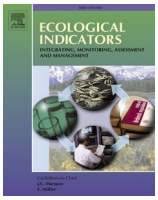
Ecological Indicators 139 (2022) 108955

Contents lists available at ScienceDirect

Ecological Indicators

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



Original Articles

Toward the use of protists as bioindicators of multiple stresses in agricultural soils: A case study in vineyard ecosystems

Bertrand Fournier [a,b], Magdalena Steiner [c], Xavier Brochet [d,e], Florine Degrune [a,b], Jibril Mammeri [d], Diogo Leite Carvalho [d,e,1], Sara Leal Siliceo [d], Sven Bacher [c], Carlos Andrés Peña-Reyes [d,e], Thierry J. Heger [b,*]

[a] Institute of Environmental Science and Geography, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany
[b] Soil Science and Environment Group, CHANGINS, HES-SO University of Applied Sciences and Arts Western Switzerland, Route de Duillier 50, 1260 Nyon, Switzerland
[c] Applied Ecology Group, Department of Biology, Ch. du Musée 10, CH-1700 Fribourg, Switzerland
[d] School of Business and Engineering Vaud (HEIG-VD), HES-SO University of Applied Sciences and Arts Western Switzerland, Switzerland
[e] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

ARTICLE INFO

Keywords:
Biomonitoring
Machine learning
Predictive model
Soil function
Soil quality
Microbial ecology

ABSTRACT

Management of agricultural soil quality requires fast and cost-efficient methods to identify multiple stressors that can affect soil organisms and associated ecological processes. Here, we propose to use soil protists which have a great yet poorly explored potential for bioindication. They are ubiquitous, highly diverse, and respond to various stresses to agricultural soils caused by frequent management or environmental changes.

We test an approach that combines metabarcoding data and machine learning algorithms to identify potential stressors of soil protist community composition and diversity. We measured 17 key variables that reflect various potential stresses on soil protists across 132 plots in 28 Swiss vineyards over 2 years. We identified the taxa showing strong responses to the selected soil variables (potential bioindicator taxa) and tested for their predictive power.

Changes in protist taxa occurrence and, to a lesser extent, diversity metrics exhibited great predictive power for the considered soil variables. Soil copper concentration, moisture, pH, and basal respiration were the best predicted soil variables, suggesting that protists are particularly responsive to stresses caused by these variables. The most responsive taxa were found within the clades Rhizaria and Alveolata. Our results also reveal that a majority of the potential bioindicators identified in this study can be used across years, in different regions and across different grape varieties.

Altogether, soil protist metabarcoding data combined with machine learning can help identifying specific abiotic stresses on microbial communities caused by agricultural management. Such an approach provides complementary information to existing soil monitoring tools that can help manage the impact of agricultural practices on soil biodiversity and quality.

1. Introduction

In a context of ongoing human population growth where agricultural landscapes occupy an increasing proportion of the total land area on Earth, there is a growing need for assessing the impact of agriculture on soil quality and functioning. Management of agricultural soil quality relies mostly on in situ measurements of multiple variables such as pH, bulk density, nutrient, and pesticide concentration. Despite providing quantitative information on soil quality and how it changes through space and time, such in situ measurements do not directly reflect the stresses caused by agricultural practices on soil organisms and, thereby, do not integrate the key role that soil organisms play in soil functions such as plant productivity and nutrient cycling (Giller et al., 1997, Wagg et al., 2014). Bioindicators that can reflect several abiotic stressors affecting soil biodiversity and its associated functions are thus needed for the management of agricultural soil quality.

* Corresponding author.
E-mail address: thierry.heger@changins.ch (T.J. Heger).
[1] Deceased.

https://doi.org/10.1016/j.ecolind.2022.108955
Received 13 March 2022; Received in revised form 4 May 2022; Accepted 7 May 2022

A number of bioindicators have been suggested for monitoring soil quality such as: soil microbial biomass, soil animals, plants, and soil enzymes (Killham, 2002). Bioindicators should be commonly found and easily sampled, represent the local environmental conditions (as opposed to regional), and show some specific responses to target environmental variables that can be reliably measured. Protists fulfil these criteria (Payne, 2013). They are present in all agricultural soils worldwide. Progresses in metabarcoding technology make it possible to characterize most of their biodiversity in a reproducible and cost-effective way (Taberlet et al., 2018). Because soil protists are ecologically distinct from bacteria, fungi and other soil taxa (e.g. different physiology and ecological preferences), the information they provide on soil quality is complementary to that provided by other bioindicators. Their responsiveness to soil conditions – such as pH (Dupont et al., 2016), nitrogen level (Zhao et al., 2019, Zhao et al., 2020), soil moisture (Geisen et al., 2014), and quantity of pesticides (Fournier et al., 2020) – point toward a great potential for bioindication of agricultural soil ecosystem quality. In addition, protists are more sensitive than other microbial taxa to abiotic stresses such as changes in soil nitrogen fertilization (Zhao et al., 2019) or synthetic pesticides (Fournier et al., 2020). Such responsiveness to abiotic stresses caused by fertilization or pesticides – usually associated with detrimental effect on soil quality – reinforce the idea of using soil protists to identify several abiotic stressors affecting soil biodiversity and quality in agroecosystems. Despite its potential, the use of soil protists as bioindicator remains underexplored.

The association of metabarcoding data with supervised machine learning (SML) (Peters et al., 2014, Thessen, 2016) represents a promising approach in the context of bioindication and biomonitoring (Cordier et al., 2019). This approach allows for quantitative predictions of individual soil variables contrary to other bioindication methods that produce qualitative approximations. It has been successfully applied to protists in aquatic ecosystems to quantitatively predict several variables related to ecosystem quality (Cordier et al., 2017; Aylagas et al., 2018) (see Pawlowski et al., 2016 for a review). The great advantage of SML is that it is not constrained by strict statistical assumptions about the distribution of data and can fully exploit the large mass of data generated by metabarcoding by automatically sorting the ecological signal from the background noise. This approach can thus (i) identify responsive taxa without the need for strong taxonomic or ecological knowledge (taxonomy-free approaches: e.g. Apothéloz-Perret-Gentil et al., 2017) and (ii) produce quantitative predictions of individual variables relevant for soil quality. Here we assume that quantitative predictions can be used to identify a stress or a disruption of the structure and functioning of soil microbial communities resulting from agricultural soil management or changes in environmental conditions. We believe that the metabarcoding-SML quantitative approach can be implemented in agricultural soils to provide direct information about the response of soil microbes (here protists) that would complement classical soil abiotic monitoring based on in situ measurements.

