Quantifying the information content of lake microbiomes using a machine learning-based framework

Theodor Sperlea
Philipps-Universitat Marburg
Nico Kreuder
Universitat Duisburg-Essen
Daniela Beisser
Universitat Duisburg-Essen
Georges Hattab
Philipps-Universitat Marburg
Jens Boenigk
Universitat Duisburg-Essen
Dominik Heider (dominik.heider@uni-marburg.de)
Philipps-Universität Marburg https://orcid.org/0000-0002-3108-8311

Research

Keywords: microbial ecology, machine learning, bioindicators, lake ecosystems

Posted Date: August 6th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-52629/v1

Version of Record: A version of this preprint was published on March 7th, 2021. See the published version at https://doi.org/10.1111/mec.15872.

RESEARCH

Quantifying the information content of lake microbiomes using a machine learning-based framework

Theodor Sperlea[1], Nico Kreuder[2], Daniela Beisser[2], Georges Hattab[1], Jens Boenigk[2] and Dominik Heider[1*]

E-mail adresses: TS - theodor.sperlea@staff.uni-marburg.de, NK - nico.kreuder@stud.uni-due.de, DB - daniela.beisser@uni-due.de, GH - Georges.hattab@uni-marburg.de, JB - jens.boenigk@uni-due.de, DH - dominik.heider@uni-marburg.de (corresponding author)

Adresses: [1]Faculty of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Str. 6, D-35032 Marburg (Lahn), Germany; [2]Department of Biodiversity, Center for Water and Environmental Research, University of Duisburg-Essen, Hans-Meerwein-Str. 6, D-45141 Essen, Germany

*Correspondence: dominik.heider@uni-marburg.de
[1]Faculty of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Str. 6, D-35032 Marburg (Lahn), Germany
Full list of author information is available at the end of the article

Abstract

Background: Bacteria and microbial eukaryotes occupy a wide range of ecological niches and are essential for the functioning of ecosystems. The advent of next-generation sequencing methods enabled the study of environmental microbial community compositions. Yet, many questions regarding the stability and functioning of environmental microbiomes remain open.

Results: In the current study, we present a methodological framework to quantify the information shared between the microbial community of a habitat and the abiotic parameters of this habitat. It is built on theoretical considerations of systems ecology and makes use of state-of-the-art machine learning techniques. It can also be used to identify bioindicators. We apply the framework to a dataset containing operational taxonomic units (OTUs) as well as more than twenty physico-chemical and geographic parameters measured in a large-scale survey of European lakes. While a large part of variation (up to 61%) in many physico-chemical parameters can be explained by microbial community composition, some of the examined parameters only share little information with the microbiome. Moreover, we have identified OTUs that act as 'multi-task' bioindicators that could be potential candidates for lake water monitoring schemes.

Conclusions: This study demonstrates the benefits of machine learning approaches in microbial ecology. Our results represent, for the first time, a quantification of information shared between the lake microbiome and a wide array of ecosystem parameters. Building on the results and methodology presented here, it will be possible to identify microbial taxa and processes central for the functioning and stability of lake ecosystems.

Keywords: microbial ecology; machine learning; bioindicators; lake ecosystems

Introduction

Anthropogenic changes to the environment are threatening the stability of ecosystems globally and contribute to unprecedented rates of species extinction with catastrophic consequences for life as we know it [1–6]. To mitigate the destabilization and the collapse of ecosystems, we need a more refined understanding of how they function. Systems ecology offers a paradigm that describes ecosystems as dynamic and complex networks of interactions both among organisms as well as between the biotic and abiotic aspects of an ecosystem [7–10].

Through interactions and the flow of energy and nutrients, different parts of an ecosystem share information. This is not limited to direct interactions as, for example, the number of predators in an ecosystem has both an effect on the number of prey as well as on the plants eaten by the prey [11, 12]. The network structure of ecosystems and its dynamic adaptability to changes in the environment makes it possible to identify bioindicators, i.e., organisms whose presence and prevalence can be used to estimate other parameters of the ecosystem [13, 14].

Bioindicators are used in biosphere-based ecosystem monitoring schemes such as the ones implemented in European countries under the Water Framework Directive [15, 16] but also hold insights into the autecology of organisms (i.e., their specific ecological needs and actions) as well as the functioning of an ecosystem as a whole

[17]. This is the case since organisms will only emerge as indicative for parameters they react to directly (because of their ecological niche) or indirectly (since they interact closely with organisms that are, in turn, reactive to changes in the respective parameter). Due to their functional diversity, high growth rates, large population sizes, and high surface-to-volume ratio, bacteria and microeukaryotes are very responsive to environmental changes and represent optimal bioindicators [14, 18–20].

The advent of next-generation sequencing (NGS) has greatly facilitated the use of microbial bioindicators. Firstly, it made it possible to identify organisms based on their genetic makeup instead of visual features [20, 21]. Secondly, techniques such as amplicon sequencing have made it feasible to capture microbial community compositions present in environmental samples [22]. As different microorganisms exhibit different reactions to changes in a parameter, and these reactions are modulated by other microorganisms, the whole microbial community composition will be more indicative of the status of the ecosystem than a list of bioindicator species.

However, while being rather intuitive, the systems ecology paradigm also exposes theoretical and methodical obstacles for the study of microbial ecology. For example, the assumption of variable independence, which is a requirement for many statistical approaches, does not hold for ecosystem parameters or processes. Similarly, in a system, processes are influencing and modulating each other, rendering the distinction between direct and indirect interactions hard or even infeasible [8]. This is especially the case for microbial ecology, where interaction networks are hard to measure and validate [13, 23–25] and the distinction between indirect and direct interactions is an open question [26, 27]. Indeed, many studies prove a high relevance of indirect effects [28, 29].

Additional issues for the study of microbial ecology stem from the high sparsity and very high dimensionality of OTU tables [27, 30]. With a number of samples vastly lower than the number of features, regression is ill-defined and the correction of the R^2 value for sample dimensionality is impossible. Usually, both the collection of more data as well as very stringent feature selection are suggested to counteract this. Both measures, however, are only of limited use in microbial ecology, as sampling and sequencing remain expensive and the high number of different microorganisms is a non-reducible property of the study object.

In this study, we developed methodological tools to study microbial ecology in the context of systems ecology while acknowledging the aforementioned theoretical obstacles. Our main contribution is a machine learning-based framework for the quantification of the information shared between the microbiome and a total of 25 physico-chemical and positional (i.e., GPS coordinates and altitude) parameters of an ecosystem. It builds upon a wealth of studies that elucidate the role of the microbiome in ecology using machine learning [19, 31–39]. In our framework, a model learns a projection of the microbial prevalence space to a single dimension for each of the parameters, which makes it able to handle the extremely high dimensionality of amplicon-based microbiome datasets. The coefficient of correlation R^2 between the projected microbial community composition and the measured parameter is, then, used as a metric of shared information. This corresponds to the co-variation of the abiotic parameter and the whole microbiome, which is intuitively interpretable.

We applied this framework to a dataset from a large-scale survey of European lakes [40–42]. Lakes are considered as sentinels of ecosystem change at different temporal and geographical scales [43, 44]. This is, in part, because lakes aggregate water from their catchments, and with it, pollutants and high nutrient concentrations. Furthermore, lakes are also directly affected by various anthropogenic stressors, such as overfishing, eutrophication, climate change, and invasive species [45, 46].

