



# Predicting soil fungal communities from chemical and physical properties

Natacha Bodenhausen<sup>1,2</sup>  | Julia Hess<sup>2</sup> | Alain Valzano<sup>2</sup> |  
 Gabriel Deslandes-Héroid<sup>3</sup> | Jan Waelchli<sup>4</sup> | Reinhard Furrer<sup>5,6</sup> |  
 Marcel G. A. van der Heijden<sup>2,7</sup> | Klaus Schlaeppi<sup>2,3,4</sup> 

<sup>1</sup>Department of Soil Sciences, Research Institute of Organic Agriculture (FiBL), Frick, Switzerland

<sup>2</sup>Department of Agroecology and Environment, Agroscope, Zürich, Switzerland

<sup>3</sup>Institute of Plant Sciences, University of Bern, Bern, Switzerland

<sup>4</sup>Department of Environmental Sciences, University of Basel, Basel, Switzerland

<sup>5</sup>Department of Mathematics, University of Zürich, Zürich, Switzerland

<sup>6</sup>Institute of Computational Science, University of Zürich, Zürich, Switzerland

<sup>7</sup>Department of Plant and Microbial Biology, University of Zürich, Zürich, Switzerland

## Correspondence

Natacha Bodenhausen.  
 Email: [Natacha.bodenhausen@fibl.org](mailto:Natacha.bodenhausen@fibl.org)

## Funding information

Universität Basel, Grant/Award Number: core funding; Gebert RUF Foundation, Grant/Award Number: GRS-072/17

## Abstract

**Introduction:** Biogeography describes spatial patterns of diversity and explains why organisms occur in given conditions. While it is well established that the diversity of soil microbes is largely controlled by edaphic environmental variables, microbiome community prediction from soil properties has received less attention. In this study, we specifically investigated whether it is possible to predict the composition of soil fungal communities based on physicochemical soil data using multivariate ordination.

**Materials and Methods:** We sampled soil from 59 arable fields in Switzerland and assembled paired data of physicochemical soil properties as well as profiles of soil fungal communities. Fungal communities were characterized using long-read sequencing of the entire ribosomal internal transcribed spacer. We used redundancy analysis to combine the physical and chemical soil measurements with the fungal community data.

**Results:** We identified a reduced set of 10 soil properties that explained fungal community composition. Soil properties with the strongest impact on the fungal community included pH, potassium and sand content. Finally, we evaluated the model for its suitability for prediction using leave-one-out validation. The prediction of community composition was successful for most soils, and only 3/59 soils could not be well predicted (Pearson correlation coefficients between observed and predicted communities of <0.5). Further, we successfully validated our prediction approach with a publicly available data set. With both data sets, prediction was less successful for soils characterized by very unique properties or diverging fungal communities, while it was successful for soils with similar characteristics and microbiome.

**Conclusions:** Reliable prediction of microbial communities from chemical soil properties could bypass the complex and laborious sequencing-based generation of microbiota data, thereby making soil microbiome information available for agricultural purposes such as pathogen monitoring, field inoculation or yield projections.

## KEYWORDS

leave-one-out validation, microbiome, prediction, soil fertility, soil properties

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Journal of Sustainable Agriculture and Environment* published by Global Initiative of Crop Microbiome and Sustainable Agriculture and John Wiley & Sons Australia, Ltd.