Here, we test the metabarcoding-SML approach using a case study in Swiss vineyard soils. We used data from 132 soil samples collected over two years in 28 Swiss vineyards from a winegrowing region with relatively homogenous climatic conditions but differing in agricultural practices and soil characteristics. Our specific goals are 1) to assess which soil variables can be best predicted based on soil protists (i.e. identify variables that can cause a stress on soil protist communities); 2) to identify the taxa and diversity metrics that best predict these variables (i.e. identify responsive taxa and diversity metrics with high potential as bioindicators) and assess whether individual taxa and diversity metrics provide redundant or complementary information; and 3) to determine whether predictions are context-dependent.

2. Methods

2.1. Sampling sites and treatments

A total of 132 soil samples were taken in 28 vineyards located between 500 m and 800 m a.s.l. in Wallis (Switzerland) in 2016 (N = 74) and 2017 (N = 58). With ca. 4,800 ha of vineyards, the region is the highest producer of wine in Switzerland. This region has a continental climate with Mediterranean influences (mean annual temperature ~ 9.2 °C; mean annual precipitations ~ 690 cm y^{-1} in Sion at ~ 500 m a.s.l.). The 28 vineyards have similar climatic conditions but differed in soil texture and chemistry, in the use of pesticide (principally copper content), and in the cover and diversity of the vegetation in the vineyard inter-rows. In 10 vineyards, vegetation in all inter-rows was removed by application of glyphosate-based herbicides 2–3 times a year. In 8 vineyards, the inter-rows were managed by vegetation removal in every other row by herbicide application. In 10 vineyards, spontaneous vegetation was maintained, and the vegetation was cut 2–4 times per season (Steiner et al., Personal communication).

2.2. Soil sampling

Soil sampling was carried out over two consecutive years on the 13th of June 2016 and the 26th of May 2017. Sampling was done in two adjacent rows, starting five meters from the edge of the south eastern end of each vineyard. Ten subsamples per inter-row were collected until 10 cm depth within a span of 8–10 m. Subsamples were pooled per inter-row, sieved with a 2 mm mesh and stored at 4 °C before DNA extraction, which was performed within 3–4 days after sampling.

2.3. Environmental variables

A total of 17 soil variables were selected based on their importance for vineyard soil functioning and their potential impact on soil microbial communities (Table 1). These variables were classified into five broad categories: vegetation-related variables, soil texture, soil chemistry, microbial variables, and pesticide use. The *bioavailable copper content* was determined by extraction using diethylenetriaminepentaacetic acid (DTPA). A total of 20 ml DTPA was added to 10 g of soil. The solution was stirred for 2 h, filtered, and analyzed using an atomic absorption spectrometer (Schaller 2000). *Soil microbial biomass* was estimated as the microbial carbon per gram of dried soil [μg $C_{microbial}$ g^{-1}] using substrate induced (400 μl 40% glucose solution) respiration and calculation of microbial biomass C (according to Beck et al., 1997 & Anderson and Domsch, 1978). *Basal respiration rate* was measured as the O_2

Table 1
List of the measured soil variables.

Variable category	Description	Short name
Vegetation	Plant cover [%]	P_cover
	Plant species richness	P_rich
	Plant diversity (Simpson index)	P_sim
	Plant diversity (Shannon index)	P_sha
	Vegetation treatment (three categories: herbicide, vegetation removal, spontaneous vegetation)	Veg_trt
	Age of the vegetation cover [years]	age_veg
Soil texture	Clay content [%]	Clay
	Sand content [%]	Sand
	Stone content [%]	Stone
Soil chemistry	pH	pH
	Nitrogen content unit [%]	N
	Carbon content unit [%]	Ctot
	Soil organic matter content [%]	SOM
	C/N ratio	CN
	Soil moisture content [%]	H2O
Microbial	Basal respiration [μg O_2 h^{-1} g^{-1}]	BR
Pesticide	Copper concentration [mg/kg]	Cu

consumption per hour per gram of dry soil [μg O_2 h^{-1} g^{-1}] following the method of Scheu (1992). *Total carbon and nitrogen* content were measured using the Dumas combustion technique (Dumas, 1826; see Bremner and Mulvaney, 1982) with the "vario MAX CNS" analyser of the ELEMETA company. *C/N ratio* is the ratio of the total carbon content over the total nitrogen content. *Soil organic matter content* (SOM) was derived from the total soil carbon content. For calcareous soils (pH < 6.9), the calcium-carbonate fraction (i.e. inorganic carbon) was determined and subtracted from the total carbon content. *Soil moisture content* was estimated using 3.5 g of soil as the weight loss after 12 h drying at 80 °C. *Soil pH* was measured from a solution of 10 g of dried soil mixed with 25 ml of 0.01 M $CaCl_2$ (ratio = 1:1.5). Prior to measurement, the solution was stirred for 5 min and left for three hours to allow the suspended solids to settle. *Soil clay (*2 and 0.02 mm*), sand (*2 and 0.02 mm*) and stone contents* [%] were estimated as the relative weight of these soil fractions. Finally, the *percentage of soil covered by vegetation* was visually estimated in 1 m^2 quadrats. *Plant species richness* was estimated as the total number of species present in each individual plot. And *plant species diversity* was estimated using both the Shannon (lower importance of rare species) and Simpson (higher importance of rare species) diversity indices. Two variables describing vegetation management in each vineyard were derived from interviews with the winemakers. *Vegetation treatment* is a categorical variable representing how vegetation was managed in the inter-rows of each vineyard: herbicide, alternate herbicide, spontaneous vegetation. The *age of the vegetation cover* is the number of years the vegetation in the inter-rows has been present.

