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 (Ceballos et al., 2015; Isbell et al., 2017; Steffen et al., 2011; Tilman et al., 2017; Williams et al., 2015). 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 (Evans et al., 2011; Tilman et al., 2017; Williams et al., 2015).
Through interactions and the flow of energy and nutrients, different parts of an ecosystem are connected. 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 (Krikorian, 1979; Ulanowicz, 2001). The dynamic adaptability of ecosystems to environmental changes makes it possible to identify bioindicators, i.e., organisms whose presence and prevalence can be used to estimate other variables of the ecosystem (Heink & Kowarik, 2010; Karimi et al., 2017).

Bioindicators are used in biosphere-based ecosystem monitoring schemes such as the ones implemented in European countries under the Water Framework Directive (Birk et al., 2012; Hering et al., 2010) 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 (Plassart et al., 2019). This is the case since organisms will only emerge as indicative for environmental variables they respond to directly (because of their ecological niche) or indirectly (since they interact closely with organisms that are, in turn, responsive to changes in the respective environmental variable). 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 (Cordier et al., 2019; Frühe et al., 2020; Karimi et al., 2017; Merkley et al., 2004).

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 (Frühe et al., 2020; Kermarrec et al., 2014). Secondly, techniques such as amplicon sequencing have made it feasible to capture microbial community compositions present in environmental samples (Parks et al., 2017). As different microorganisms exhibit different responses to changes in a environmental variable, and these responses are modulated by other microorganisms, the microbial community composition as a whole is more indicative of the status of the ecosystem than a selection of bioindicator species separately.

However, while being rather intuitive, the systems ecology paradigm also exposes theoretical and methodical obstacles for the study of microbial communities. For example, the assumption of variable independence, which is a requirement for many statistical approaches, does not hold for all environmental variables or processes of an ecosystem. Similarly, in a system, processes are influencing and modulating each other, rendering the distinction between direct and indirect interactions hard or even infeasible (Jørgensen, 2016). This is especially the case for microbial communities, where interaction networks are hard to measure and validate (Cazelles et al., 2015; Harris, 2016; Heink & Kowarik, 2010; Röttjers & Faust, 2018a) and the distinction between indirect and direct interactions is an open question (Guimarães et al., 2017; Röttjers & Faust, 2018b). Indeed, many studies prove a high ecological relevance of indirect microbial interactions (Deltedesco et al., 2020; Miller & Travis, 1996).

Additional issues for the study of microbial communities stem from the sparsity and very high dimensionality of OTU tables (Röttjers & Faust, 2018b; Weiss et al., 2017). With a number of samples vastly lower than the number of regressors (in our case: taxa or OTUs), regression is ill-defined and the adjustment of the $R^2$ value for the number of regressors 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 for the study of microbial communities, as sampling and sequencing remain expensive and the high number of different microorganisms is a nonreducible property of the study object.

In this study, we developed methodological tools to study microbial communities 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 covariation between the microbiome and a total of 27 physicochemical and positional (i.e., GPS coordinates and altitude) variables of lake ecosystems (for an overview, see Figure 1, and for a list of variables, see Table 1). It builds upon a wealth of studies that elucidate the role of the microbiome in ecology using machine learning (Cordier, 2019; Cordier et al., 2018, 2019; Glasl et al., 2019; Grossmann, Beisser, et al., 2016; Han et al., 2019; Kiersztn et al., 2019; Mikhailov et al., 2018; Sperlea et al., 2018; Tan et al., 2015). In our covariation framework, a machine learning model is trained to approximate a projection of the microbial prevalence space to a single dimension for each of the environmental variables, which makes it able to handle the extremely high dimensionality of amplicon-based microbiome data sets. The coefficient of correlation $R^2$ between the projected microbial community composition and the measured environmental variable is, then, used as a metric of covariation. This corresponds to the covariation of the environmental variable and the whole microbiome, which is intuitively interpretable.

We applied this framework to a data set from a large-scale survey of European lakes (Bock et al., 2018, 2020; Boenigk et al., 2018). Lakes are considered as sentinels of ecosystem change at different temporal and geographical scales (García-García et al., 2019; Williamson et al., 2008). 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 (Dudgeon et al., 2005; World Wildlife Fund, 2018).

The use of nonlinear ensemble models facilitated a dimensionality reduction of up to six orders of magnitude while retaining important relationships in the amplicon data set. Comparing two feature selection methods that are motivated by ecology, we found that filtering for bioindicators leads to a favourable behaviour of the framework. Analysing the operational taxonomic units (OTUs) identified as bioindicators in the feature selection step, we identified bacteria and microbial eukaryotes indicative of multiple environmental variables of lakes, which support the notion of high interdependency between ecological variables.

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 physicochemical variables in lake ecosystems. To that end, we report a comprehensive quantification of the covariation of the complete microbiome and the environmental variables of lake ecosystems. At its core, our framework makes use of supervised machine learning methods to reduce the dimensionality of the microbial community composition. As the environmental variables are numerical features, the $R^2$ metric serves as metric for covariation. Two different feature selection methods were compared, and microbial bioindicators were extracted from the IndVal method. This work paves the way for future endeavours to better uncover the functional workings of ecosystems.