The use of non-linear ensemble models facilitated a dimensionality reduction of up to 6 orders of magnitude while retaining important non-linear relationships in the amplicon dataset. Comparing two feature selection methods that were motivated by ecology, we found that filtering for bioindicators leads to a favorable behavior of the framework. Analyzing the operational taxonomic units (OTUs) identified as bioindicators in the feature selection step, we identified bacteria and microbial eukaryotes indicative of multiple lake parameters, which support the notion of high inter-dependency between ecological parameters.

At the time of writing and to our knowledge, we provide the first large-scale, sequencing-based analysis of the potential of the full microbial community composition as an indicator for physico-chemical parameters in lake ecosystems. To that end, we report a comprehensive quantification of the co-variation of the complete microbiome concerning these parameters. Our results highlight the advantages of machine learning methods for the study of microbial ecology in a systems ecology paradigm. Furthermore, they underscore the importance of including bacteria and microeukaryotes at the species or OTU level into ecological monitoring schemes. We believe this work paves the way for future endeavors to better uncover the functional workings of ecosystems.

Methods

Sample collection

Sampling was part of a pan-European study conducted in August 2012 (eukaryotic sequences are published in [40]; NCBI Bioproject PRJNA414052, prokaryotic sequences are published and described in [47] and [42]; NCBI Bioproject PRJNA559862). To analyze the effects of bio-geo-chemical factors on bacterial and protist freshwater communities on a large scale, 280 lakes were sampled, covering a broad latitudinal gradient ranging from Spain to the South of Scandinavia and altitudes from sea level to 3110 m.a.s.l. The samples were taken in daylight from the shore of each lake or pond collecting epilimnial water up to 0.5 m depth. Sampling details and information on measured physico-chemical and geographic factors can be found in [40]. For DNA analyses filtered samples were air-dried and frozen in liquid nitrogen (Cryoshippers) and stored at -80 °C until further processing.

DNA extraction and sequencing

Genomic DNA was extracted using the my-Budget DNA Mini Kit (Bio-Budget Technologies GmbH, Krefeld, Germany) following the protocol of the manufacturer and modifications after [40]. Bacterial amplicon sequencing targeted the V2-V3 region of the 16S rRNA gene, eukaryotic amplicon sequencing targeted the V9 region of the 18S, and the ITS1 gene in the SSU genomic region. Samples were commercially sequenced (Fasteris, Geneva, Switzerland) using paired-end Illumina HiSeq 2500 sequencing in the 'rapid run' mode to generate 2 x 300 bp reads. For details, please see [40, 47] and [42].

DNA extraction and sequencing

Adapter removal, quality trimming, and demultiplexing using index sequences were performed by the sequencing company (Fasteris). Base quality of raw sequence reads was rechecked using the FastQC software (v0.11.5; [48]) and reads with an average Phred quality score below 25 or with at least one base with a Phred quality score below 15 were removed using PRINSEQ-lite (v0.20.4; [49]). The paired-end reads were assembled and quality filtered with the tool PANDASeq (v2.10; [50]). Reads with uncalled bases, an assembly quality score below 0.9, a read overlap below 20 bases, or a base with a recalculated Phread-score below 1 were discarded. Assembled sequences were dereplicated and chimeras identified using UCHIME (usearch v7.0.1090; [51]). Additionally, a split-sample filtering protocol (AmpliconDuo; [52]) was used to discard sequences that were not found in both technical replicates (A and B variant). Remaining sequences were clustered using SWARM (v2.2.2; [53]). The eukaryotic representative sequences were further clustered by identical V9 sequences (V9_Clust.R; [54]). The taxonomic assignment of the eukaryotic sequences was performed by searching the NCBI database using BLAST (BLAST + v2.7.1; NCBI nt sequences from Dec 5, 2017). For the prokaryotic sequences SILVA (SILVA SSURef release 132) was used.

Data preparation

Values for temperature (T) and conductivity (LF), measured in field in triplicates, were averaged for each sample. For the analyses at different taxonomic levels, for each taxon at each taxonomic level, OTU counts belonging to this taxon were aggregated. OTUs missing a taxonomic annotation at a taxonomic level were not counted.

To circumvent the problem of missing data in data analysis, two sub-datasets were created, namely the *all_samples* and *all_features* sub-datasets. The *all_samples* sub-dataset contains the parameters measured in the field (altitude, GPS coordinates, pH, conductivity, temperature, and time of sampling) and OTUs for 241 lakes. An additional set of 21 physico-chemical parameters had been measured for a subset of 47 lakes. Excluding the positional parameters and the time measurement, lakes with the extended parameter set and the corresponding OTUs constitute the *all_features* dataset.

Outliers in the lake parameters were defined as data points falling outside of a range of 1.5 times the interquartile range below the first or above the third quartile (as calculated using the R function `boxplot.stats()`). Samples that contain at least one outlier in any of the lake water parameters relevant for the sub-dataset were excluded from further analysis, leading to 201 and 42 samples in the *all_samples* and *all_features* dataset, respectively (see supplementary table 1 for a list of lakes present in the sub-datasets).

Machine learning and data analysis

To quantify the amount of information shared by the microbial community and a lake parameter, a machine learning model was trained to approximate this parameter based on the OTU table or taxonomically aggregated prevalence table. Models were trained using either 10-fold cross-validation (*all_samples* sub-dataset) or leave-one-out cross-validation (*all_features* sub-dataset) to avoid over-fitting at low sample

numbers. In our framework, the model prediction is seen as the projection of the microbial community composition to one dimension that is comparable to the lake parameter in question. As metric for shared information, the coefficient of determination R^2 was used. A higher R^2 value here indicates that the used model is better able to perform the dimensionality reduction while retaining information shared between the microbial community composition and the abiotic factor in question. However, the R^2 is not expected to assume values close to 1 as the microbiome is not expected to fully share information with any of the lake parameters.

Pre-model training feature selection was performed in each fold using either a fast correlation-based filter (FCBF) [55] or the `multipatt()` function (IndVal method, 999 random permutations) from the R package indicspecies (v1.7.9, [56]). The choice of the former was motivated by the widespread use of correlation networks as proxies for microbial interactions [57]. In these, nodes represent species and are connected with an edge if their prevalence correlates across a range of samples. Along these lines, FCBF groups OTUs or taxa that are neighbors in a correlation network into syntaxa, i.e., groups of organisms that act as one unit in environmental changes [58]. Pearson correlation and a correlation coefficient cutoff of 0.6 were used for this filter. For the IndVal analysis, samples were separated by tertiles of the parameter in question and OTU and taxon occurence numbers were standardized using the Hellinger transformation to decrease the influence of highly abundant OTUs [59].

A total of 7 machine learning models from the R package caret (v6.0.86, [60]) were used in this study with default parameters: random forest (rf), stochastic gradient boosting (gbm), extreme gradient boosting (xgbTree), support vector machines with linear and radial kernel (svmLinear, svmRadial), generalized linear model (glmnet), and k-nearest neighbors (knn). OTU counts were centered and scaled using the R function `scale()` before training. Confidence intervals were determined by 1,000x re-sampling by bootstrapping of predicted and measured values. Lake parameters were clustered according to their Pearson correlation using the `hclust()` function from the R package stats (v4.0.1). Variable importances were extracted from rf models using the `varImp()` function from the R package caret (v6.0.86, [60]) and averaged over the training folds. The `ttest()` function was used to assess significance in difference between the variable importances.