2.4. DNA extraction, amplification and sequencing

DNA was extracted from 0.25 g of homogenized mixed samples with the DNeasyPowerSoil Kit ® following the provided protocol (QIAGEN N. V., Netherlands). DNA extracts were quantified using a Nano Drop 1000 Spectrophotometer (Thermo Fisher Scientific, USA), stored at −20 °C, and sent to McGill University, Génome Québec Innovation Center, for PCR amplification and sequencing. The V9 SSU rRNA hypervariable region was amplified with the general eukaryotic primer pair 1380f/1510r (Amaral-Zettler et al., 2009). Paired-end DNA sequencing was performed on an Illumina MiSeq platform. Sequencing data and metadata are stored at the European Nucleotide Archive retrievable, upon acceptance of this article, with the accession number PRJEB32992.

2.5. Sequence data processing and taxonomic assignment

The absence of sequencing primers in the dataset was verified using cutadapt (Martin, 2011). The analysis of the reads was then done using the Divisive Amplicon Denoising Algorithm (DADA2) software (Callahan et al., 2016). The DADA2 pipeline infers exact amplicon sequence variants (ASVs) from sequencing data using the following steps: filtering, dereplication, sample inference, chimera identification, and merging of paired-end reads. We then assigned taxonomy to the ASVs with QIIME2 (Bolyen et al., 2019) using a pre-trained Naive Bayes classifier (Wang et al., 2007) and the Silva database (Ref NR 99, release 132) for protists (Quast et al., 2012). ASVs which were not assigned to protists were discarded. Through this approach, we obtained a total of 714′516 protist sequences classified into 2298 distinct ASVs. The main groups of protists were characterized by a relative abundance of around 35% of Alveolata, 32% of Rhizaria, 14% of Stramenopiles, 10% of Amoebozoa, and 6% of Chloroplastida.

2.6. Data analyses

We first calculated 23 metrics representing different aspects of protist alpha diversity: diversity, richness, evenness, rarity, and dominance (R package Microbiome (Lahti et al., 2017); see Table S1 for further explanations). We selected metrics that are relevant for each diversity aspects and that differ in the importance given to the number of taxa present and the total number and abundance of rare and/or dominant taxa (Table S1). We then built and tested various machine learning models using protist ASV occurrence and diversity metrics as predictors of soil variables: *k-nearest neighbors regression* (Hechenbichler and Schliep, 2004), Support Vector Machines with Linear Kernel (svmLinear, R package kernlab; Karatzoglou et al., 2004), extreme Gradient Boosting (xgbdart; R package xgboost; Chen and Guestrin, 2016), cubist (R package Cubist; see e.g. Kotsiantis et al., 2007 for a review), and Neural network regression (R nnet; Venables and Ripley, 2002). Environmental data were scaled and centered before analysis. Cu, M_bio, P_bio, C_tot, and H_2O were log-transformed prior to analyses to facilitate the visualization of the results. ASVs and diversity metrics with near zero variance or very high correlation were removed from the dataset as they do not improve the quality of the predictions. Data splitting was carried out in two ways. We first divided the data into a training (70%) and a test (30%) set by randomly sampling an equal number of data points above and below the median of the response variable (hereafter random splitting). We then used the 2016 data as the train set and the 2017 data as the test set (temporal splitting). Bootstrap resampling was used for model training and tuning. We also compared the predictive performances of models based on ASV alone, diversity metrics alone, or both to assess whether individual taxa abundance and diversity metrics provide complementary information.

The correlations between observed values and values predicted based on protists were measured using Kendall's Tau where 0 represents no relationship and 1 represents perfect correlation. This metric provided an estimation of the strength of the response of protists to specific soil characteristics that reflects the general bioindication potential of protists for a specific variable. The contribution of all soil taxa and diversity metrics to these predictions was estimated using importance measures (mean decrease in accuracy). These importance metrics identify the taxa and diversity metrics showing a strong relationship to specific variables and having an important contribution to the model. Ranking the taxa or diversity metrics based on importance metrics allows us to identify the best potential bioindicators.

To evaluate the dependency of predictive performance on the inherent variability of the measured soil variables, the tau values were correlated with the standard deviations of each soil variable. A positive correlation means that performance increases with variability whereas a negative one indicates the opposite. To explore the potential transferability of the method to other geographical and agricultural context, the correlations between model residuals and longitude, elevation, and grape variety were calculated. Strong correlation suggest that the suitability of the method depends on geographic location and/or grape variety. Grape variety represents slightly different agricultural practices and microclimatic conditions as winemakers optimize the variety given local conditions. Because climatic conditions in our study change more strongly with longitude than latitude, we focused only on longitude in this analysis. Data preprocessing, model calibration, and model validation were done using the R package 'Caret' (Kuhn, 2008). All analyses were done in R version 4.0.2 (R Development core team, 2019).

3. Results

3.1. Which soil variables can be predicted based on soil protists?

Correlations between predicted and observed values (Tau) were overall high but dependent on the type of variables. Focusing on the best model for each variable, the correlations between observed and predicted values based on test data were all significant ($P < 0.01$) with Kendall's Tau values ranging between 0.24 and 0.62 (Fig. 1; see Table S2 for the detailed results of each model). Edaphic variables were the best predicted variables with soil moisture (Tau = 0.62), basal respiration (Tau = 0.57), pH (Tau = 0.54), and copper concentration (Tau = 0.52) showing the highest Tau values. To the contrary, vegetation-related variables showed the lowest Tau values (plant Shannon: Tau = 0.27;