2 | MATERIALS AND METHODS

2.1 | Sample collection

Sampling was part of a pan-European study conducted in August 2012 (eukaryotic sequences are published in Boenigk et al., 2018; NCBI Bioproject PRJNA414052, prokaryotic sequences are published and described in Nuy et al., (2020), Bock et al., (2020); NCBI Bioproject PRJNA559862). To analyse the effects of biogeochemical 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 physicochemical and geographic factors can be found in Boenigk et al., 2018. For DNA analyses filtered samples were air-dried and frozen in liquid nitrogen (Cryoshippers) and stored at $-80^\circ$C until further processing.

2.2 | 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 Boenigk et al.,
2018. 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 × 300 bp reads. For details, please see Boenigk et al., (2018), Nuy et al., (2020) and Bock et al., (2020).

2.3 | Sequence analysis

Adapter removal, quality trimming, and demultiplexing using index sequences were performed by the sequencing company (Fasteris). Sequence processing was performed using a provisional version of the Natrix pipeline (Welzel et al., 2020). Base quality of raw sequence reads was rechecked using the fastqc software (v0.11.5; Andrews, 2010) 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; Schmieder & Edwards, 2011). The paired-end reads were assembled and quality filtered with the tool pandaseq (v2.10; Masella et al., 2012). 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 usinguchime (usearch v7.0.1090; Edgar et al., 2011). Additionally, a split-sample filtering protocol (AmpliconDuo; Lange et al., 2015) 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; Mahé et al., 2014) and OTU tables were generated based on this clustering. The eukaryotic representative sequences were further clustered by identical V9 sequences (V9_Clust.R; Jensen, 2017). 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.

2.4 | 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 values in the environmental parameter data set, two subdata sets were created, namely the all_samples and all_features subdata sets. The all_samples subdata set contains the environmental variables measured in the field (altitude, GPS coordinates, pH, conductivity, temperature, and time of sampling) and OTUs for 241 lakes. An additional set of 21 physicochemical variables had been measured for a subset of 47 lakes. Excluding the positional variables and the time measurement, lakes with the extended feature set and the corresponding OTUs constitute the all_features data set.

Outliers in the environmental variables 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 environmental variables relevant for the subdata set were excluded from further analysis, leading to 201 and 42 samples in the all_samples and all_features data set, respectively (see Table S1 for a list of lakes present in the subdata sets). This was done to reduce the variability in the data set as well as to remove potential measurement errors. OTUs and taxa absent from all samples in one of these subdata sets were removed. OTU and taxon counts were centred and scaled using the r function scale() before training.

2.5 | Covariation framework and machine learning

At the core of the covariation framework is a model that is trained to approximate this environmental variable based on the OTU table or taxonomically aggregated prevalence table. In the covariation framework, however, the common supervised machine learning approach is interpreted in a novel way, that is consistent with the theory behind machine learning as well as systems ecology: The prediction of the framework is interpreted as a projection of the microbial community composition to a single dimension that is comparable to the environmental variable the model was trained on. As metric for covariation, the coefficient of determination $R^2$ between the dimensionality-reduced microbiome and the measured values of the variable for the held-out samples was used. As a secondary metric, the root-mean-square error was also calculated.

The full model used to quantify the amount of covariation of the microbial community and an environmental variable consists of a feature selection method and a machine learning model, both of which will be described in the following paragraphs in more detail. These two steps form the full model of the covariation framework and are evaluated as one, for example the feature selection as well as the machine learning are evaluated based on the full model performance. Because of the low number of samples analysed here, a cross-validation scheme was used for model training and prediction as this results in final models with low bias even for small data sets (Bishop, 2006). This will be described at the end of the subsection.

Feature selection was performed using either a fast correlation-based filter (FCBF) (Yu & Liu, 2003) or the multipatt() function (IndVal method, 999 random permutations) from the r package indic species (v1.7.9: Cáceres & Legendre, 2009). The choice of the former was motivated by the widespread use of correlation networks as proxies for microbial interactions (Proulx et al., 2005). 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 neighbours in a correlation network into syntaxa, i.e., groups of organisms that act as one unit in environmental changes (Chaffron et al., 2010). For this filter, a cutoff of 0.6 was chosen for the Pearson correlation coefficient, because, consistently over taxonomic levels, only around one percent of intertaxa correlations showed higher correlation coefficients. For the IndVal analysis, which is of widespread use in ecological studies, samples were separated by tertiles of the variable in question and OTU and taxon occurrence numbers were standardised using the Hellinger transformation to decrease the influence of highly abundant OTUs (Legendre & Gallagher, 2001).