Results

Nonlinear models capture relevant patterns in microbial community composition

While the most widespread use of regression models is predictive in nature, they can also be seen as learning and approximating a dimensionality reduction function that projects the input feature space to an one-dimensional target space. Based on this notion, we developed a framework to quantify the information shared between an ecosystem's microbial community composition and an abiotic parameter. The model is trained to, in a sense, model the interactions between the microorganisms as well as the interactions between the microbiome and the target parameter. As metric of shared information, we used the R^2 metric between the dimensionality-reduced "prediction" and the measured parameter values as this can straightforwardly be interpreted as co-variation between the microbiome and the parameter. In a first implementation of the framework, we employed a fast correlation-based filter (FCBF)

to reduce the dimensionality before machine learning, and trained machine learning models on the *all_samples* dataset (and, therefore, only for a reduced number of parameters) using a 10-fold cross-validation evaluation scheme. This choice of feature selection method was motivated by the use of correlation networks in microbial ecology [57].

To test the hypothesis that non-linear, as well as linear, relationships between microorganisms are important for their reaction to environmental changes, we compared the performance of different regression models. Higher R^2 values indicate a higher propensity of the model to capture relevant patterns in the microbial community composition. The maximum attainable R^2 is determined by the information shared between the microbiome and the target lake parameter; a value of 1 would only be possible if no other factors influence the microbiome or the parameter level.

In our results, models that can learn both linear and non-linear relationships between features (Random Forest and xgbTree) outperform other models, supporting the notion that complex relationships are present between microbial community structure and ecosystem parameters (see figure 1). Based on these results, we focus the presentation and discussion of further results in this paper to Random Forest models.

Additionally, the results show that, consistently for all models and parameters, lower (i.e., more detailed) levels of taxonomic hierarchy share more information with the lake parameters. Nevertheless, FCBF does not reduce the dimensionality of the microbial community composition sufficiently to enable the training of regression models for all levels. Especially at the OTU level, around 89% of the initial features were still left after feature selection (table 1). This disproportion between sample number and feature space dimensionality made the application of the framework impossible for some of the parameters (see missing values in figure 1).

Indicator species analysis as feature selection for microbiome dimensionality reduction

As an alternative filtering method, we employed the IndVal method [61]. This calculates a composite indicator value based on the specificity and fidelity of a given species concerning a predefined set of sites. Its use in the identification of bioindicators suggests that it should be able to select OTUs or taxa that share information with a given lake parameter. Applying IndVal as a feature selection method in our framework to the *all_features* dataset resulted in more stringent models for OTUs and taxa (table 1). Comparing the results of the framework developed earlier using either FCBF or IndVal as feature selection method shows that, for some parameters and taxonomic levels, using IndVal leads to better results, albeit not significantly (see Fig. 2A). Furthermore, for some combinations of taxonomic levels and parameters, the use of FCBF outperformed the use of IndVal. On the other hand, some FCBF runs were not computable (highlighted by the missing values in Fig. 2A) this was never the case for IndVal runs. Finally, as the models trained using IndVal selected features are more sparse, this filter method is, in general, preferable to FCBF for microbial ecology. Based on these results, we conclude that most of the taxa or OTUs sharing information with the respective abiotic factor are contained in the IndVal selection.

Information shared between different taxonomic levels and lake parameters

Random Forest models trained with IndVal-selected features at the OTU level lead to median R^2 values of more than 0.3 for more than half of the physico-chemical parameters present in the *all_features* dataset (figure 2B). As seen for FCBF (see figure 1), lower taxonomic levels share more information with the physico-chemical parameters, supporting the notion that the diversity of niches occupied by OTUs belonging to the same higher-order taxa has an ecological significance.

To test the hypothesis that different levels of microbial taxonomy interact with physico-chemical parameters in different ways, we aggregated the IndVals over different levels of taxonomy and used this data to train machine learning models (lines labelled "all" in figure 2B). These models do not significantly outperform the models trained on OTU prevalence tables. Therefore, we conclude that different taxonomic levels do not contribute to ecologically relevant patterns not already present at the OTU level.

Analysis of microbial multi-task bioindicators

The results presented to this point support the use of the IndVal to identify ecologically relevant OTUs from amplicon sequencing data. After Bonferroni correction, the numbers of bioindicators for different parameters ranged over four orders of magnitude (see table 2 and supplementary table 3). We analyzed these bioindicator OTUs by focusing on multi-task bioindicators, i.e., OTUs that emerged as indicative for multiple physico-chemical parameters and might, therefore, act as general indicators of lake ecosystem status.

All of the bioindicators indicative of more than 7 parameters are annotated as Bacteria (see table 3 and figure 3A) except for two OTUs that are annotated as chloroplasts of the green algae *Phacotus lenticularis*. This organism has been described as a bioindicator for freshwater ecosystems before [62, 63]. Most of the other OTUs are from the Phyla Bacteroidetes and Proteobacteria. Many of the lowest distinct taxa we identified have previously been discussed as bioindicators for general ecosystem quality (Ignavibacteriales [36], *Limnobacter* [64], and Sandaracinaceae [65]), certain environmental parameters (*Opitutus* [17, 66], Alcaligenaceae [67], *Novosphingobium* [68, 69], and NS11-12 marine group [70, 71]), and human interference/impact/pollution (*Actibacter* [72], Fluviicola [73], and SC-I-84 [74]). However, not all of these taxa have previously been identified in lake ecosystems, and most of the OTUs among these bacterial multi-task bioindicators are assigned to taxa originally isolated from soil ecosystems (see table 3).

The multi-task bioindicators among the eukaryotes are, at most, indicative for five parameters. Among the 32 OTUs that are indicative for more than two parameters, six are annotated as Ciliophora or Chlorophyta. These classes are ubiquitous in lakes [34, 75, 76], contain many species that inhabit specific ecological niches and have been used as bioindicators [77–79]. Similarly, many of the eukaryotic multi-task OTUs identified here belong to genera that have been described as ubiquitous in freshwater ecosystems (e.g., Chytridiomycota [80], *Desmodesmus* [81] or *Gymnodinium* [82]). However, most of the species we identified have, to our knowledge, not yet been described as bioindicators at lower taxonomic levels.

Based on our finding that bacterial OTUs can be indicative for more than five parameters at the same time, we speculated that bacteria are, in general, better

suited as bioindicators than eukaryotes. To test this hypothesis, we extracted feature importance values from the Random Forest models used in the machine learning framework described earlier. Comparing the feature importances assigned to bacterial OTUs with those from eukaryotic OTUs using t-tests lead to significant p-values for the lake parameters dissolved organic carbon (DOC), dissolved reactive silica (DRSi), hydrogen (H), potassium (K), ammonium (NH_4), and temperature (T) (see supplementary table 5). This suggests that bacteria and eukaryotes play different roles in lake ecosystems. However, for the other parameters, we observed no significant difference in feature importances between bacterial and eukaryotic OTUs. This supports the notion that in an ecosystem, groups of interacting organisms cannot be seen as independent with regards to their ecological function. The network structure of ecosystems is also supported by our finding that physico-chemical parameters both the bacterial and eukaryotic multi-task bioindicators respond to are distributed in accordance to the Pearson correlation between the physico-chemical parameters (see figure 3).

Discussion

To arrive at a fuller image of the functioning of ecosystems, methodological approaches and theoretical paradigms have to be integrated. In this study, we combined bioindicator analysis, machine learning techniques, and the systems ecology paradigm to quantify the information shared between the microbiome and physico-chemical parameters of lake ecosystems. We present a framework that acknowledges the technical obstacles presented by ecological data in general and molecular microbial ecology datasets in particular.