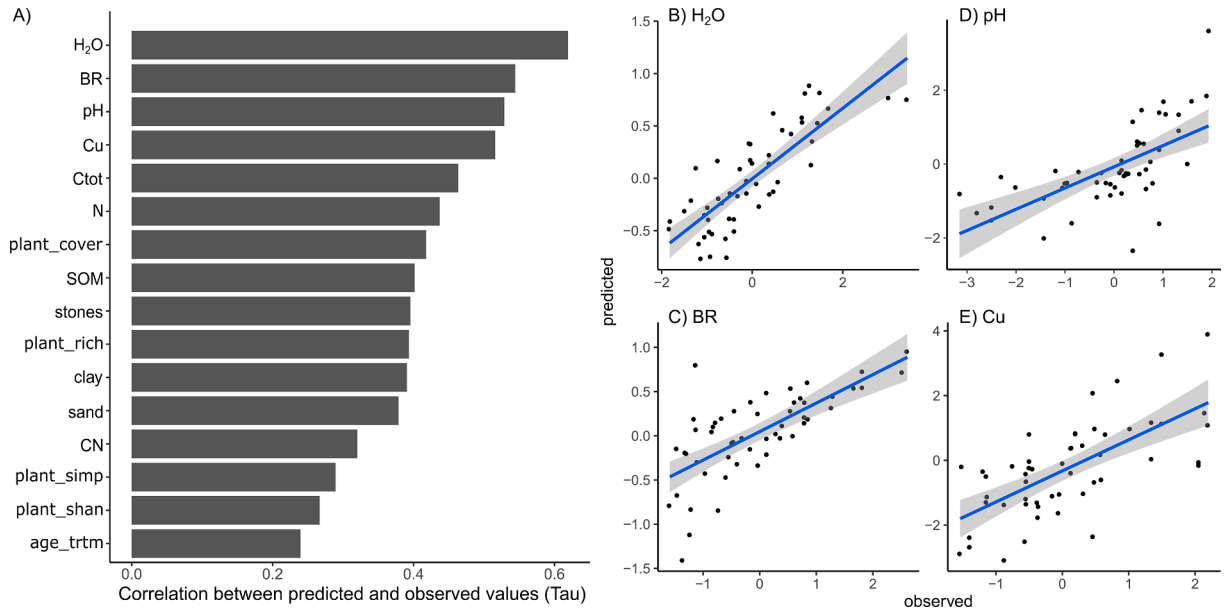


Fig. 1. Protist-based predictions of environmental variables. (A) Coefficients of correlation (Kendall's Tau) between observed and predicted values. All correlations were significant ($p < 0.01$). See descriptions of the abbreviations of the variable names in the Table 1. (B-E) Scatterplots of observed *versus* predicted values for the four best predicted variables. Linear regression lines are fitted, and the grey area represents 95% confidence intervals.

plant Simpson: Tau = 0.29; and time since the vegetation cover was installed: Tau = 0.24). However, this contrasts with vegetation treatment (Veg_trt), the only categorical variable in our dataset, that showed a kappa value of 0.62 with an accuracy of 0.81, a sensitivity of 0.79, and a specificity of 0.83.

3.2. *Which taxa and diversity metrics best predict soil variables?*

Rhizaria and Alveolata were the two taxonomic groups showing the strongest correlations with environmental variables (based on importance metrics; Fig. 2A). Diversity metrics, especially those metrics related to evenness and rarity, showed significant correlations as well (Fig. 2B). In addition to these general trends, importance metrics allowed identifying the most responsive individual taxa and diversity

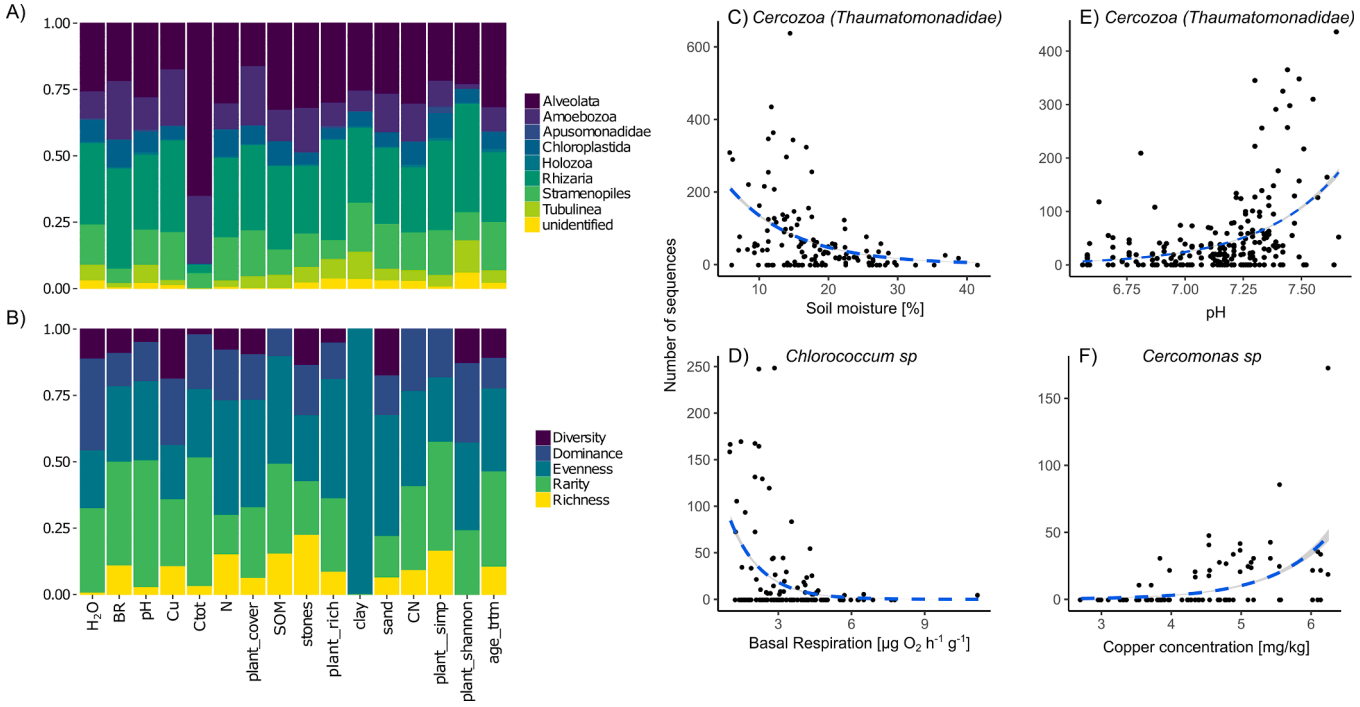


Fig. 2. Importance of (A) ASV and (B) diversity metrics as predictors of agricultural soil variables. Variable importance was measured by the residual sum of squares (RSS) in models based on various machine learning algorithms. (C-F) Relationship between the four best predicted soil variables and the respective most important predictors. Dashed blue lines are local weighted regressions (loess). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

metrics (Fig. 2C-F). For example, increased soil moisture was associated with a decrease in the relative abundance of an ASV within the clade Thaumatomonadidae (Fig. 2C). This ASV is thus a potential bioindicator of wetter soil conditions. Increased basal respiration was associated with a decrease in the relative abundance of ASV identified as Chlorococcum (Fig. 2D). Higher pH values were associated with an increase in the relative abundance of another ASV within the clade Thaumatomonadidae. Finally, increased copper concentration was associated with an increase in an ASV within the clade Cercomonas (Fig. 2F). Further information about the responsiveness of species ASV and diversity metric are given in Fig. S2-4. Our results also reveal a high variability in the correlations between taxa abundance and environmental variables among and within taxonomic groups (Fig. S5).