A total of seven machine learning models from the \texttt{caret} package \cite{Kuhn2008} were used as base learners in this study: random forest (\texttt{rf}), stochastic gradient boosting (\texttt{gbm}), extreme gradient boosting (\texttt{xgbTree}), support vector machines with linear and radial kernel (\texttt{svmLinear}, \texttt{svmRadial}), generalised linear model (\texttt{glmnet}), and \texttt{k}-nearest neighbours (\texttt{knn}). These models were trained using the \texttt{train()} function with default parameters, which includes hyperparameter tuning by grid search. Model predictions were generated using the \texttt{predict()} function.

To use a cross-validation scheme, the full set of samples was split into \textit{k} subsets of approximately equal size, and \textit{k} models are trained and used for prediction separately. While higher values of \textit{k} are known to reduce the bias in the evaluation, the runtime of the whole training process is also greatly affected by \textit{k}. Thus, for the all\_samples subdata set a 10-fold cross-validation and for the all\_features subdata set a leave-one-out cross-validation scheme was used as follows. For fold \textit{i}, all subsets except for subset \textit{i} were used in the training phase. The training phase consisted in, firstly, feature selection of the input features (i.e., taxa and OTU tables), and, secondly, fitting a model to approximate the target variable based on the selected features. Then, the fitted model was used to predict the target variables based in the held-out subset \textit{i}. These predictions were collected and compared to the respective measured value of the environmental variable in question. As performance metrics, the coefficient of determination, \( R^2 \), calculated between the one-dimensional microbiome (i.e., the prediction in a traditional machine learning scheme) and the measured values of the environmental variable values. We also calculated the root-mean-square error as a secondary metric for the performance of the framework (see Table S2). However, in the context of the covariation framework, the \( R^2 \) metric lends itself to a more straightforward interpretation, i.e., as the amount of variation in the environmental variable explained by the projected, one-dimensional microbial community composition.

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 data set (and, therefore, only for a reduced number of variables) using a 10-fold cross-validation evaluation scheme. The choice of this feature selection method was motivated by the use of correlation networks for microbial communities (Proulx et al., 2005).

To test the hypothesis that nonlinear, as well as linear, relationships between microorganisms are important for their response 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. In our results, models that can approximate both linear and nonlinear relationships between features (i.e., Random Forest and \texttt{xgbTree}) outperform linear models. This result suggests that nonlinear projections are necessary to capture environmentally relevant patterns in the microbiome in a single dimension and thus supports the notion that complex relationships are present between microbial community structure and environmental variables (see Figure 2). Based on this, we focus the presentation and discussion of further results to Random Forest models.

However, 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 2). This disproportion between sample number and feature space (i.e., taxon or OTU table) dimensionality made the application of the framework impossible for some of the environmental variables (see missing values in Figure 2).

### 3 | RESULTS

#### 3.1 | Nonlinear models capture relevant patterns in microbial community composition

In general, regression models can be seen as approximating a function that projects the input feature space to a one-dimensional target space, thus performing a supervised dimensionality reduction. Based on this notion, we developed a framework to quantify the covariation of an ecosystem’s microbial community composition and an environmental variable. In it, we train supervised machine learning models to project the OTU- or taxonomy-table microbiome obtained by amplicon sequencing to a single dimension that is comparable to the target variable in question. As a metric of covariation, we used the coefficient of determination, \( R^2 \), calculated between the one-dimensional microbiome (i.e., the prediction in a traditional machine learning scheme) and the measured values of the environmental variable values. We also calculated the root-mean-square error as a secondary metric for the performance of the framework (see Table S2). However, in the context of the covariation framework, the \( R^2 \) metric lends itself to a more straightforward interpretation, i.e., as the amount of variation in the environmental variable explained by the projected, one-dimensional microbial community composition.

In the context of the covariation framework, the \( R^2 \) metric lends itself to a more straightforward interpretation, i.e., as the amount of variation in the environmental variable explained by the projected, one-dimensional microbial community composition. In our results, models that can approximate both linear and nonlinear relationships between features (i.e., Random Forest and \texttt{xgbTree}) outperform linear models. This result suggests that nonlinear projections are necessary to capture environmentally relevant patterns in the microbiome in a single dimension and thus supports the notion that complex relationships are present between microbial community structure and environmental variables (see Figure 2). Based on this, we focus the presentation and discussion of further results to Random Forest models.

However, 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 2). This disproportion between sample number and feature space (i.e., taxon or OTU table) dimensionality made the application of the framework impossible for some of the environmental variables (see missing values in Figure 2).
As an alternative filtering method, we employed the IndVal method (Dufrêne & Legendre, 1997). 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 covary with a given lake variable. Applying IndVal as a feature selection method in our framework to the all_features data set 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 environmental variables and taxonomic levels, using IndVal leads to better results, albeit not significantly (see Figure 3a). Furthermore, for some combinations of taxonomic levels and environmental variables, the use of FCBF outperformed the use of IndVal. On the other hand, while some FCBF runs were not computable (highlighted by the missing values in Figure 3a), 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 communities. Based on these results, we conclude that most of the taxa or OTUs that covary with the respective environmental variable are contained in the IndVal selection.