In this framework, we compared different machine learning models and found that ensembles of decision trees (such as Random Forest and xgbTree models) were best able to project the microbiome to a one-dimensional space (figure 1). This is most likely due to their ability to learn highly non-linear relationships and cope with large feature spaces [83]. Additionally, ensembles of decision trees are, in principle, capable of learning from data for which the independence assumption does not hold [83]. We were also able to show that while using FCBF and IndVal as feature selection methods leads to comparable results, the IndVal method results in sparser models that allow the use of the framework even for extremely high-dimensional datasets at low levels of the taxonomy (see table 1). While IndVal has been used for molecular datasets collected, for example, at the Great Barrier Reef [37], this study is first in applying it to molecular data in the context of lake ecology.

Applying our framework to a microbial ecology dataset collected in a large-scale survey of European lakes, the results suggest that lower levels of microbial taxonomy share a higher level of information with lake parameters (figures 1 and 2B). This supports the hypothesis that OTUs from the same species can react to environmental change differently and might, therefore, inhabit different ecological niches [44, 47, 84–86]. Furthermore, our results show that, for most physico-chemical parameters, higher levels of taxonomy do not contain information not already present on the OTU level (see figure 3B), which is in contrast to the findings of others [87].

In the analysis of bioindicator OTUs identified in this study, we focused on multi-task bioindicators. Among the OTUs identified as bioindicators for more than 7

abiotic lake parameters, most have been taxonomically assigned to uncultured soil bacteria (see figure 3A and table 3). Similarly, most high-ranked eukaryotic multi-task bioindicators (see figure 3B and supplementary table 4) have been first identified in freshwater biomes, but not necessarily been found in lake samples yet. As the dataset analyzed here stems from lakes, this is most probably an artifact of imprecise taxonomic annotation [88], but might also point to the diversity of ecological niches inhabited by bacterial subspecies grouped into one OTU or species [44]. Although soil and freshwater microbial community compositions differ significantly [75], microorganisms can enter lakes from soil ecosystems directly or, e.g., via rivers that feed the lake. The emergence of *Phacotus lenticularis* as a multi-task organism in both groups of organisms (see table 3 and figure 3B) underscores its role as a bioindicator.

Recent studies have argued for differences in ecological function between bacteria and microbial eukaryotes in lake ecosystems [42, 89, 90]. More specifically, it has been argued that bacteria are more responsive to environmental changes than eukaryotes [14, 18, 20, 42]. This is supported by our result that bacteria that are multi-task bioindicators can be indicative of more lake parameters than eukaryotic multi-task bioindicators (see figure 3). We also found significant differences in the variable importances assigned to bacterial and eukaryotic OTUs by the Random Forest model used in the framework for some lake parameters (see supplementary table 5). However, this is not the case for all parameters. We see two main reasons for this: Firstly, at the Domain level, aggregated prevalence numbers do not share much information with lake parameters (see figure 1, 2B). Secondly, the interactions between organisms lead to indirect effects that would inhibit such a simple distinction between eukaryotes and bacteria. In the context of systems ecology, we would not expect groups of organisms to be independent in a manner relevant to this question.

Unsurprisingly, the parameters these multi-task bioindicators are indicative of show a high degree of correlation (see figure 3). Aside from underscoring the need for functional diversity in bioindicators if aiming at covering all parameters, this indicates that there are "main factors" among lake parameters that influence a high number of other parameters strongly. Altitude has been described as one of them, as it is directly or indirectly related to, among others, temperature, radiation, salinity, conductivity, and nutrient concentration [91]. This is the case because lakes in the lowland mainly arise from rivers, which have their source in mountain chains and get enriched with nutrients during their courses. Especially for eukaryotic multi-task bioindicators, our analyses suggest that temperature, conductivity (as measured in the field, displayed in this study under the label "LF"), and pH might also act as "main factors".

Conclusion

The results and methods presented in this study represent an important contribution to the discussion around the use of microorganisms in lake ecosystem monitoring schemes. Firstly, they indicate that the physico-chemical status of a lake cannot fully be predicted by its microbiome (see figures 1, 2B), even if the microbiome can be used for biomonitoring [33]. Nevertheless, up to around 60% of the

variation in certain parameters can be predicted by the lakes microbial community composition, which is comparable to results from soil ecosystems [92]. Secondly, the predominance of bacteria among multi-task bioindicators (see figure 3) supports the view that, in lake ecosystems, bacteria are more responsive to changes in abiotic than eukaryotes [42]. This underscores the importance of including prokaryotes into official ecosystem monitoring schemes. Thirdly, the results that OTUs share more information with environmental parameters than any of the other taxonomic level calls for the use of bioindicators at or below the OTU level. While higher-order analyses might suffice for monitoring, the precise delineation of taxonomic groups at the OTU level is especially necessary to gain insight into the microorganisms' autecology.

Abbreviations

Alk.Gram: Alkalinity; Cat.Sum: sum of cations; COND: conductivity; Coord.N: latitude; Coord.O: longitude; DOC: dissolved organic carbon; dissolved reactive silica (DRSi); FCBF: fast correlation-based filter; LF: conductivity, measured in the field; Sum.Ions: sum of ions; T: temperature; TP: total phosphorus.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the LOEWE program of the State of Hesse (Germany) in the MOSLA research cluster. We also acknowledge funding by the Bauer-Foundation and Stemmler-Foundation for the project "Differential potential of metabarcoding, metatranscriptomics, and metagenomics for the assessment of lake water quality" and of the DFG project BO 3245/19-1.

Author's contributions

TS designed and performed the data analyses, NK and DB contributed substantially to the bioindicator analysis, DB and JB provided the datasets, JB, GH, and DH supervised the study. All authors discussed the results and wrote and revised the manuscript.

Acknowledgements

Calculations on the MaRC2 high-performance computer of the University of Marburg were conducted for this research. We would like to thank René Sitt of HPC-Hessen, funded by the State Ministry of Higher Education, Research and the Arts, for installation and maintenance of software on the MaRC2 high-performance computer. We would like to thank Julia Nuy and Marius Welzel for helping with data availability and Nils Richber for discussions on theoretical properties of systems.

Availability of data

Raw sequencing data are available under the NCBI BioProject IDs PRJNA414052 and PRJNA559862.

Author details

[1]Faculty of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Str. 6, D-35032 Marburg (Lahn), Germany. [2]Department of Biodiversity, Center for Water and Environmental Research, University of Duisburg-Essen, D-45141 Essen, Germany.