## 1 | INTRODUCTION

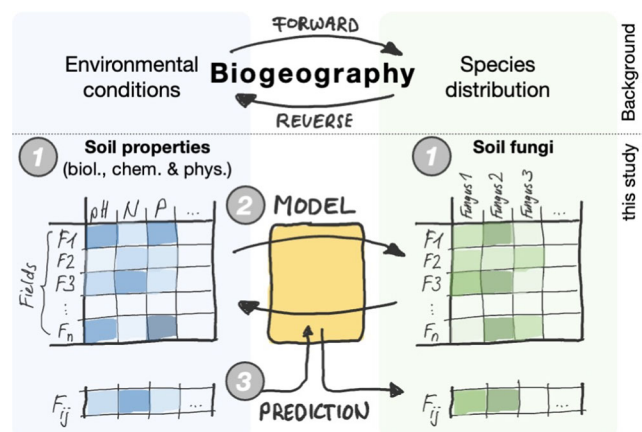
Soils host a large community of microbes with an estimated diversity of several thousands of species per gram of soil (Roesch et al., 2007). The global microbial biomass (bacteria, fungi, archaea, protists and viruses) is estimated to be >1000 kg per hectare (Fierer, 2017). Soil microorganisms participate in important soil functions such as nutrient cycling and soil carbon sequestration, thereby contributing to soil fertility. Fungal communities notably play a key role in the decomposition of organic matter, thus driving the carbon cycle. In addition to their role as saprotrophs, soil fungi include symbiotic organisms such as mycorrhizal fungi or major crop pathogens, thus contributing to plant growth and soil, plant and animal health (Banerjee & van der Heijden, 2023).

Biogeography studies patterns of species diversity over space and times (Martiny et al., 2006). Biogeography seeks to understand the relationship between ecological factors and species distributions (Figure 1, Section 'Background'). Similar to plants and animals, biogeographical patterns are also studied for microorganisms, which follow distribution principles other than those of larger organisms (Martiny et al., 2006). While the biogeography of macroorganisms is often dependent on spatial distance or dispersal, microbial biogeography is often controlled by the physical and chemical variables of the immediate environment (Fierer & Jackson, 2006). Greatly facilitated with modern DNA sequencing, the main drivers of global microbial biogeography have been demonstrated across different ecosystems around the globe (Delgado-Baquerizo et al., 2018; Tedersoo et al., 2014). Particular for soils, the immediate physicochemical environment, including pH, nutrient availability and humidity, comprises the critical drivers explaining the local microbial

diversity and richness (Chu et al., 2020). For instance, the biogeography of soil bacterial communities, even across different ecosystems, was largely explained by soil pH (Fierer & Jackson, 2006; Karimi et al., 2018). Karimi et al. defined the hierarchy of the main drivers of soil bacterial and archaeal diversity with soil pH > land management > soil texture > soil nutrients > climate (Karimi et al., 2018). By contrast, soil fungal biogeography is primarily driven by climate, followed by edaphic variables as well as by spatial drivers (Tedersoo et al., 2014). In addition to soil pH, prominent edaphic drivers of soil microbial biogeography across ecosystems are levels of organic carbon and redox status, which shape bacterial communities (Fierer, 2017), while calcium and phosphorus are the strongest drivers of fungal diversity (Tedersoo et al., 2014). Soil microbial diversity patterns also vary *within* ecosystems; for example, distinct microbial communities are found in the same field soils after different cropping management practices (Hartmann et al., 2015; Hartman et al., 2018). Such observations exemplify that the immediate physicochemical conditions also operate at a very fine scale to drive soil microbial biogeography, such as in agroecosystems.

Our motivation for this work is rooted in the vision that microbiome information of field soils, that is, abundance patterns of certain bacteria or fungi, can be deployed for agricultural purposes (Schlaeppli & Bulgarelli, 2015). Chemical soil information is often taken as a basis for decisions about agricultural management; for example, in Switzerland, farmers who receive government subsidies need to adjust fertilizer regimes based on soil nutrient and texture analyses (Richner & Sinaj, 2017). On the other hand, microbiome information on agricultural soils remains untapped for rational management decisions. For instance, the presence or absence of soil microbes could be used to predict the likelihood of pathogen development or symbiont establishment, to explain over- and underyielding fields or to estimate the efficiency of nutrient cycling. Obtaining the biological information of soils remains more complex and laborious than generating chemical soil data, which is standard practice in analytical soil laboratories. The analysis of biological soil parameters is more demanding because they are time-consuming in the laboratory (e.g., microbial biomass, microbial community analysis) or because they rely on experimental bioassays (e.g., microbial respiration, soil suppressiveness). A solution to obtain microbiological soil information is to exploit biogeographical principles and predict soil microbiomes from chemical data.