### Table 2 Dimensionality of taxonomic levels, as well as average dimensionality after dimensionality reduction via fast correlation-based filter (FCBF) and the IndVal method for the all_features data set

| 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   | 2353 | 2242 | 69.41  |
| Species | 5384 | 4967 | 80.49  |
| OTU     | 315,731 | 279,952 | 721.07 |

### Figure 2 Covariation of the microbial community composition of a lake and its variables, for the all_samples data set using fast correlation-based filter (FCBF) as feature selection method. Lines represent 95% confidence intervals calculated from resampling, dots represent the median of resampled values. Some of the model-variable combinations are not computable because of too high microbial community dimensionality.

#### 3.2 Indicator species analysis as feature selection for microbiome dimensionality reduction

As an alternative filtering method, we employed the IndVal method (Dufrêne & Legendre, 1997). This calculates a composite indicator
3.3 Covariation at different taxonomic levels

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 physicochemical variables present in the all_features data set (Figure 3b). As seen for FCBF (see Figure 2), lower taxonomic levels covary more with the physicochemical variables than do higher levels. However, the results from different levels of taxonomy should be compared with care, because the number of regressors (i.e., taxa or OTUs) increases strongly with falling taxonomic level and the $R^2$ is known to increase monotonically with the number of regressors. The usual way to alleviate this is to adjust $R^2$ values to the feature space dimensionality, but this is not possible here because the dimensionality is much higher than the numbers of samples of the analysed data sets.
To test the hypothesis that different levels of microbial taxonomy respond with environmental variables 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 3b). These models do not significantly outperform the models trained on OTU prevalence tables although they were trained with a higher number of regressors (i.e., the sum of all taxa and OTUs). Therefore, we conclude that higher taxonomic levels do not contribute to ecologically relevant patterns not already present at the OTU level.

3.4 Analysis of microbial multitask bioindicators

The results presented to this point support the use of the IndVal to identify ecologically relevant taxa and OTUs from amplicon sequencing data. The numbers of bioindicators for different variables at different levels of taxonomy obtained this way ranged over four orders of magnitude (see Table 1 and Table S3). We analysed these bioindicator OTUs by focusing on multitask bioindicators, i.e., OTUs that emerged as indicative for multiple environmental variables and might, therefore, act as general indicators of lake ecosystem status.

All of the bioindicators indicative of more than seven variables are annotated as Bacteria (see Table 3 and Figure 4a) 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 (Jiang & Shen, 2005; Schlegel et al., 1998). 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 (Ignivibacteriales; Cordier, 2019), Limnobacter (Yang et al., 2019), and Sandaracinaceae (Wei et al., 2019), certain environmental variables (Opitutus; Puranik et al., 2016; Plassart et al., 2019), Alcaligenaceae (Sharuddin et al., 2017), Novosphingobium (Astudillo-García et al., 2019; Reis et al., 2020), and NS11-12 marine group (Coclet et al., 2019; Henson et al., 2018), and human interference/impact/pollution (Acbitacter (Kegler et al., 2018), Fluviicola (Chen et al., 2019), and SC-I-84 (Pershina et al., 2015)). However, not all of these taxa have previously been identified in lake ecosystems, and most of the OTUs among these bacterial multitask bioindicators are assigned to taxa originally isolated from soil ecosystems (see Table 3).

The multitask bioindicators among the eukaryotes are, at most, indicative for five environmental variables. Among the 32 OTUs that are indicative for more than two variables, six are annotated as Ciliophora or Chlorophyta. These classes are ubiquitous in lakes (Grossmann, Jensen, Heider, et al., 2016; Grossmann, Jensen, Pandey, et al., 2016; Mikhailov et al., 2018), contain many species that inhabit specific ecological niches and have been used as bioindicators (Bellinger & Sigee, 2015; Foissner & Berger, 1996; Lee et al., 2004). Similarly, many of the eukaryotic multitask OTUs identified here belong to genera that have been described as ubiquitous in freshwater ecosystems (e.g., Chytridiomycota (Bai et al., 2018), Desmodesmus (Johnson et al., 2007) or Gymnodinium (Thessen et al., 2012)). However, most of the species we identified have, to our knowledge, not yet been described as bioindicators at lower taxonomic levels. Notably, clustering the environmental variables according to their pairwise Pearson correlation results in patterns of multitask bioindicators in Figure 4. This further supports the notion that an interaction network underlies the microbial community structure of lake ecosystems and this network is shaped by environmental variables.

Based on our finding that bacterial OTUs can be indicative for more than five environmental variables 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 covariation framework. With the null hypothesis that the feature importances of eukaryotes and prokaryotes have the same means, we ran two-sample t tests and found significant differences (p < .05 after Bonferroni-Hochberg correction with n = 23) for the bioindicators for altitude, dissolved organic carbon (DOC), dissolved reactive silica (DRSi), hydrogen (H), potassium (K), ammonium (NH₄), nitrate (NO₃), sum of ions (Sum. Ions), and temperature (T) (see Table S5). For these environmental variables, thus, the mean feature importance of eukaryotic and prokaryotic bioindicators can be regarded as different. This result suggests that, at least with regard to these variables, bacteria and eukaryotes play different roles in lake ecosystems. However, for the other variables, 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 fully independent with regards to their ecological function.