References

1. Steffen, W., Persson, Å., Deutsch, L., Zalasiewicz, J., Williams, M., Richardson, K., Crumley, C., Crutzen, P., Folke, C., Gordon, L., Molina, M., Ramanathan, V., Rockström, J., Scheffer, M., Schellnhuber, H.J., Svedin, U.: The anthropocene: From global change to planetary stewardship. AMBIO **40**(7), 739–761 (2011). doi:10.1007/s13280-011-0185-x
2. Ceballos, G., Ehrlich, P.R., Barnosky, A.D., Garcia, A., Pringle, R.M., Palmer, T.M.: Accelerated modern human-induced species losses: Entering the sixth mass extinction. Science Advances **1**(5), 1400253–1400253 (2015). doi:10.1126/sciadv.1400253
3. Williams, M., Zalasiewicz, J., Haff, P., Schwägerl, C., Barnosky, A.D., Ellis, E.C.: The anthropocene biosphere. The Anthropocene Review **2**(3), 196–219 (2015). doi:10.1177/2053019615591020
4. Isbell, F., Gonzalez, A., Loreau, M., Cowles, J., Díaz, S., Hector, A., Mace, G.M., Wardle, D.A., O'Connor, M.I., Duffy, J.E., Turnbull, L.A., Thompson, P.L., Larigauderie, A.: Linking the influence and dependence of people on biodiversity across scales. Nature **546**(7656), 65–72 (2017). doi:10.1038/nature22899
5. Tilman, D., Clark, M., Williams, D.R., Kimmel, K., Polasky, S., Packer, C.: Future threats to biodiversity and pathways to their prevention. Nature **546**(7656), 73–81 (2017). doi:10.1038/nature22900
6. IPBES: Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. IPBES secretariat, Bonn, Germany, ??? (2019)

7. Evans, M.R., Bithell, M., Cornell, S.J., Dall, S.R.X., Díaz, S., Emmott, S., Ernande, B., Grimm, V., Hodgson, D.J., Lewis, S.L., Mace, G.M., Morecroft, M., Moustakas, A., Murphy, E., Newbold, T., Norris, K.J., Petchey, O., Smith, M., Travis, J.M.J., Benton, T.G.: Predictive systems ecology. Proceedings of the Royal Society B: Biological Sciences **280**(1771), 20131452 (2013). doi:10.1098/rspb.2013.1452
8. Jorgensen, S.E.: Introduction to Systems Ecology. Applied Ecology and Environmental Management. CRC Press, ??? (2016). https://books.google.de/books?id=8oTRBQAAQBAJ
9. Otwell, A.E., de Lomana, A.L.G., Gibbons, S.M., Orellana, M.V., Baliga, N.S.: Systems biology approaches towards predictive microbial ecology. Environmental Microbiology **20**(12), 4197–4209 (2018). doi:10.1111/1462-2920.14378
10. Webster, N.S., Wagner, M., Negri, A.P.: Microbial conservation in the anthropocene. Environmental Microbiology **20**(6), 1925–1928 (2018). doi:10.1111/1462-2920.14124
11. Krikorian, N.: The volterra model for three species predator-prey systems: Boundedness and stability. Journal of Mathematical Biology **7**(2), 117–132 (1979). doi:10.1007/bf00276925
12. Ulanowicz, R.E.: Information theory in ecology. Computers & Chemistry **25**(4), 393–399 (2001). doi:10.1016/s0097-8485(01)00073-0
13. Heink, U., Kowarik, I.: What are indicators? On the definition of indicators in ecology and environmental planning. Ecological Indicators **10**(3), 584–593 (2010). doi:10.1016/j.ecolind.2009.09.009
14. Karimi, B., Maron, P.A., Boure, N.C.-P., Bernard, N., Gilbert, D., Ranjard, L.: Microbial diversity and ecological networks as indicators of environmental quality. Environmental Chemistry Letters **15**(2), 265–281 (2017). doi:10.1007/s10311-017-0614-6
15. Hering, D., Borja, A., Carstensen, J., Carvalho, L., Elliott, M., Feld, C.K., Heiskanen, A.-S., Johnson, R.K., Moe, J., Pont, D.: The european water framework directive at the age of 10: A critical review of the achievements with recommendations for the future. Science of The Total Environment **408**(19), 4007–4019 (2010). doi:10.1016/j.scitotenv.2010.05.031
16. Birk, S., Bonne, W., Borja, A., Brucet, S., Courrat, A., Poikane, S., Solimini, A., van de Bund, W., Zampoukas, N., Hering, D.: Three hundred ways to assess europe's surface waters: An almost complete overview of biological methods to implement the water framework directive. Ecological Indicators **18**, 31–41 (2012). doi:10.1016/j.ecolind.2011.10.009
17. Plassart, P., Prévost-Bouré, N.C., Uroz, S., Dequiedt, S., Stone, D., Creamer, R., Griffiths, R.I., Bailey, M.J., Ranjard, L., Lemanceau, P.: Soil parameters, land use, and geographical distance drive soil bacterial communities along a european transect. Scientific Reports **9**(1) (2019). doi:10.1038/s41598-018-36867-2
18. Merkley, M., Rader, R.B., McArthur, J.V., Eggett, D.: Bacteria as bioindicators in wetlands: Bioassessment in the Bonneville Basin of Utah, USA. Wetlands **24**(3), 600–607 (2004). doi:10.1672/0277-5212(2004)024[0600:babiwb]2.0.co;2
19. Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., Pawlowski, J.: Embracing environmental genomics and machine learning for routine biomonitoring. Trends in Microbiology **27**(5), 387–397 (2019). doi:10.1016/j.tim.2018.10.012
20. Frühe, L., Cordier, T., Dully, V., Breiner, H.-W., Lentendu, G., Pawlowski, J., Martins, C., Wilding, T.A., Stoeck, T.: Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. Molecular Ecology (2020). doi:10.1111/mec.15434
21. Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Frigerio, J.-M., Humbert, J.-F., Bouchez, A.: A next-generation sequencing approach to river biomonitoring using benthic diatoms. Freshwater Science **33**(1), 349–363 (2014). doi:10.1086/675079
22. Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., Tyson, G.W.: Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology (2017). doi:10.1038/s41564-017-0012-7
23. Cazelles, K., Araújo, M.B., Mouquet, N., Gravel, D.: A theory for species co-occurrence in interaction networks. Theoretical Ecology **9**(1), 39–48 (2015). doi:10.1007/s12080-015-0281-9
24. Harris, D.J.: Inferring species interactions from co-occurrence data with Markov networks. Ecology **97**(12), 3308–3314 (2016). doi:10.1002/ecy.1605
25. Röttjers, L., Faust, K.: Can we predict keystones? Nature Reviews Microbiology **17**(3), 193–193 (2018). doi:10.1038/s41579-018-0132-y
26. Guimarães, P.R., Pires, M.M., Jordano, P., Bascompte, J., Thompson, J.N.: Indirect effects drive coevolution in mutualistic networks. Nature **550**(7677), 511–514 (2017). doi:10.1038/nature24273
27. Röttjers, L., Faust, K.: From hairballs to hypotheses–biological insights from microbial networks. FEMS Microbiology Reviews **42**(6), 761–780 (2018). doi:10.1093/femsre/fuy030
28. Miller, T.E., Travis, J.: The evolutionary role of indirect effects in communities. Ecology **77**(5), 1329–1335 (1996). doi:10.2307/2265530
29. Deltedesco, E., Keiblinger, K.M., Piepho, H.-P., Antonielli, L., Pötsch, E.M., Zechmeister-Boltenstern, S., Gorfer, M.: Soil microbial community structure and function mainly respond to indirect effects in a multifactorial climate manipulation experiment. Soil Biology and Biochemistry **142**, 107704 (2020). doi:10.1016/j.soilbio.2020.107704
30. Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E.R., Knight, R.: Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome **5**(1) (2017). doi:10.1186/s40168-017-0237-y
31. Tan, B., Ng, C., Nshimyimana, J.P., Loh, L.L., Gin, K.Y.-H., Thompson, J.R.: Next-generation sequencing (NGS) for assessment of microbial water quality: current progresschallenges and future opportunities. Frontiers in Microbiology **6** (2015). doi:10.3389/fmicb.2015.01027
32. Grossmann, L., Beisser, D., Bock, C., Chatzinotas, A., Jensen, M., Preisfeld, A., Psenner, R., Rahmann, S., Wodniok, S., Boenigk, J.: Trade-off between taxon diversity and functional diversity in european lake ecosystems. Molecular Ecology **25**(23), 5876–5888 (2016). doi:10.1111/mec.13878