The complementarity between ASV and diversity metrics was low. Comparisons of models based on ASV alone and biodiversity metrics alone showed that ASV have a higher potential as bioindicators (42 % decrease in Tau values when using diversity metrics alone; $P < 0.001$; Fig. 3A). The combined use of ASV and diversity metrics resulted in a marginal increase in Tau values as compared to models based on ASV alone (7%; $P = 0.33$; Fig. 3A). Only few variables such as pH, vegetation cover, and plant richness were better predicted by diversity metrics (Fig. 3B-C).

3.3. Are protist-based predictions context-dependent?

The comparisons of the performances obtained using randomly split of training and test data to a temporal split revealed a significant decrease of 23 % ($P < 0.001$) in the correlation between observed and predicted values (Fig. 4A) when considering all variables and models. However, most variables showed no to very modest decrease in predictive performance (<10%) with the exception of soil moisture that showed the most important decrease (Fig. 4B).

The analysis of model residuals revealed no significant effect of elevation ($p = 0.11$) and longitude ($p = 0.56$) on predictive performances, but a significant effect of grape variety ($p = 0.003$) where residuals were on average 10% larger for "Pinot noir" ($N = 121$) than for "Chasselas" ($N = 115$). These trends varied among environmental variables. For example, in the case of soil moisture, pH and copper concentration, model residuals were not significantly correlated to either of grape variety ($p = 0.06$; $p = 0.15$; and $p = 0.36$, respectively), elevation ($p = 0.1$; $p = 0.3$ and $p = 0.6$) or longitude ($p = 0.64$; $p = 0.9$ and $p = 0.38$) whereas, for basal respiration, model residuals decreased

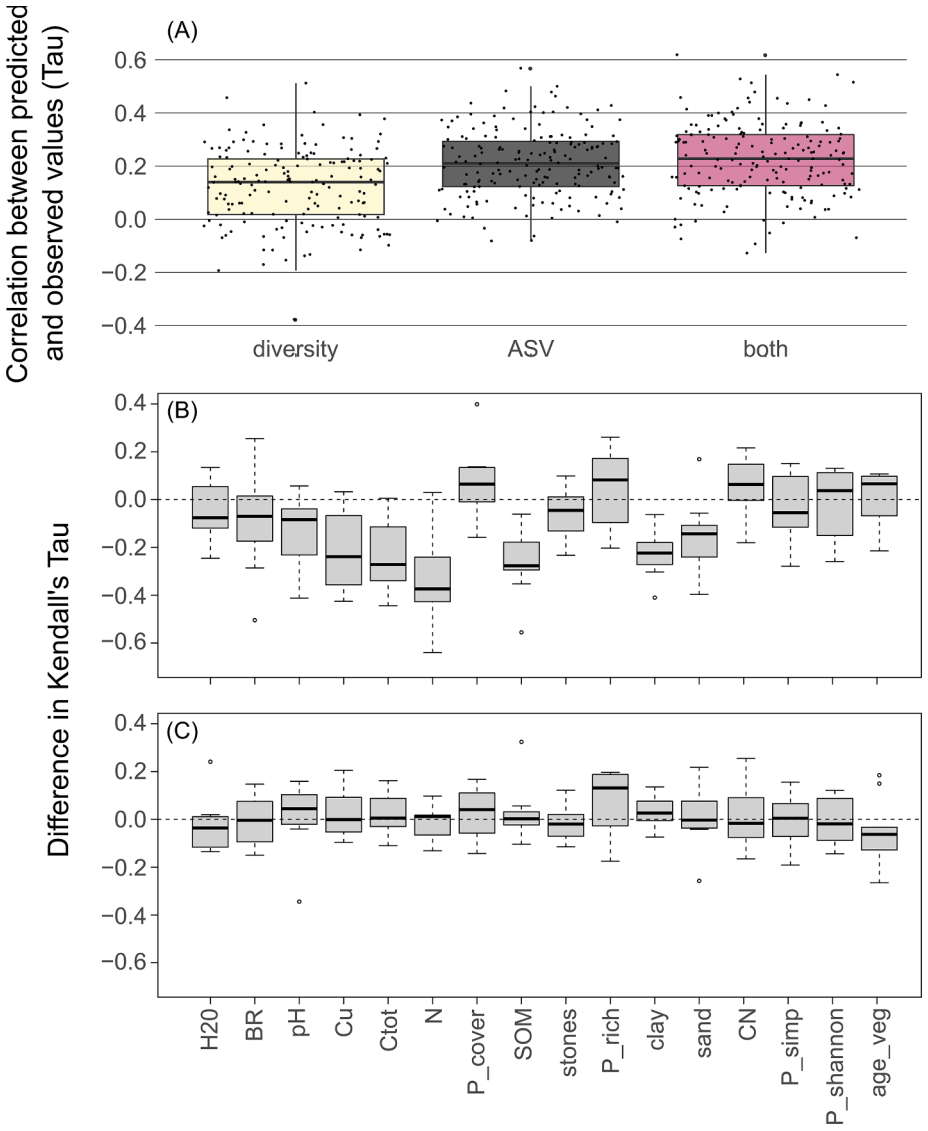


Fig. 3. Complementarity among ASV abundance and diversity metrics. (A) Model predictive performances as a function of different predictors (diversity metrics, individual ASV, or both). (B-C) Differences in model predictive performance (Tau) between (B) models using only ASV as predictors *versus* using only diversity metrics (baseline = ASV), and (C) models using only ASV as predictors *versus* using both ASV and diversity metrics (baseline = ASV).