4 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 covariation of the microbiome and environmental variables of lake ecosystems. We present a framework that acknowledges the technical obstacles presented by ecological data in general and molecular microbial community data sets in particular.

For the design of the covariation 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 as judged by the R² metric (Figure 1). This is most probably due to their ability to approximate highly nonlinear relationships and cope with large feature spaces (Breiman, 2001). Additionally, ensembles of decision trees are, in principle, capable of learning from data for which the independence assumption does not hold (Breiman, 2001). We were also able to show that while using FCBF and IndVal as feature selection methods leads to comparable results, the IndVal method results
**Table 3** Multitask bioindicators that have been identified for more than seven environmental variables. Highlighted rows contain chloroplasts identified based on 16 s rRNA sequence. For an overview of variables and indicator statistic for each of these operational taxonomic units (OTUs), see Figure 4.

| ID   | Freq | Phylum            | Class             | Order            | Family            | Genus                  | Species                        |
|------|------|-------------------|-------------------|------------------|-------------------|------------------------|--------------------------------|
| N1077| 10   | Bacteroidetes     | Cytophagia        | Cytophagales     | Cyclobacteriaceae | Uncultured bacterium  |                                |
| N3553| 10   | Proteobacteria    | Alphaproteobacteria| Rhodospirillales | Acetobacteriaceae | Roseomonas             | Roseomonas sp. S08             |
| N513 | 10   | Bacteroidetes     | Flavobacteriia    | Flavobacteriales | Flavobacteriaceae | Actibacter             | Uncultured bacterium           |
| N2267| 9    | Ignavibacteria    | Ignavibacteriaceae| Ignavibacteriales| PHOS-HE36         | Uncultured soil bacterium |                                |
| N2497| 9    | Proteobacteria    | Alphaproteobacteria| Rhodospirillales | Acetobacteriaceae | 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    | Fluvicola              | Uncultured bacterium           |
| N3608| 8    | Proteobacteria    | Betaproteobacteria| SC-1-84          | uncultured bacterium|                       |                                |
| N395 | 8    | Bacteroidetes     | Flavobacteriia    | Flavobacteriales | Cryomorphaceae    | Fluvicola              | Uncultured Bacteroidetes bacterium |
| N426 | 8    | Proteobacteria    | Betaproteobacteria| Burkholderiales  | Limnobacter        | Uncultured bacterium  |                                |
| N533 | 8    | Proteobacteria    | Alphaproteobacteria| Sphingomonadace  | 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   | Gaelliales       | uncultured        | uncultured bacterium  |                                |
| N6836| 8    | Proteobacteria    | Deltaproteobacteria| Myxococcales     | Sandaracinaceae   | uncultured bacterium  |                                |
| N735 | 8    | Chloroplast of Phacotus lenticularis |              |                  |                   |                       |                                |
in sparser models that allow the use of the framework even for extremely high-dimensional data sets at low levels of the taxonomy (see Table 2). While IndVal has been used for molecular data sets collected, for example, at the Great Barrier Reef (Glasl et al., 2019), this study is first in applying it to molecular data in the context of lake ecology.

Applying our framework to a data set of microbial communities collected in a large-scale survey of European lakes, we were able to quantify the covariation between the lake microbiome and a list of environmental parameters for different levels of taxonomy (Figure 3b). Due to the high dimensionality of environmental microbiomes, we are not able to conclude whether OTUs show a covariation significantly higher than any of the other levels of taxonomy. Nevertheless, our results show that, for most environmental variables, higher levels of taxonomy do not contain relevant patterns not already present on the OTU level (see Figure 3b), which is in contrast to the findings of others (Washburne et al., 2017).

In the analysis of bioindicator OTUs identified in this study, we focused on multitask bioindicators. Among the OTUs identified as bioindicators for more than seven environmental variables, most have been taxonomically assigned to uncultured soil bacteria (see Figure 4a and Table 3). Similarly, most high-ranked eukaryotic multitask bioindicators (see Figure 4b and Table S4) have been first identified in freshwater biomes, but not necessarily been found in lake samples yet. As the data set analysed here stems from lakes, this is most probably an artefact of imprecise taxonomic annotation (Chen et al., 2013), but might also point to the diversity of ecological niches inhabited by bacterial subspecies grouped into one OTU or species (García-García et al., 2019). Although soil and freshwater microbial community compositions differ significantly (Grossmann, Jensen, Heider, et al., 2016), microorganisms can enter lakes from soil ecosystems directly or, for example, via rivers that feed the lake. The emergence of *Phacotus lenticularis* as a multitask organism in both groups of organisms (see Table 3 and Figure 4b) underscores its role as a bioindicator.