Biogeography can be interpreted in a *forward* or *reverse* manner (Figure 1, Section 'Background'). This distinction of *forward* versus *reverse* interpretation of biogeography becomes particularly important for prediction. Most studies cited above model the relationships between edaphic soil factors and soil microbiomes but often without performing actual species predictions from their models, for example (Karimi et al., 2018; Tedersoo et al., 2014). In those studies, the word 'prediction' is used for explaining biogeographic relationships, but not to predict species distribution from environmental conditions. By contrast, *reverse* prediction uses species distributions to explain the environmental conditions. In this case, soil microbiomes serve as an explanatory variable to predict, for example, crop



**FIGURE 1** Forward and reverse biogeography. Forward-based interpretation of biogeography explains species distribution from environmental conditions while reverse-based interpretation uses species distribution to explain the environment. In this study, we measured the biological, chemical and physical properties alongside with the fungal species in soil samples from different fields (Step 1). Next, we modelled the biogeography of soil fungi based on the soil properties (Step 2). Finally, we used this model for prediction of soil fungi from the soil property data of a test field (Step 3).

productivity (Chang et al., 2017), type of land use (Hermans et al., 2020) or ecosystem functions (Wagg et al., 2019). We propose to use microbial biogeography in a *forward* manner, that is, explaining species distribution from the environmental conditions. In that case, the soil microbiome is the response variable. To the best of our knowledge, no forward biogeography studies have been addressing predictions of soil fungal communities from soil properties.

Microbial ecology traditionally displays microbial diversity using ordination techniques to which the environmental factors can be fitted to describe the multifactorial relationships; see Kundel et al. (2020) or Chen et al. (2021) as examples. Other methods, including deep learning approaches, have been used to predict soil microbial community composition from soil properties (García-Jiménez et al., 2020); however, multivariate ordination has rarely been used for prediction. The environmental properties can participate in the ordination; in this case, the ordination is called 'constrained'. Redundancy analysis (RDA) is a commonly used constrained ordination technique that combines regression with principal component analysis (PCA) (Borcard et al., 2018). A key advantage of RDA is that covariates can be included in the model to account for confounding factors. This statistical method is called partial RDA (Borcard et al., 1992) and is the multivariate equivalent of partial linear regression (Borcard et al., 2018). For example, the researcher is interested to display the relationship between species composition and spatial variables, when the effect of temporal variables is held constant. Recently, RDA has been successfully used as a tool to predict genomic composition (Capblancq & Forester, 2021). Here, we explored the use of multivariate ordination to predict microbial community composition from environmental properties.

In this study, we tested the proof-of-principle that composition of soil fungal communities can be predicted based on measured physicochemical soil properties. For this, we sampled soils from 59 arable fields and characterized their fungal communities using long-read amplicon sequencing as well as their chemical, physical and biological properties. Using partial multivariate ordination, a reduced set of 10 variables was identified to model the relationship between soil properties and fungal community composition. Furthermore, we followed the same approach using a publicly available data set. Both models allowed successful prediction of species composition, except for soils with unique properties or few representatives in the data set.

## 2 | MATERIALS AND METHODS

### 2.1 | Soil sampling

Soil samples were collected from 22 fields in 2018, 25 fields in 2019 and 12 fields in 2020 (Supporting Information: Figure S1). These fields belong to farmers who agreed to participate in an inoculation experiment with arbuscular mycorrhizal fungi, which took place over three years (Lutz et al., Submitted). The exact global positioning system coordinates of those sites are available but not provided here for confidentiality reasons. In 2018, fields were sampled from April

23 to May 16. In 2019, fields were sampled from April 18 to June 7. In 2020, fields were sampled from April 22 to May 16. Soils were sampled before fertilization with a soil auger (Eijkkamp; diameter 3 cm, depth 20 cm). Approximately 20 soil cores were mixed to form a composite sample. In the laboratory, soils were sieved with a 2 mm sieve to remove stones and decomposing plant material. Subsamples for DNA extraction were stored at  $-20^{\circ}\text{C}$  until extraction. Soils were stored at  $4^{\circ}\text{C}$  for a maximum of 2 weeks before further processing.

