted and measured values of dissolved inorganic phosphate, water transparency (Secchi depth), total nitrogen, and total phosphorus. The calculations were performed using the HELCOM Eutrophication Assessment Tool (HEAT), which was updated for the Third Holistic Assessment of the Baltic Sea by the Baltic Marine Environment Protection Commission (HELCOM) and is publicly available at <https://github.com/ices-tools-prod/HEAT>.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The eDNA datasets analysed were previously published^{35,36} and are available as raw data in the European Nucleotide Archive (ENA) under accessions PRJEB55296 and PRJEB84926, and as processed data in the Swedish ASV portal at <https://asv-portal.biodiversitydata.se> and at the Global Biodiversity Information Facility (GBIF) under the accessions PRJEB55296-16S, PRJEB84926-16S, PRJEB55296-18S, and PRJEB84926-18S. The physicochemical data and microscopy data were accessed through the National archive for Swedish oceanographic data (SHARK, <https://shark.smhi.se/hamta-data/>). Code, processed metabarcoding and environmental data are available at <https://github.com/EnvGen/EnvPredict>.

Received: 28 January 2026; Accepted: 28 May 2026;

Published online: 15 June 2026

References

1. Alneberg, J. et al. Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. *Commun. Biol.* **3**, 119 (2020).
2. Nowinski, B. & Moran, M. A. Niche dimensions of a marine bacterium are identified using invasion studies in coastal seawater. *Nat. Microbiol.* **6**, 524–532 (2021).