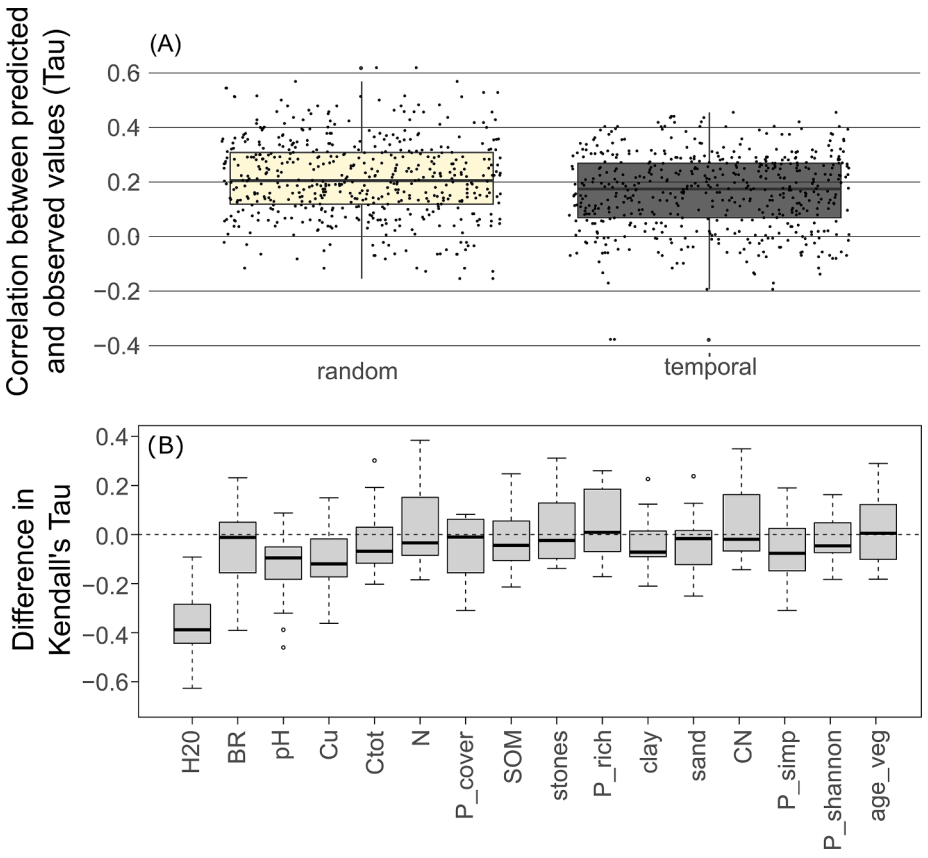


Fig. 4. Model temporal transferability. (A) Model predictive performances as a function for different splitting of the data into training and test sets (random or temporal). (B) Differences in model predictive performance (Tau) between random splitting of the data *versus* temporal splitting (baseline = random splitting) for each variable.

significantly with elevation ($p = 0.04$) but were not significantly correlated with longitude ($p = 0.16$) or grape variety ($p = 0.08$).

Finally, predictive performances were not correlated with environmental variability. The correlation between the standard deviation and Kendall's Tau values of each variable was non-significant ($p = 0.59$).

4. Discussion

Our results largely confirm the suitability of soil protist community composition and diversity as bioindicators of biotic stress in agricultural soils. They highlight the high variability in the responses of individual taxa and diversity metrics to changes in the soil ecosystem. This variability allowed the prediction of a broad range of variables relevant for soil quality assessments. Protist community composition and diversity were particularly responsive to soil copper content, pH, soil moisture, and the activity of soil microbes (basal respiration). Predictions of these variables based on protists can thus reflect stresses on the soil microbial food web related, for example, to the use of pesticides (copper), drought (soil moisture), soil acidification (pH), or decrease in microbial activity (basal respiration). Our results also suggest that our approach is able to identify small differences in environmental conditions and has a generally low but variable-specific sensitivity to different temporal, spatial or environmental contexts.

4.1. Protists exhibit a great bioindication potential across multiple abiotic stressors

Our results based on the metabarcoding-SML approach show that changes in protist communities can be used to quantitatively predict the selected environmental variables, confirming protist responsiveness to multiple stressors of the soil ecosystem. The correlations between observed and predicted values were all significant which further supports the idea that protists have a great bioindicator potential. The correlations ranged between 0.24 and 0.62 (Tau), and were highest for copper concentration, soil moisture, microbial activity, and soil chemistry. These predictions are based on the observed responses of soil protists to the selected environmental variables which were taxa-specific and agreed with previous studies about the impact of viticulture on soil microbial communities. For example, several studies pointed at the negative impact of Cu on soil microbes (e.g. Ekelund et al., 2003; Du Plessis et al., 2005). In our study, increased copper concentration showed negative correlations with protist richness and diversity metrics suggesting an overall negative impact of copper on protist communities. Despite this general negative impact of copper, several taxa such as flagellate ASVs of the genus *Cercomonas* (Rhizaria) responded positively to increased copper concentration suggesting the existence of a broad range of tolerance to copper among protist taxa. Increased pH led to a decrease in evenness (Fig. S3) as observed in Öztoprak et al. (2020). As for other environmental variables, our results highlight an important variability in the effect of pH among taxa. For example, most taxa within the clade Cavosteliida were positively associated with acidic conditions whereas a majority of taxa within the clades Alveolata and Tubulinea were positively associated with alkaline soils (Fig. S5). All other variables provided similar results confirming the variability in ecological strategies among protists and their potential as bioindicators of agricultural soil quality.

Individual ASVs, especially within the clades Rhizaria and Alveolata, were overall better predictors of soil variables than the diversity metrics. The complementarity between individual ASV and diversity metrics was low suggesting that the information provided by diversity metrics is largely redundant to that provided by individual ASVs. Nevertheless, a few variables such as pH, vegetation cover, and plant richness were

better predicted by diversity metrics suggesting a direct impact of these variables on microbial diversity. Overall, our results confirm that protists ASV and, to a lesser extent, diversity metrics can provide quantitative information on a broad range of relevant soil variables. This, in turn, enables the identification of multiple potential stressors of agricultural soil microbial diversity (here based on protists).