### 2.2 | Soil analysis

Soil texture, cation-exchange capacity, base saturation, humus and water holding capacity were determined at the Environmental Analytics laboratory at Agroscope according to the Swiss reference methods (FAL R. FAW, 2004). Microbial biomass and respiration were measured by the Soil Biology Laboratory at Agroscope with chloroform fumigation extraction and substrate-induced respiration, respectively, both according to Swiss reference methods (FAL R. FAW, 2004). Nitrate and ammonium concentrations were determined photometrically according to the reference procedure (FAL R. FAW, 2004); Nmin is the sum of nitrate and ammonium. Macro- and micronutrients as well as pH were measured at Labor für Boden- und Umweltanalytik (Eric Schweizer AG) according to their standard protocols. Data are provided in Supporting Information: Table S1.

### 2.3 | Profiling fungal communities

DNA was extracted from approximately 250 mg soil with the NucleoSpin Soil kit (Macherey-Nagel) from four subsamples from each soil. Samples from each year were extracted independently and sequenced in a separate library. Finally, the DNA extracts from the 3 years were used as templates for polymerase chain reaction (PCR) to sequence a fourth library. For this library, the DNA from the four subsamples was pooled in equimolar ratios. After quantification with a Picogreen Quant-iT PicoGreen dsDNA Assay kit (Invitrogen), DNA was diluted to 1 ng/ $\mu\text{L}$ .

The PCR primers ITS1F (Gardes & Bruns, 1993) and ITS4 (White et al., 1990) were used to amplify the entire internal transcribed spacer (ITS) region for long-read sequencing (Bodenhausen et al., 2019). Amplicon libraries were prepared using a two-step PCR protocol. The first step amplifies the ITS region from genomic DNA, while the second step adds barcodes specific for each sample. PCRs were prepared in the same way for each library: 5 Prime Hot Master Mix (Quantabio Beverly) was used with a total reaction volume of 20  $\mu\text{L}$  with 0.3% BSA and 500 nM of each primer. The PCR programme consisted of an initial denaturation step of 2 min at  $94^{\circ}\text{C}$ , followed by 25 cycles of denaturation at  $94^{\circ}\text{C}$  for 45 s, annealing at  $55^{\circ}\text{C}$  for 1 min, and elongation at  $72^{\circ}\text{C}$  for 1 min with a final elongation step of 10 min at  $72^{\circ}\text{C}$ . Clean-up was followed by solid-phase reversible immobilization with SPRIselect beads (Beckman Coulter). The second PCR step consisted of the same PCR but

with barcoded ITS1F and ITS4 primers and without BSA and the same cycling programme but with only 10 steps. Finally, DNA was quantified with Picogreen and pooled in equimolar ratios. Negative controls were included for PCR and sequenced along with the other samples.

Four libraries were sequenced with the long-read sequencing technology Single Molecule, Real-Time (SMRT) of PacBio (Supporting Information: Table S2). The first SMRT library was sequenced on a Sequel I instrument at the Functional Genomics Centre in Zürich (<https://fgcz.ch>) according to standard PacBio protocols. That library was sequenced twice to obtain more coverage. The second and third libraries were sequenced on a Sequel I instrument, while the fourth library was sequenced on a Sequel IIe, all at the Next-Generation Sequencing Platform of the University of Bern (<https://www.ngs.unibe.ch>) according to standard PacBio protocols. The raw data were converted to circular consensus sequences (min. passes = 5) and demultiplexed with SMRT software (v. 9.0.0, Pacific Biosciences of California).