33. Cordier, T., Forster, D., Dufresne, Y., Martins, C.I.M., Stoeck, T., Pawlowski, J.: Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. Molecular Ecology Resources **18**(6), 1381–1391 (2018). doi:10.1111/1755-0998.12926
34. Mikhailov, I.S., Zakharova, Y.R., Bukin, Y.S., Galachyants, Y.P., Petrova, D.P., Sakirko, M.V., Likhoshway, Y.V.: Co-occurrence networks among bacteria and microbial eukaryotes of Lake Baikal during a spring phytoplankton bloom. Microbial Ecology (2018). doi:10.1007/s00248-018-1212-2
35. Sperlea, T., Füser, S., Boenigk, J., Heider, D.: SEDE-GPS: socio-economic data enrichment based on GPS information. BMC Bioinformatics **19**(S15) (2018). doi:10.1186/s12859-018-2419-4
36. Cordier, T.: Bacterial communities' taxonomic and functional turnovers both accurately predict marine benthic ecological quality status. Environmental DNA (2019). doi:10.1002/edn3.55
37. Glasl, B., Bourne, D.G., Frade, P.R., Thomas, T., Schaffelke, B., Webster, N.S.: Microbial indicators of environmental perturbations in coral reef ecosystems. Microbiome **7**(1), (2019). doi:10.1186/s40168-019-0705-7
38. Kiersztyn, B., Chróst, R., Kaliński, T., Siuda, W., Bukowska, A., Kowalczyk, G., Grabowska, K.: Structural and functional microbial diversity along a eutrophication gradient of interconnected lakes undergoing anthropopressure. Scientific Reports **9**(1) (2019). doi:10.1038/s41598-019-47577-8
39. Han, M., Dsouza, M., Zhou, C., Li, H., Zhang, J., Chen, C., Yao, Q., Zhong, C., Zhou, H., Gilbert, J.A., Wang, Z., Ning, K.: Agricultural risk factors influence microbial ecology in Honghu lake. GenomicsProteomics & Bioinformatics **17**(1), 76–90 (2019). doi:10.1016/j.gpb.2018.04.008
40. Boenigk, J., Wodniok, S., Bock, C., Beisser, D., Hempel, C., Grossmann, L., Lange, A., Jensen, M.: Geographic distance and mountain ranges structure freshwater protist communities on a european scale. Metabarcoding and Metagenomics **2**, 21519 (2018). doi:10.3897/mbmg.2.21519
41. Bock, C., Salcher, M., Jensen, M., Pandey, R.V., Boenigk, J.: Synchrony of eukaryotic and prokaryotic planktonic communities in three seasonally sampled austrian lakes. Frontiers in Microbiology **9** (2018). doi:10.3389/fmicb.2018.01290
42. Bock, C., Jensen, M., Forster, D., Marks, S., Nuy, J., Psenner, R., Beisser, D., Boenigk, J.: Factors shaping community patterns of protists and bacteria on a european scale. Environmental Microbiology (2020). doi:10.1111/1462-2920.14992
43. Williamson, C.E., Dodds, W., Kratz, T.K., Palmer, M.A.: Lakes and streams as sentinels of environmental change in terrestrial and atmospheric processes. Frontiers in Ecology and the Environment **6**(5), 247–254 (2008). doi:10.1890/070140
44. García-García, N., Tamames, J., Linz, A.M., Pedrós-Alió, C., Puente-Sánchez, F.: Microdiversity ensures the maintenance of functional microbial communities under changing environmental conditions. The ISME Journal (2019). doi:10.1038/s41396-019-0487-8
45. Dudgeon, D., Arthington, A.H., Gessner, M.O., Kawabata, Z.-I., Knowler, D.J., Lévêque, C., Naiman, R.J., Prieur-Richard, A.-H., Soto, D., Stiassny, M.L.J., Sullivan, C.A.: Freshwater biodiversity: importance, threats, status and conservation challenges. Biological Reviews **81**(02), 163 (2005). doi:10.1017/s1464793105006950
46. WWF: Living Planet Report - 2018: Aiming Higher. WWF, Gland, Switzerland (2018)
47. Nuy, J.K., Hoetzinger, M., Hahn, M.W., Beisser, D., Boenigk, J.: Ecological differentiation in two major freshwater bacterial taxa along environmental gradients. Frontiers in Microbiology **11** (2020). doi:10.3389/fmicb.2020.00154
48. Andrews, S.: FASTQC. A quality control tool for high throughput sequence data (2010)
49. Schmieder, R., Edwards, R.: Quality control and preprocessing of metagenomic datasets. Bioinformatics **27**(6), 863–864 (2011). doi:10.1093/bioinformatics/btr026
50. Masella, A.P., Bartram, A.K., Truszkowski, J.M., Brown, D.G., Neufeld, J.D.: PANDAseq: paired-end assembler for Illumina sequences. BMC Bioinformatics **13**(1), 31 (2012). doi:10.1186/1471-2105-13-31
51. Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., Knight, R.: UCHIME improves sensitivity and speed of chimera detection. Bioinformatics **27**(16), 2194–2200 (2011). doi:10.1093/bioinformatics/btr381
52. Lange, A., Jost, S., Heider, D., Bock, C., Budeus, B., Schilling, E., Strittmatter, A., Boenigk, J., Hoffmann, D.: AmpliconDuo: A split-sample filtering protocol for high-throughput amplicon sequencing of microbial communities. PLOS ONE **10**(11), 0141590 (2015). doi:10.1371/journal.pone.0141590
53. Mahé, F., Rognes, T., Quince, C., de Vargas, C., Dunthorn, M.: Swarm: robust and fast clustering method for amplicon-based studies. PeerJ **2**, 593 (2014). doi:10.7717/peerj.593
54. Jensen, M.: V9_Clust.R – R-Script for modifying DNA-sequence-abundance tables: clustering of related sequences (e.g. SSU-ITS1) according to 100% identical sub-sequences. (2017). https://github.com/manfred-uni-essen/V9-cluster
55. Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Fawcett, T., Mishra, N. (eds.) Proceedings of the Twentieth International Conference on Machine Learning. Proceedings of the Twentieth International Conference on Machine Learning, pp. 856–863 (2003)
56. De Cáceres, M., Legendre, P.: Associations between species and groups of sites: indices and statistical inference (2009). http://sites.google.com/site/miqueldecaceres/
57. Proulx, S.R., Promislow, D.E.L., Phillips, P.C.: Network thinking in ecology and evolution. Trends in Ecology & Evolution **20**(6), 345–353 (2005). doi:10.1016/j.tree.2005.04.004
58. Chaffron, S., Rehrauer, H., Pernthaler, J., von Mering, C.: A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Research **20**(7), 947–959 (2010). doi:10.1101/gr.104521.109
59. Legendre, P., Gallagher, E.D.: Ecologically meaningful transformations for ordination of species data. Oecologia **129**(2), 271–280 (2001). doi:10.1007/s004420100716
60. Kuhn, M.: Building predictive models in R using the caret package. Journal of Statistical Software **28**(5) (2008). doi:10.18637/jss.v028.i05
61. Dufrêne, M., Legendre, P.: Species assemblages and indicator species: The need for a flexible asymmetrical approach. Ecological Monographs **67**(3), 345–366 (1997).