Recent studies have argued for differences in ecological function between bacteria and microbial eukaryotes in lake ecosystems (Bock et al., 2020; Logares et al., 2018; Massana & Logares, 2012). More specifically, it has been argued that bacteria are more responsive to environmental changes than eukaryotes (Bock et al., 2020; Frühe et al., 2020; Karimi et al., 2017; Merkley et al., 2004). This is supported by the results of our study that bacteria that are multitask bioindicators can be indicative of more environmental variables of lakes than eukaryotic multitask bioindicators (see Figure 4). Moreover, 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 environmental variables (see Table S5). However, this is not the case for all variables. There are two main reasons for this. First, at the domain level, aggregated prevalence numbers do not covary much with environmental variables (see Figures 2, 3b). Second, 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 variables these multitask bioindicators are indicative of show a high degree of correlation (see Figure 4). Aside from underscoring the need for functional diversity in bioindicators if aiming at covering all environmental variables, this indicates that there are "main factors" among lake variables that influence a high number of other variables 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 (Karlsson et al., 2005). This is the case because lakes in the lowland mainly arise from rivers that have their source in mountain chains and get enriched with nutrients during their courses. In particular for eukaryotic multitask 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".

Nevertheless, our results also underscore the need for further studies that include large-scale amplicon sequencing surveys of ecosystems. This is mainly the case because the natural variability of environmental samples in general and lake ecosystems specifically is very high, leading to the rather large confidence intervals observed in this study. Thus, including more samples in analyses such as ours would enable to better model the heterogeneity of natural ecosystems and lead to more robust and powerful statistical results. Further studies that are based on larger data sets should also allow for analyses based on less-stringent outlier removal than applied in this study, representing a wider array of natural variation of lake ecosystems.

In principle, both the covariation framework as well as the bioindicator analysis can easily be applied to samples from different environmental sources and other sequencing methods, such as metagenomic and metatranscriptomic data sets, as long as there is a straightforward interpretation for the results of the IndVal method. This is especially noteworthy as the importance and popularity of metagenomic assays in microbial ecology has risen fast in the last years (Awasthi et al., 2020; Huguerth et al., 2015; Panwar et al., 2020; Vishnivetskaya et al., 2020; Zeng et al., 2016).

Taken together, our results represent an important contribution to the discussion around the use of microorganisms in lake ecosystem monitoring schemes. First, they indicate that the physicochemical status of a lake cannot fully be predicted by its microbiome (see Figures 2, 3b). Nevertheless, up to around 60% of the variation in certain variables can be predicted by the lakes microbial community composition, which is comparable to results from soil ecosystems (Hermans et al., 2016). Second, the predominance of bacteria among multitask bioindicators (see Figure 4) supports the view that, in lake ecosystems, bacteria are more responsive to changes in environmental variables than eukaryotes (Bock et al., 2020). This underscores the importance of including prokaryotes into official ecosystem monitoring schemes. Third, our results offer an insight into the autecology of microbial taxa and OTUs in their natural habitats by indicating which microbes react strongly to changes different environmental variables. These insights can lay the groundwork for novel, niche-based analyses of environmental microbiomes (Chase & Leibold, 2003).

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. 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. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST
The authors declare that they have no competing financial interests.

AUTHOR CONTRIBUTIONS
Theodor Sperlea designed and performed the data analyses, Nico Kreuder and Daniela Beisser contributed substantially to the bioindicator analysis, Daniela Beisser and Jens Boenigk provided the data sets, Jens Boenigk, Georges Hattab, and Dominik Heider supervised the study. All authors discussed the results and wrote and revised the manuscript.

DATA AVAILABILITY STATEMENT
Raw sequencing data are available under the NCBI BioProject IDs PRJNA414052 and PRJNA559862, and the physico-chemical parameter data under the https://doi.org/10.6084/m9.figshare.14039312.

ORCID
Theodor Sperlea https://orcid.org/0000-0003-4307-2963
Georges Hattab https://orcid.org/0000-0003-4168-8254
Dominik Heider https://orcid.org/0000-0002-3108-8311

REFERENCES
Andrews, S. (2010). FASTQC. A quality control tool for high-throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
Astudillo-Garcia, C., Hermans, S. M., Stevenson, B., Buckley, H. L., & Lear, G. (2019). Microbial assemblages and bioindicators as proxies for ecosystem health status: potential and limitations. Applied Microbiology and Biotechnology, 103(16), 6407–6421.
Awasthi, M. K., Ravindran, B., Sarasaiya, S., Chen, H., Wainaina, S., Singh, E., Liu, T., Kumar, S., Pandey, A., Singh, L., & Zhang, Z. (2020). Metagenomics for taxonomy profiling: tools and approaches. Bioengineering, 11(1), 356–374.
Bai, Y., Wang, Q., Liao, K., Jian, Z., Zhao, C., & Qu, J. (2018). Fungal community as a bioindicator to reflect anthropogenic activities in a river ecosystem. Frontiers in Microbiology, 9, 3152.
Bellinger, E. G., & Signe, D. C. (2015). Freshwater algae: Identification, enumeration and use as Bioindicators. Wiley.