## 2.4 | Bioinformatics

Calculations were performed at the SciCORE (<http://scicore.unibas.ch/>) scientific computing centre at the University of Basel. The R statistical environment (R version 4.0.0) was used for data analysis (Team Rs, 2016). All the following steps were performed using the R package dada2 (1.16.0) (Callahan et al., 2016). Although all samples from each year were sequenced separately, they were analyzed together to form one operational taxonomic unit (OTU) table. After orienting all sequences in the same direction, primer sequences were removed. Sequences were quality filtered (max. expected errors: 2, min. length: 500 bp), truncated (after 1800 bp or at the first instance of a quality score <3) and dereplicated. Next, sequencing errors were denoised by the DADA algorithm using a parametric error model. A count table of amplicon sequence variants (ASVs) was created, chimeras were removed and ASVs were clustered by 97% similarity with the R package DECIPHER (v. 2.16.1). Finally, the naïve Bayesian classifier from the ribosomal database project RDP (Wang et al., 2007) was used to assign taxonomy based on the UNITE database (Nilsson et al., 2019) with `utax_reference_dataset_10.05.2021.fasta`. Because the assignments were ambiguous for some ASVs, we assigned the data with an alternative classifier (IDTAXA, minimal bootstrap = 40%) again using the UNITE database (UNITE\_v2020\_February2020.RData provided by DECIPHER). We combined the assignments of both classifiers, choosing for each ASV the classifier that could assign to a deeper taxonomic rank. In the case of equality, the assignment of the RDP classifier was chosen.

## 2.5 | Statistical analysis

The R package ggplot2 was used for plotting (Wickham, 2016), except when otherwise noted.

The soil variables were correlated with each other using the Pearson correlation coefficient and displayed with `ggcorrplot::correlogram()` (Kassambara, 2022). PCA was conducted with the function `base::prcomp()` using scaled and centred variables. Soil texture was classified with the R package `soiltexture` (Moeys, 2018) using the classification system 'USDA.TT' and displayed with the `soiltexture::TT.plot()` function.

The R package `phyloseq` was used to process the microbiome data (McMurdie & Holmes, 2013) with the R package `vegan` for community analyses (Oksanen et al., 2019). Negative controls were filtered out since the number of sequences was very low. One sample with 23 sequences was considered failed and removed (F35, replicate 3). After normalization (relative abundance), data were split into three sublibraries for each year, and absent OTUs were removed from OTU tables for each sublibrary. PCA was performed with `vegan::rda()` on Hellinger-transformed data (Legendre & Gallagher, 2001). The four replicates for each field generally clustered close together (Supporting Information: Figure S2), indicating that they have similar communities. Therefore, the four replicates were merged by summing the reads. Supporting Information: Table S2 shows summary statistics for the three merged libraries.

Rarefaction curves were prepared with `vegan::rarecurve()`, confirming that the fungal communities were sequenced sufficiently deep to capture the existing diversity (Supporting Information: Figure S3). Rarefaction analysis also revealed that the number of sequences per sample depended on the sampling year. Of note, the factor 'year' is confounded with the preparation of separate sequencing libraries for each year (see below, where we disentangle the 'year/library' effect). Summary statistics are presented in Supporting Information: Table S3. The number of generated sequences per sample was highest in 2018 because that library was sequenced twice in two subsequent runs. The OTU table was rarefied to the lowest number of sequences per sample (4272) `phyloseq::rarefy_even_depth()` for subsequent data analysis, as recommended by (Weiss et al., 2015).

The data was filtered based on prevalence, keeping only OTUs present in at least five samples. After filtering, the OTU tables comprised 227,801 sequences and 452 OTUs. To plot the taxonomy profile with a bar chart, OTUs were merged based on phylum name with `phyloseq::tax_glom()`. The abundant community was defined as the OTUs that were present in 90% of the samples. PCA and partial PCA were performed with `vegan::rda()` on Hellinger-transformed data (Legendre & Gallagher, 2001), conditioning for library (Borcard et al., 2018). The Bray-Curtis dissimilarity matrix of the Hellinger-transformed data was calculated with `vegan::vegdist()` and the maximum distance was extracted from this distance matrix.