doi:10.1890/0012-9615(1997)067[0345:saaist]2.0.co;2
62. Schlegel, I., Koschel, R., Krienitz, L.: On the occurrence of Phacotus lenticularis (Chlorophyta) in lakes of different trophic state. Hydrobiologia **369/370**, 353–361 (1998). doi:10.1023/a:1017019925927
63. Jiang, J.-G., Shen, Y.-F.: Use of the aquatic protozoa to formulate a community biotic index for an urban water system. Science of The Total Environment **346**(1-3), 99–111 (2005). doi:10.1016/j.scitotenv.2004.12.001
64. Yang, Y., Li, S., Gao, Y., Chen, Y., Zhan, A.: Environment-driven geographical distribution of bacterial communities and identification of indicator taxa in Songhua River. Ecological Indicators **101**, 62–70 (2019). doi:10.1016/j.ecolind.2018.12.047
65. Wei, J., Gao, J., Wang, N., Liu, Y., Wang, Y., Bai, Z., Zhuang, X., Zhuang, G.: Differences in soil microbial response to anthropogenic disturbances in Sanjiang and Momoge Wetlands, China. FEMS Microbiology Ecology (2019). doi:10.1093/femsec/fiz110
66. Puranik, S., Pal, R.R., More, R.P., Purohit, H.J.: Metagenomic approach to characterize soil microbial diversity of Phumdi at Loktak Lake. Water Science and Technology **74**(9), 2075–2086 (2016). doi:10.2166/wst.2016.370
67. Sharuddin, S.S., Ramli, N., Hassan, M.A., Mustapha, N.A., Amran, A., Mohd-Nor, D., Sakai, K., Tashiro, Y., Shirai, Y., Maeda, T.: Bacterial community shift revealed Chromatiaceae and Alcaligenaceae as potential bioindicators in the receiving river due to palm oil mill effluent final discharge. Ecological Indicators **82**, 526–529 (2017). doi:10.1016/j.ecolind.2017.07.038
68. 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. Applied Microbiology and Biotechnology **103**(16), 6407–6421 (2019). doi:10.1007/s00253-019-09963-0
69. Reis, M.P., Suhadolnik, M.L.S., Dias, M.F., Ávila, M.P., Motta, A.M., Barbosa, F.A.R., Nascimento, A.M.A.: Characterizing a riverine microbiome impacted by extreme disturbance caused by a mining sludge tsunami. Chemosphere **253**, 126584 (2020). doi:10.1016/j.chemosphere.2020.126584
70. Henson, M.W., Hanssen, J., Spooner, G., Fleming, P., Pukonen, M., Stahr, F., Thrash, J.C.: Nutrient dynamics and stream order influence microbial community patterns along a 2914 kilometer transect of the Mississippi River. Limnology and Oceanography **63**(5), 1837–1855 (2018). doi:10.1002/lno.10811
71. Coclet, C., Garnier, C., Durrieu, G., Omanović, D., D'Onofrio, S., Poupon, C.L., Mullot, J.-U., Briand, J.-F., Misson, B.: Changes in bacterioplankton communities resulting from direct and indirect interactions with trace metal gradients in an urbanized marine coastal area. Frontiers in Microbiology **10** (2019). doi:10.3389/fmicb.2019.00257
72. Kegler, H.F., Hassenrück, C., Kegler, P., Jennerjahn, T.C., Lukman, M., Jompa, J., Gärdes, A.: Small tropical islands with dense human population: differences in water quality of near-shore waters are associated with distinct bacterial communities. PeerJ **6**, 4555 (2018). doi:10.7717/peerj.4555
73. Chen, L., Tsui, M.M.P., Lam, J.C.W., Hu, C., Wang, Q., Zhou, B., Lam, P.K.S.: Variation in microbial community structure in surface seawater from pearl river delta: Discerning the influencing factors. Science of The Total Environment **660**, 136–144 (2019). doi:10.1016/j.scitotenv.2018.12.480
74. Pershina, E., Valkonen, J., Kurki, P., Ivanova, E., Chirak, E., Korvigo, I., Provorov, N., Andronov, E.: Comparative analysis of prokaryotic communities associated with organic and conventional farming systems. PLOS ONE **10**(12), 0145072 (2015). doi:10.1371/journal.pone.0145072
75. Grossmann, L., Jensen, M., Heider, D., Jost, S., Glücksman, E., Hartikainen, H., Mahamdallie, S.S., Gardner, M., Hoffmann, D., Bass, D., Boenigk, J.: Protistan community analysis: key findings of a large-scale molecular sampling. The ISME Journal **10**(9), 2269–2279 (2016). doi:10.1038/ismej.2016.10
76. Grossmann, L., Jensen, M., Pandey, R.V., Jost, S., Bass, D., Psenner, R., Boenigk, J.: Molecular investigation of protistan diversity along an elevation transect of alpine lakes. Aquatic Microbial Ecology **78**(1), 25–37 (2016). doi:10.3354/ame01798
77. Foissner, W., Berger, H.: A user-friendly guide to the ciliates (Protozoa Ciliophora) commonly used by hydrobiologists as bioindicators in rivers, lakes, and waste waters, with notes on their ecology. Freshwater Biology **35**(2), 375–482 (1996). doi:10.1111/j.1365-2427.1996.tb01775.x. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2427.1996.tb01775.x
78. Lee, S., Basu, S., Tyler, C.W., Wei, I.W.: Ciliate populations as bio-indicators at Deer Island Treatment Plant. Advances in Environmental Research **8**(3-4), 371–378 (2004). doi:10.1016/s1093-0191(02)00118-1
79. Bellinger, E.G., Sigee, D.C.: Freshwater Algae: Identification, Enumeration and Use as Bioindicators. Wiley, ??? (2015). https://books.google.de/books?id=rhMmBgAAQBAJ
80. Bai, Y., Wang, Q., Liao, K., Jian, Z., Zhao, C., Qu, J.: Fungal community as a bioindicator to reflect anthropogenic activities in a river ecosystem. Frontiers in Microbiology **9** (2018). doi:10.3389/fmicb.2018.03152
81. Johnson, J.L., Fawley, M.W., Fawley, K.P.: The diversity of Scenedesmus and Desmodesmus (Chlorophyceae) in Itasca State Park, Minnesota, USA. Phycologia **46**(2), 214–229 (2007). doi:10.2216/05-69.1
82. Thessen, A.E., Patterson, D.J., Murray, S.A.: The taxonomic significance of species that have only been observed once: The genus Gymnodinium (Dinoflagellata) as an example. PLoS ONE **7**(8), 44015 (2012). doi:10.1371/journal.pone.0044015
83. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001). doi:10.1023/a:1010933404324
84. Edgar, R.C.: Updating the 97% identity threshold for 16s ribosomal RNA OTUs. Bioinformatics **34**(14), 2371–2375 (2018). doi:10.1093/bioinformatics/bty113
85. Maistrenko, O.M., Mende, D.R., Luetge, M., Hildebrand, F., Schmidt, T.S.B., Li, S.S., Rodrigues, J.F.M., von Mering, C., Coelho, L.P., Huerta-Cepas, J., Sunagawa, S., Bork, P.: Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. The ISME Journal **14**(5), 1247–1259 (2020). doi:10.1038/s41396-020-0600-z
86. Piwosz, K., Shabarova, T., Pernthaler, J., Posch, T., Šimek, K., Porcal, P., Salcher, M.M.: Bacterial and eukaryotic small-subunit amplicon data do not provide a quantitative picture of microbial communities, but they are reliable in the context of ecological interpretations. mSphere **5**(2) (2020). doi:10.1128/msphere.00052-20
87. Washburne, A.D., Silverman, J.D., Leff, J.W., Bennett, D.J., Darcy, J.L., Mukherjee, S., Fierer, N., David,