Birk, S., Bonne, W., Borja, A., Brucket, S., Courrat, A., Poikane, S., Sollini, A., van de Bund, W., Zampoukas, N., & Hering, D. (2012). 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.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer Verlag.

Bock, C., Jensen, M., Forster, D., Marks, S., Nuy, J., Psenner, R., Beisser, D., & Boenigk, J. (2020). Factors shaping community patterns of protists and bacteria on a European scale. Environmental Microbiology, 22(6), 2243–2260.

Bock, C., Salcher, M., Jensen, M., Pandey, R. V., & Boenigk, J. (2018). Synchrony of eukaryotic and prokaryotic planktonic communities in three seasonally sampled austrian lakes. Frontiers in Microbiology, 9, 1290.

Boenigk, J., Wodniok, S., Bock, C., Beisser, D., Hempel, C., Grossmann, L., Lange, A., & Jensen, M. (2018). Geographic distance and mountain ranges structure freshwater protist communities on a European scale. Metabarcoding and Metagenomics, 2, e21519.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Cáceres, M. D., & Legendre, P. (2009). Associations between species and groups of sites: indices and statistical inference. Ecology, 90(12), 3566–3574. https://doi.org/10.1890/08-1823.1

Cazelles, K., Araújo, M. B., Mouquet, N., & Gravel, D. (2015). A theory for species co-occurrence in interaction networks. Theoretical Ecology, 9(1), 39–48.

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human-induced species losses: Entering the sixth mass extinction. Science Advances, 1(5), e1400253.

Chaffron, S., Rehrauer, H., Pernthaler, J., & von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Research, 20(7), 947–959.

Chase, J. M., & Leibold, M. A. (2003). Ecological Niches - Linking classical and contemporary approaches. University of Chicago Press.

Chen, L., Tsui, M. M., Lam, J. C., Hu, C., Wang, Q., Zhou, B., & Lam, P. K. (2019). Variation in microbial community structure in surface seawater from pearl river delta: Discerning the influencing factors. Science of the Total Environment, 660, 136–144.

Chen, W., Zhang, C. K., Zhang, C., Zhou, H., Gilbert, J. A., Wang, Z., & Ning, K. (2019). Agricultural risk factors influence microbial ecology in Honghu Lake. Genomics, Proteomics & Bioinformatics, 17(1), 76–90.

Dudnagon, 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. (2005). Freshwater biodiversity: importance, threats, status and conservation challenges. Biological Reviews, 81(2), 163.

Dufrêne, M., & Legendre, P. (1997). Species assemblages and indicator species: The need for a flexible asymmetrical approach. Ecological Monographs, 67(3), 345–366.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics, 27(16), 2194–2200.

Evans, M. R., Bithell, M., Cornell, S. J., Dall, S. R. X., Diaz, 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. (2013). Predictive systems ecology. Proceedings of the Royal Society B: Biological Sciences, 280(1771), 20131452.

Foisner, W., & Berger, H. (1996). 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.

Früh, L., Cordier, T., Dully, V., Breiner, H.-W., Lantendu, G., Pawlowski, J., Martins, C., Wilding, T. A., & Stoeck, T. (2020). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. Molecular Ecology, 1–19.

García-García, N., Tamames, J., Linz, A. M., Pedrós-Aliò, C., & Puente-Sánchez, F. (2019). Microdiversity ensures the maintenance of functional microbial communities under changing environmental conditions. The ISME Journal, 13(12), 2969–2983. https://doi.org/10.1038/s41396-019-0487-8

Glasi, B., Bourne, D. G., Frade, P. R., Thomas, T., Schaffelke, B., & Webster, N. S. (2019). Microbial indicators of environmental perturbations in coral reef ecosystems. Microbiome, 7, 94.

Grossmann, L., Beisser, D., Bock, C., Chatzinotas, A., Jensen, M., Preisfeld, A., Psenner, R., Rahmann, S., Wodniok, S., & Boenigk, J. (2016). Trade-off between taxon diversity and functional diversity in European lake ecosystems. Molecular Ecology, 25(23), 5876–5888.

Grossmann, L., Jensen, M., Heider, D., Jost, S., Glöcksmann, E., Hartikainen, H., Mahamadalie, S. S., Gardner, M., Hoffmann, D., Bass, D., & Boenigk, J. (2016). Protistan community analysis: key findings of a large-scale molecular sampling. The ISME Journal, 10(9), 2269–2279.

Grossmann, L., Jensen, M., Pandey, R. V., Jost, S., Bass, D., Psenner, R., & Boenigk, J. (2016). Molecular investigation of protistan diversity along an elevation transect of alpine lakes. Aquatic Microbial Ecology, 78(1), 25–37.

Guimarães, P. R., Pires, M. M., Jordano, P., Bascompte, J., & Thompson, J. N. (2017). Indirect effects drive coevolution in mutualistic networks. Nature, 550(7677), 511–514.