4.2. The environmental context has limited influence on protist-based predictions

Our results revealed an overall low context-dependency of the predictions based on protists which highlights the potential transferability of the proposed approach to other agricultural and environmental contexts. Our approach using protist metabarcoding data and machine learning algorithms was able to distinguish small differences in soil conditions and is relatively robust to large temporal variability and different spatial and environmental contexts. For instance, the lack of correlation between predictive performances (Tau) and variable standard deviations shows that the predictive capacity of our approach was independent from the variability of the selected variables. For example, the range of pH values was narrow (between 6.7 and 7.6), but pH was among the best predicted variables.

With respect to inter-annual variability, temporal instead of random data splitting leads to a 23% decrease in predictive power which can be considered emodest given the inter-annual variability in our data (e.g. 43% decrease in soil moisture and 33% decrease in basal respiration in 2017 compared to 2016; see Fig. S7). The highest decrease was observed for soil moisture with Tau values decreasing by about 0.4. This result is not surprising given the high temporal variability in soil moisture between the two years. However, most of the other variables showed no to very modest decrease in predictive performances. Among them, basal respiration, for example, also showed an important inter-annual variability. This suggests that, for most of the tested variables, inter-annual variability in soil conditions does not influence the response of soil protists.

With respect to environmental and spatial transferability, our results point toward an increase in prediction error (as shown by the analysis of model residuals) in vineyards with the "Pinot noir" variety as compared to vineyards with the "Chasselas" variety. This effect was, however, limited (10% increase in model residuals) and no effect of longitude was observed. Our results thus suggest that, although some of the bioindicators identified might be specific to a particular grape variety and/or its associated environmental conditions, a majority can be used across different regions and grape varieties.

4.3. Perspectives

Our analysis raises several questions related to the effectiveness of using protists as bioindicators of stress in agricultural soils. *Can the identified bioindicators be used over much larger spatial and temporal scales?* The temporal and spatial extent of our study is limited, and the distribution range and temporal variability of the identified bioindicators is still little documented. As a consequence, the extent to which the identified bioindicators can be used over larger spatial scales (continental to global) and/or over longer time periods is still uncertain. Further studies are needed to tease apart context-specific bioindicators from those that can be used over broader latitudinal gradients, longer time periods, and across different types of cultures. *What are the ecological mechanisms driving the association between potential bioindicators and environmental variables?* Correlations between taxon abundance and environmental variables can reflect a direct impact (e.g. physiological processes can only be completed within a given range of pH), an indirect association (e.g. a predator is indirectly impacted through changes in abundance of its preferred prey), or spurious correlations (no causative interactions). In addition, the absence of a taxon might be due to processes unrelated to environmental variables such as dispersal limitations. In such a case, including zeros in the analyses can lead to an underestimation of the importance of the environmental factor. However, dispersal limitations are unlikely to play a key role in our study because of the relatively small spatial extent considered and using a presence-only approach should solve this potential issue when applying the approach over larger spatial extents.

5. Conclusion

Soil protists are a key component of the microbial food web that can be strongly impacted by agricultural practices and changes in soil conditions. An approach combining metabarcoding data with machine learning can provide quantitative information on these impacts thereby identifying a potential stress on the soil microbial foodweb. This approach thus offers an integrative and quantitative way to identify several stressors of microbial communities and provides complementary information to existing soil monitoring tools. Overall, the proposed approach can improve our understanding of agricultural soil ecosystems and help limit the impact of anthropogenic activities on soil quality through the development of more sustainable agricultural practices.

CRediT authorship contribution statement

Bertrand Fournier: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft. **Magdalena Steiner:** Conceptualization, Formal analysis, Writing – review & editing. **Xavier Brochet:** Methodology, Writing – review & editing. **Florine Degrune:** Validation, Writing – review & editing. **Jibril Mammeri:** Methodology, Writing – review & editing. **Diogo Leite Carvalho:** Methodology, Writing – review & editing. **Sara Leal Siliceo:** Methodology, Writing – review & editing. **Sven Bacher:** Project administration, Supervision, Conceptualization, Writing – review & editing. **Carlos Andrés Peña-Reyes:** Supervision, Methodology, Writing – review & editing. **Thierry J. Heger:** Project administration, Supervision, Conceptualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge financial support from HES-SO (project 78046, MaLDIveS) to TH and CRP, the Swiss Federal Office for the Environment (00.5005.PZ / A58E8CC1A; 00.5005.PZ / 3A97E39C8; 19.0061.PJ.PZ and D-91173401/988, MiDiBo_2) to TH; from the DFG (FO 1420/1-1; project FunShift) and the WISNA program from the German Federal Ministry of Education and Research to BF; and from the PromESSinG project funded through the 2013-2014 BiodivERsA/FACCE-JPI joint call for research proposals and the Swiss National Science Foundation [Grant Number 40FA40_158390] to SB. We thank Frédéric Lamy, Matteo Mota and Dorothea Noll for thoughtful discussions and data interpretation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2022.108955.

References

Amaral-Zettler, L.A., McCliment, E.A., et al., 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS One 4 (7), e6372.

Anderson, J.P., Domsch, K.H., 1978. A physiological method for the quantitative measurement of microbial biomass in soils. Soil Biol. Biochem. 10 (3), 215–221.