3. Ranheim Sveen, T., Hannula, S. E. & Bahram, M. Microbial regulation of feedbacks to ecosystem change. *Trends Microbiol.* **32**, 68–78 (2024).
4. Astudillo-García, C., Hermans, S. M., Stevenson, B., Buckley, H. L. & Lear, G. Microbial assemblages and bioindicators as proxies for ecosystem health status: potential and limitations. *Appl. Microbiol. Biotechnol.* **103**, 6407–6421 (2019).
5. Glasl, B. et al. Microbial indicators of environmental perturbations in coral reef ecosystems. *Microbiome* **7**, 94 (2019).
6. Bowman, J. S. Making sense of a scent-sensing metaphor for microbes and environmental predictions. *mSystems* **6**, e0099321 (2021).
7. SHARK. *SHARK - National Marine Environmental Monitoring of Phytoplankton in Sweden Since 1983* (The Swedish Meteorological and Hydrological Institute; 2024).
8. Reid, P. C., Colebrook, J. M., Matthews, J. B. L. & Aiken, J. The Continuous Plankton Recorder: concepts and history, from Plankton Indicator to undulating recorders. *Prog. Oceanogr.* **58**, 117–173 (2003).
9. Guan, W., Bao, M., Lou, X., Zhou, Z. & Yin, K. Monitoring, modeling and projection of harmful algal blooms in China. *Harmful Algae* **111**, 102164 (2022).
10. HELCOM. *Development of tools for assessment of eutrophication in the Baltic Sea* (HELCOM, 2006).
11. Hugerth, L. W. & Andersson, A. F. Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. *Front. Microbiol.* **8**, 1561 (2017).
12. Hugenholtz, P. & Tyson, G. W. Microbiology: metagenomics. *Nature* **455**, 481–483 (2008).
13. Sunagawa, S. et al. Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
14. Pinhassi, J. et al. Functional responses of key marine bacteria to environmental change - toward genetic counselling for coastal waters. *Front. Microbiol.* **13**, 869093 (2022).
15. Bourlat, S. J. et al. Genomics in marine monitoring: new opportunities for assessing marine health status. *Mar. Pollut. Bull.* **74**, 19–31 (2013).
16. McElhinney, J. M. W. R., Catacutan, M. K., Mawart, A., Hasan, A. & Dias, J. Interfacing machine learning and microbial omics: a promising means to address environmental challenges. *Front. Microbiol.* **13**, 851450 (2022).
17. Ghannam, R. B. & Techtman, S. M. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Comput. Struct. Biotechnol. J.* **19**, 1092–1107 (2021).
18. Smith, M. B. et al. Natural bacterial communities serve as quantitative geochemical biosensors. *MBio* **6**, e00326-15 (2015).
19. Machine learning modeling using microbiome data reveal microbial indicator for oil-contaminated groundwater. *J. Water Process Eng.* **53**, 103610 (2023).
20. JanBen, R. et al. Machine learning predicts the presence of 2,4,6-trinitrotoluene in sediments of a Baltic Sea munitions dumpsite using microbial community compositions. *Front. Microbiol.* **12**, 626048 (2021).
21. Sperlea, T. et al. Quantification of the covariation of lake microbiomes and environmental variables using a machine learning-based framework. *Mol. Ecol.* **30**, 2131–2144 (2021).
22. Thompson, J., Johansen, R., Dunbar, J. & Munsky, B. Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition. *PLoS One* **14**, e0215502 (2019).
23. Hu, H. et al. Supervised machine learning improves general applicability of eDNA metabarcoding for reservoir health monitoring. *Water Res.* **246**, 120686 (2023).
24. Hilderbrand, R. H. et al. Microbial communities can predict the ecological condition of headwater streams. *PLoS One* **15**, e0236932 (2020).
25. Oh, M. & Zhang, L. DeepMicro: deep representation learning for disease prediction based on microbiome data. *Sci. Rep.* **10**, 6026 (2020).
26. Lozupone, C. A. & Knight, R. Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. USA.* **104**, 11436–11440 (2007).
27. Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
28. Philippot, L., Griffiths, B. S. & Langenheder, S. Microbial community resilience across ecosystems and multiple disturbances. *Microbiol. Mol. Biol. Rev.* **85**, e00026-20 (2021).
29. Hernández Medina, R. et al. Machine learning and deep learning applications in microbiome research. *ISME Commun.* **2**, 98 (2022).
30. Cordier, T. et al. Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning. *Environ. Sci. Technol.* **51**, 9118–9126 (2017).
31. Cordier, T. et al. Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Mol. Ecol. Resour.* **18**, 1381–1391 (2018).
32. Frühe, L. et al. Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Mol. Ecol.* **30**, 2988–3006 (2021).
33. Bowman, J. S. et al. Bacterial community segmentation facilitates the prediction of ecosystem function along the coast of the western Antarctic Peninsula. *ISME J.* **11**, 1460–1471 (2017).
34. Connors, E. et al. Microbial community composition predicts bacterial production across ocean ecosystems. *ISME J.* **18**, wrae158 (2024).
35. Latz, M. A. C. et al. A comprehensive dataset on spatiotemporal variation of microbial plankton communities in the Baltic Sea. *Sci. Data* **11**, 1–9 (2024).
36. Jurdzinski, K. T. et al. Distinct bacterial and protist plankton diversity dynamics uncovered through DNA-based monitoring in the Baltic Sea area. *bioRxiv* <https://www.biorxiv.org/content/10.1101/2024.08.14.607742v1> (2024).
37. HELCOM Indicator Manual - 2025 update. *Baltic Sea Environment Proceedings n° 175* (HELCOM, 2025).
38. Reygondeau, G. et al. Dynamic biogeochemical provinces in the global ocean: DYNAMIC BIOGEOCHEMICAL PROVINCES. *Glob. Biogeochem. Cycles* **27**, 1046–1058 (2013).
39. Brun, P. et al. Ecological niches of open ocean phytoplankton taxa: niches of open ocean phytoplankton. *Limnol. Oceanogr.* **60**, 1020–1038 (2015).
40. Melo-Merino, S. M., Reyes-Bonilla, H. & Lira-Noriega, A. Ecological niche models and species distribution models in marine environments: a literature review and spatial analysis of evidence. *Ecol. Model.* **415**, 108837 (2020).
41. Backer, H. et al. HELCOM Baltic Sea Action Plan-a regional programme of measures for the marine environment based on the Ecosystem Approach. *Mar. Pollut. Bull.* **60**, 642–649 (2010).
42. Andersen, J. H. et al. Getting the measure of eutrophication in the Baltic Sea: towards improved assessment principles and methods. *Biogeochemistry* **106**, 137–156 (2011).
43. Zschaubitz, E. et al. A benchmark analysis of feature selection and machine learning methods for environmental metabarcoding datasets. *Comput. Struct. Biotechnol. J.* **27**, 1636–1647 (2025).
44. Herlemann, D. P. et al. Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J.* **5**, 1571–1579 (2011).
45. Raes, J., Letunic, I., Yamada, T., Jensen, L. J. & Bork, P. Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol. Syst. Biol.* **7**, 473 (2011).
46. Blauw, A. N., Benincà, E. Laane, R. W. P. M., Greenwood, N. & Huisman, J. Dancing with the tides: fluctuations of coastal phytoplankton orchestrated by different oscillatory modes of the tidal cycle. *PLoS One* **7**, e49319 (2012).