L.A.: Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ **5**, 2969 (2017). doi:10.7717/peerj.2969
88. Chen, W., Zhang, C.K., Cheng, Y., Zhang, S., Zhao, H.: A comparison of methods for clustering 16s rRNA sequences into OTUs. PLoS ONE **8**(8), 70837 (2013). doi:10.1371/journal.pone.0070837
89. Massana, R., Logares, R.: Eukaryotic versus prokaryotic marine picoplankton ecology. Environmental Microbiology **15**(5), 1254–1261 (2012). doi:10.1111/1462-2920.12043
90. Logares, R., Tesson, S.V.M., Canbäck, B., Pontarp, M., Hedlund, K., Rengefors, K.: Contrasting prevalence of selection and drift in the community structuring of bacteria and microbial eukaryotes. Environmental Microbiology (2018). doi:10.1111/1462-2920.14265
91. Karlsson, J., Jonsson, A., Jansson, M.: Productivity of high-latitude lakes: climate effect inferred from altitude gradient. Global Change Biology **11**(5), 710–715 (2005). doi:10.1111/j.1365-2486.2005.00945.x
92. Hermans, S.M., Buckley, H.L., Case, B.S., Curran-Cournane, F., Taylor, M., Lear, G.: Bacteria as emerging indicators of soil condition. Applied and Environmental Microbiology (2016). doi:10.1128/aem.02826-16

Figures

Figure 1 Information shared between the microbial community composition of a lake and its parameters. Results are shown for the *all_samples* dataset using FCBF as feature selection method. Lines represent 95% confidence intervals calculated from re-sampling, dots represent the median of re-sampled values. Some of the model-parameter combinations are not computable because of too high microbial community dimensionality.

Figure 2 IndVal as feature selection method for the all_features dataset. (A) Difference of median R^2 between models trained on FCBF- and IndVal-filtered microbial community composition. Negative values indicate better performance using FCBF, positive values indicate better performance using IndVal. Missing data points indicate combinations of taxonomic level and parameter that were not computable because of too high dimensionality after using FCBF. (B) Quantification of information shared between the microbial community composition and physico-chemical parameters of a lake using IndVal as feature selection method. Lines represent 95% confidence intervals calculated from re-sampling, dots represent the median of re-sampled values. Grey lines (labelled "all") represent results of models trained on a concatenation of all data from different taxonomic levels. Ion names represent concentrations. For the results for other models and taxonomic levels, see supplementary table 2.

Figure 3 Multi-task bioindicators for lake ecosystem parameters. (A) OTUs indicative for more than 7 parameters, (B) Eukaryotic OTUs indicative for more than 2 parameters. Dot size represents indicator statistic magnitude. Dendrogram and parameter order are derived from all-vs-all Pearson correlations in the *all_features* dataset. For taxonomic annotation of the OTUs, see table 3 and supplementary table 4, for (A) and (B), respectively.

Tables

Table 1 Dimensionality of taxonomic levels, as well as average dimensionality after dimensionality reduction via FCBF and the IndVal method for the *all_features* dataset.

Level	Taxa	FCBF	IndVal
Domain	3	3	-
Phylum	76	76	3.22
Class	253	244	10.10
Order	752	714	24.16
Family	885	857	33.72
Genus	2 353	2 242	69.41
Species	5 384	4 967	80.49
OTU	315 731	279 952	721.07

Additional Files

Additional file 1 — Supplementary Table 1.
Lake/Sample IDs present in the all_samples and all_features datasets.

Table 2 Number of OTUs identified as bioindicators in the IndVal analysis for the physico-chemical lake parameters.

Parameter	Number	Parameter	Number
Alk.Gran	16	K^{+}	118
Altitude	1 595	LF	1 349
Anions	89	Mg^{+2}	70
Ca^{+2}	32	Na^{2}	86
Cations	164	NH_4	6
CatSum	5	NO_3	17
Cl^{-}	48	pH	603
COND	1	SO_4	3
DOC	43	SumIons	166
DRSi	1	T	920
H^{+}	15	TP	92
HCO_3	27		

Additional file 2 — Supplementary Table 2.
Shared information measured using the machine learning framework presented in this study, containing results for models and taxonomical levels not reported in the main paper, for both FCBF as well as IndVal as feature selection method.

Additional file 3 — Supplementary Table 3.
IndVal results for all taxonomic levels and target variables.

Additional file 4 — Supplementary Table 4.
Multi-task bioindicators for physico-chemical lake parameters with taxonomic annotation.

Additional file 5 — Supplementary Table 5.
Significance of feature importance difference. This file contains a list of p-values that resulted from t-tests assessing the difference of the distribution of feature importances (from the Random Forest model) for bacterial vs. eukaryotic OTUs for a set of environmental parameters.

Table 3 Multi-task bioindicators that have been identified for more than 7 lake parameters. Highlighted rows contain chloroplasts identified based on 16s rRNA sequence. For a overview of parameters and indicator statistic for each of these OTUs, see figure 3.

ID	Freq	Phylum	Class	Order	Family	Genus	Species
N1077	10	Bacteroidetes	Cytophagia	Cytophagales	Cyclobacteriaceae	uncultured	uncultured bacterium
N3553	10	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	*Roseomonas*	*Roseomonas sp.* S08
N513	10	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	*Actibacter*	uncultured bacterium
N2267	9	Ignavibacteriae	Ignavibacteria	Ignavibacteriales	PHOS-HE36	uncultured soil bacterium	
N2497	9	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	*Roseomonas*	groundwater biofilm bacterium H2
N569	9	Verrucomicrobia	Opitutae	Opitutales	Opitutaceae	*Opitutus*	uncultured soil bacterium
N177	8	Bacteroidetes	Flavobacteriia	Flavobacteriales	NS9 marine group	uncultured bacterium	
N1886	8	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	uncultured	uncultured soil bacterium
N209	8	Chloroplast of *Phacotus lenticularis*					
N2139	8	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	env.OPS 17	uncultured bacterium	
N313	8	Bacteroidetes	Flavobacteriia	Flavobacteriales	Cryomorphaceae	*Fluviicola*	uncultured bacterium
N3608	8	Proteobacteria	Betaproteobacteria	SC-I-84	uncultured bacterium		
N395	8	Bacteroidetes	Flavobacteriia	Flavobacteriales	Cryomorphaceae	*Fluviicola*	uncultured Bacteroidetes bacterium
N426	8	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	*Limnobacter*	uncultured bacterium
N533	8	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	*Novosphingobium*	uncultured bacterium
N60	8	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	NS11-12 marine group	uncultured Sphingobacterium sp.	
N636	8	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiales Incertae Sedis	*Rhizomicrobium*	uncultured bacterium
N642	8	Actinobacteria	Thermoleophilia	Gaiellales	uncultured	uncultured bacterium	
N6836	8	Proteobacteria	Deltaproteobacteria	Myxococcales	Sandaracinaceae	uncultured	uncultured bacterium
N735	8	Chloroplast of *Phacotus lenticularis*					

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ST1subdatasetsids.txt
- ST2frameworkresults.csv
- ST3indValsalltax.csv
- ST4multitaskbioindicators.csv
- ST5ttestresults.csv