Apothéloz-Perret-Gentil, L., Cordonier, A., Straub, F., Iseli, J., Esling, P., Pawlowski, J., 2017. Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. Mol. Ecol. Resour. 17 (6), 1231–1242.
Aylagas, E., Borja, Á., Muxika, I., Rodríguez-Ezpeleta, N., 2018. Adapting metabarcoding-based benthic biomonitoring into routine marine ecological status assessment networks. Ecol. Ind. 95, 194–202.
Beck, T., Joergensen, R.G., et al., 1997. An inter-laboratory comparison of ten different ways of measuring soil microbial biomass C. Soil Biol. Biochem. 29 (7), 1023–1032.
Bolyen, E., Rideout, J.R., et al., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37 (8), 852–857.
Bremner, J.M., Mulvaney, C., 1982. Nitrogen—Total 1. Meth. Soil Anal. Part 2. Chem. Microbiol. Properties 595–624.
Callahan, B.J., McMurdie, P.J., et al., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods 13 (7), 581–583.
Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.
Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T., Pawlowski, J., 2017. Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning. Environ. Sci. Technol. 51 (16), 9118–9126.
Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., Pawlowski, J., 2019. Embracing environmental genomics and machine learning for routine biomonitoring. Trends Microbiol. 27 (5), 387–397.
Du Plessis, K.R., Botha, A., Joubert, L., Bester, R., Conradie, W.J., Wolfaardt, G.M., 2005. Response of the microbial community to copper oxychloride in acidic sandy loam soil. J. Appl. Microbiol. 98 (4), 901–909.
Dupont, A.Ö.C., Griffiths, R.I., Bell, T., Bass, D., 2016. Differences in soil micro-eukaryotic communities over soil pH gradients are strongly driven by parasites and saprotrophs. Environ. Microbiol. 18 (6), 2010–2024.
Ekelund, F., Olsson, S., Johansen, A., 2003. Changes in the succession and diversity of protozoan and microbial populations in soil spiked with a range of copper concentrations. Soil Biol. Biochem. 35 (11), 1507–1516.
Fournier, B., Pereira Dos Santos, S., Gustavsen, J.A., Imfeld, G., Lamy, F., Mitchell, E.A. D., Mota, M., Noll, D., Planchamp, C., Heger, T.J., 2020. Impact of a synthetic fungicide (fosetyl-Al and propamocarb-hydrochloride) and a biopesticide (*Clonostachys rosea*) on soil bacterial, fungal, and protist communities. Sci. Total Environ. 738, 139635.
Geisen, S., Bandow, C., Jörg, R., Bonkowski, M., 2014. Soil water availability strongly alters the community composition of soil protists. Pedobiologia 57 (4–6), 205–213.
Giller, K.E., Beare, M.H., Lavelle, P., Izac, A.-M.-N., Swift, M.J., 1997. Agricultural intensification, soil biodiversity and agroecosystem function. Appl. Soil Ecol. 6 (1), 3–16.
Hechenbichler, K., Schliep, K., 2004. Weighted k-nearest-neighbor techniques and ordinal classification. Collaborative Research Center 386, Discussion Paper 399. DOI: 10.5282/ubm/epub.1769.
Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. kernlab - An S4 Package for Kernel Methods in R. J. Stat. Softw. 11, 1–20.
Killham, K., Staddon, W.J., 2002. Bioindicators and sensors of soil health and the application of geostatistics. In: Enzymes in the environment. Marcel Dekkerr, NY, USA, pp. 391–405.
Kotsiantis, S.B., Zaharakis, I., Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. Emerg. Artif. Intell. Appl. Comput. Eng. 160, 3–24.
Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. J. Stat. Softw. 28, 1–26.
Lahti, L., Shetty, S., Blake, T., Salojarvi, J., 2017. Tools for microbiome analysis in R. Version 1 (5), 28.
Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17 (1), 10–12.
Öztoprak, H., Walden, S., Heger, T., Bonkowski, M., Dumack, K., 2020. What drives the diversity of the most abundant terrestrial cercozoan family (Rhogostomidae, Cercozoa, Rhizaria)? Microorganisms 8, 1123.
Pawlowski, J., Lejzerowicz, F., Apotheloz-Perret-Gentil, L., Visco, J., Esling, P., 2016. Protist metabarcoding and environmental biomonitoring: Time for change. Eur. J. Protistol. 55, 12–25.
Payne, R.J., 2013. Seven reasons why protists make useful bioindicators. Acta Protozoologica 52 (3).
Peters, D.P., Havstad, K.M., Cushing, J., Tweedie, C., Fuentes, O., Villanueva-Rosales, N., 2014. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. Ecosphere 5, 1–15.
Quast, C., Pruesse, E., et al., 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41 (D1), D590–D596.
R Development core team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Scheu, S., 1992. Automated measurement of the respiratory response of soil microcompartments: active microbial biomass in earthworm faeces. Soil Biol. Biochem. 24 (11), 1113–1118.
Taberlet, P., Bonin, A., Zinger, L., Coissac, E., 2018. Environmental DNA: For biodiversity research and monitoring. Oxford University Press.
Thessen, A., 2016. Adoption of machine learning techniques in ecology and earth science. One Ecosystem 1, e8621.
Venables, W., Ripley, B., 2002. Modern applied statistics (Fourth S., editor). Springer, New York.
Wagg, C., Bender, S.F., Widmer, F., van der Heijden, M.G.A., 2014. Soil biodiversity and soil community composition determine ecosystem multifunctionality. PNAS 111 (14), 5266–5270.
Wang, Q., Garrity, G.M., et al., 2007a. Native bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73 (16), 5261–5267.
Wang, Y., Shi, J., Wang, H., Lin, Q., Chen, X., Chen, Y., 2007b. The influence of soil heavy metals pollution on soil microbial biomass, enzyme activity, and community composition near a copper smelter. Ecotoxicol. Environ. Saf. 67, 75–81.
Zhao, Z.-B., He, J.-Z., Geisen, S., Han, L.-L., Wang, J.-T., Shen, J.-P., Wei, W.-X., Fang, Y.-T., Li, P.-P., Zhang, L.-M., 2019. Protist communities are more sensitive to nitrogen fertilization than other microorganisms in diverse agricultural soils. Microbiome 7 (1).
Zhao, Z.-B., He, J.-Z., Quan, Z., Wu, C.-F., Sheng, R., Zhang, L.-M., Geisen, S., 2020. Fertilization changes soil microbiome functioning, especially phagotrophic protists. Soil Biol. Biochem. 148, 107863.

Further reading

Bonkowski, M., Clarholm, M., 2012. Stimulation of plant growth through interactions of bacteria and protozoa: testing the auxiliary microbial loop hypothesis. Acta Protozoologica 51, 237–247.