47. Blauw, A. N., Benincà, E., Laane, R.W.P.M., Greenwood, N., Huisman, J. Predictability and environmental drivers of chlorophyll fluctuations vary across different time scales and regions of the North Sea. *Prog. Oceanogr.* **161**, 1–18 (2018).
48. Zhu, Y. et al. Development of analytical methods for ammonium determination in seawater over the last two decades. *Trends Anal. Chem.* **119**, 115627 (2019).
49. HELCOM. Manual for marine monitoring in the COMBINE programme of HELCOM. Vol. 2017. (2021).
50. Utermöhl, H. Zur Vervollkommenung der quantitativen Phytoplankton-Methodik. *Mitt. Int. Ver. Theor. Angew. Limnol.* **9**, 1–38 (1958).
51. Hugerth, L. W. et al. DegePrime, a program for degenerate primer design for broad-taxonomic-range PCR in microbial ecology studies. *Appl. Environ. Microbiol.* **80**, 5116–5123 (2014).
52. Balzano, S., Abs, E. & Leterme, S. C. Protist diversity along a salinity gradient in a coastal lagoon. *Aquat. Micro. Ecol.* **74**, 263–277 (2015).
53. Straub, D. et al. Interpretations of environmental microbial community studies are biased by the selected 16S rRNA (gene) amplicon sequencing pipeline. *Front. Microbiol.* **11**, 550420 (2020).
54. Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
55. Lundin, D. & Andersson, A. *SBDI Sativa Curated 16S GTDB Database* (Linnéuniversitetet, 2025).
56. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
57. Guillou, L. et al. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41**, D597–D604 (2013).
58. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
59. Yerke, A., Fry Brumit, D. & Fodor, A. A. Proportion-based normalizations outperform compositional data transformations in machine learning applications. *Microbiome* **12**, 45 (2024).
60. Garach Vélez, I., Ortuño Guzmán, F.M., Rojas Ruiz, I. & Herrera Maldonado, L.J. Exploring the role of normalization and feature selection in microbiome disease classification pipelines. *Gigascience* **14**, gfa096 (2025).
61. Chen, T. & Guestrin, C. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (ACM, 2016).
62. Hollmann, N. et al. Accurate predictions on small data with a tabular foundation model. *Nature* **637**, 319–326 (2025).
63. Frick, H. et al. rsample: General Resampling Infrastructure. R package, <https://rsample.tidymodels.org> (2025).
64. Wright, M.N. & Ziegler, A. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77**, 1–17 (2017).
65. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Soft* **28**, 1–26 (2008).
66. Dutta, A. et al. Machine learning predicts biogeochemistry from microbial community structure in a complex model system. *Microbiol. Spectr.* **10**, e0190921 (2022).
67. Sperlea, T. et al. The relationship between land cover and microbial community composition in European lakes. *Sci. Total Environ.* **825**, 153732 (2022).
68. Yan, Y. *MLmetrics: Machine Learning Evaluation Metrics*. CRAN: Contributed Packages (The R Foundation, 2015).

Acknowledgements

The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. The authors are grateful for the constructive comments provided by the three reviewers.

Author contributions

All authors: Conceptualisation; Data Curation; Methodology; Writing (review & editing). E.B.: Formal analysis; Visualisation; Writing (original draft). K.G.: Formal analysis; Visualisation; Writing (original draft). K.T.J.: Formal analysis; Visualisation; Writing (original draft). L.F.D.: Formal analysis; Writing (original draft). F.J.: Formal analysis. B.K.: Writing (review). A.F.A.: Formal analysis; Visualisation; Writing (original draft).

Funding

The study was funded by grants from the Swedish Research Council (Vetenskapsrådet; grant number 2021-05563), the Swedish Research Council for Sustainable Development (Svenska Forskningsrådet FORMAS; grant number 2022-01515), the European Union (Marie Skłodowska-Curie Actions MSCA; grant number 101064544), and a Sverker Lerheden donation. Open access funding provided by Royal Institute of Technology.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s43247-026-03715-5>.

Correspondence and requests for materials should be addressed to Anders F. Andersson.

Peer review information *Communications Earth & Environment* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Alice Drinkwater, Marisa McDonald. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026