Han, M., Dsoouza, M., Zhou, C., Li, H., Zhang, J., Chen, C., Yao, Q., Zhong, C., Zhou, H., Gilbert, J. A., Wang, Z., & Ning, K. (2019). Agricultural risk factors influence microbial ecology in Honghu Lake. Genomics, Proteomics & Bioinformatics, 17(1), 76–90.

Harris, D. J. (2016). Inferring species interactions from co-occurrence data with Markov networks. Ecology, 97(12), 3308–3314.

Heink, U., & Kowarik, I. (2010). What are indicators? On the definition of indicators in ecology and environmental planning. Ecological Indicators, 10(3), 584–593.

Henson, M. W., Hanssen, J., Spooner, G., Fleming, P., Pokonen, M., Stahr, F., & Thrash, J. C. (2018). 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.

Hering, D., Borja, A., Carstensen, J., Carvalho, L., Elliott, M., Feld, C. K., Heiskanen, A.-S., Johnson, R. K., Moe, J., & Pont, D. (2010). The European Water Framework Directive at the age of 10: A critical
Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: Robust and fast clustering method for amplicon-based studies. PeerJ, 2, e593.

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., & Neufeld, J. D. (2012). PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics, 13(1), 31.

Massana, R., & Logares, R. (2012). Eukaryotic versus prokaryotic marine picoplankton ecology. Environmental Microbiology, 15(5), 1254–1261.

Merkley, M., Rader, R. B., McArthur, J. V., & Eggett, D. (2004). Bacteria as bioindicators in wetlands: Bioassessment in the Bonneville Basin of Utah, USA. Wetlands, 24(3), 600–607.

Mikhailov, I. S., Zakharova, Y. R., Bukin, Y. S., Galachyants, Y. P., Petrova, D. P., Sakirko, M. V., & Likhoshway, Y. V. (2018). Co-occurrence networks among bacteria and microbial eukaryotes of Lake Baikal during a spring phytoplankton bloom. Microbial Ecology, 77, 96–109.

Miller, T. E., & Travis, J. (1996). The evolutionary role of indirect effects in communities. Ecology, 77(5), 1329–1335.

Nuy, J. K., Hoetzinger, M., Hahn, M. W., Beisser, D., & Boenigk, J. (2020). Ecological differentiation in two major freshwater bacterial taxa along environmental gradients. Frontiers in Microbiology, 11, 154.

Otwell, A. E., de Lomana, A. L. G., Gibbons, S. M., Orellana, M. V., & Baliga, N. S. (2018). Systems biology approaches towards predictive microbial ecology. Environmental Microbiology, 20(12), 4197–4209.

Panwar, P., Allen, M. A., Williams, T. J., Hancock, A. M., Brazendale, S., Bevington, J., Roux, S., Páez-Espino, D., Nayfach, S., Berg, M., Schulz, F., Chen, I. A., Huntemann, M., Shapiro, N., Kyrpides, N. C., Woyke, T., Elloe-Fadros, E. A., & Cavicchioli, R. (2020). Influence of the polar light cycle on seasonal dynamics of an antarctic lake microbial community. Microbiome, 8, 116.

Parks, D. H., Rinke, C., Chuvachova, M., Chauveil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology, 2(11), 1533–1542. https://doi.org/10.1038/s41564-017-0012-7

Pershina, E., Valkonen, J., Kurki, P., Ivanova, E., Chirak, E., Korvigo, I., Provorov, N., & Andronov, E. (2015). Comparative analysis of prokaryotic communities associated with organic and conventional farming systems. PLoS One, 10(12), e0145072.

Plassart, P., Prévost-Bouré, N. C., Uroz, S., Dequiedt, S., Stone, D., Creamer, R., Griffiths, R. I., Bailey, M. J., Ranjard, L., & Lemanceau, P. (2019). Soil parameters, land use, and geographical distance drive soil bacterial communities along a European transect. Scientific Reports, 9, 605.

Proulx, S. R., Promislow, D. E., & Phillips, P. C. (2005). Network thinking in ecology and evolution. Trends in Ecology & Evolution, 20(6), 345–353.

Puranik, S., Pal, R. R., More, R. P., & Purohit, H. J. (2016). Metagenomic approach to characterize soil microbial diversity of Phumdi at Loktak Lake. Water Science and Technology, 74(9), 2075–2086.

Reis, M. P., Suhadolnik, M. L. S., Dias, M. F., Ávila, M. P., Motta, A. M., Barbosa, F. A., & Nascimento, A. M. (2020). Characterizing a riverine microbiome impacted by extreme disturbance caused by a mining sludge tsunami. Chemosphere, 253, 126584.

Röttjers, L., & Faust, K. (2018a). Can we predict keystones? Nature Reviews Microbiology, 17(3), 193.

Röttjers, L., & Faust, K. (2018b). From hairballs to hypotheses – biological insights from microbial networks. FEBS Microbiology Reviews, 42(6), 761–780.

Schlegel, I., Keschel, R., & Krienitz, L. (1998). On the occurrence of Phacotus lenticularis (Chlorophyta) in lakes of different trophic state. Hydrobiologia, 369(370), 353–361.

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics, 27(6), 863–864.