For constrained ordination, variable selection for RDA was performed with `vegan::ordiR2step()` and 999 permutations, conditioning on the library. We performed forward selection and started with an 'empty' model and a 'scope' model which contains all the candidate variables (Borcard et al., 2018). Variables are added in order of decreasing F-values, the permutations are used to test the addition of each variable. The variable selection stops when the permutation probability is larger than a predefined significance level



(0.05). In case of two variables with equal F-values, the selected set of variables is that which yields the model with the lowest Akaike Information Criterion, see Supporting Information: Table S4 for results. The significance of the relationship between the selected variables and the community matrix was tested with a global permutation test using `vegan::anova.cca()` with 999 permutations.

We used leave-one-out validation to predict community composition. Its principle is to omit the data of one site (both fungal community and soil properties), to build a model based on the data of all the remaining other sites and to predict the site scores of the omitted sample using its soil property data (Supporting Information: Figure S4). This is then repeated for each field. Model selection was performed for each reduced data set with `vegan::ordiR2step()` as before except with only 99 permutations. Each model was saved and all the models were displayed with a heatmap. Prediction was performed with `vegan::predict()`; this function can predict the species composition (with argument `type = 'response'`). Root mean square error (RMSE) is the mean of the differences between the predicted values and the observed values, and this was calculated for each iteration of the leave-one-out validation.

The fourth library contains all the samples from the 3 years, see Supporting Information: Table S3 for summary statistics. PERMANOVA analysis was performed using `vegan::adonis2()`, see Supporting Information: Table S5 for results. The effect sizes of 'year' were similar both with samples in the same library (9.1%, estimated based on  $R^2$  values from PERMANOVA) and with samples in separate libraries (12.7%). PCA was performed as before and shows that the community composition of the re-sequenced library is very similar to the community composition of the first three libraries (Supporting Information: Figure S5).

The data analysis code is made accessible through [github.com/PMI-Basel/Bodenhausen\\_et\\_al\\_Prediction\\_soil](https://github.com/PMI-Basel/Bodenhausen_et_al_Prediction_soil).

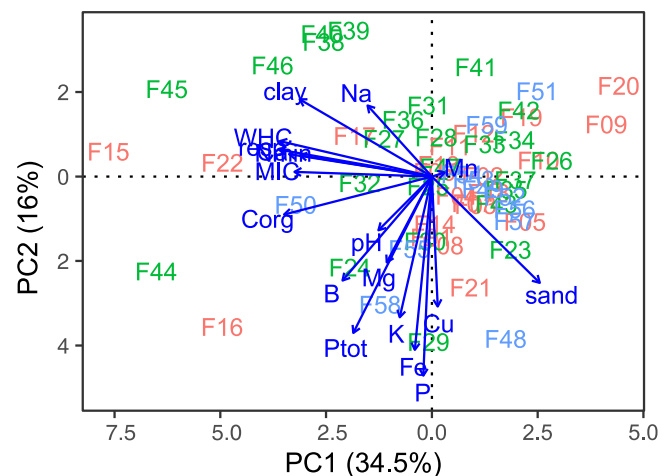
The environmental data of the Swiss Soil Monitoring Network (NABO) (<https://www.agroscope.admin.ch/agroscope/en/home/topics/environment-resources/soil-bodies-water-nutrients/nabo.html>) is available upon demand by signing a data usage agreement. The microbiome data is published (Gschwend et al., 2021). We analyzed the environmental data and the fungal community data using the same R scripts as before, so we provided only the report (Supporting Information: Data set S1).

Soil physicochemical properties were measured according to standard protocols and both bacterial and fungal communities were sequenced once a year for 5 years, but here we focus only on the fungal community of the last year. For constrained ordination, variable selection for RDA was performed as before except without conditioning, see Supporting Information: Table S6 for results.

### 3 | RESULTS

#### 3.1 | Soil properties

We focused our study on soils from 59 arable fields and collected samples from 22 fields in 2018, 25 fields in 2019 and 12 fields in 2020



**FIGURE 2** Properties of the different field soils. PCA biplot of 18 soil variables. F stands for field which are numbered from 01 to 59. The soil samples were collected over 3 years, in 2018 (red), 2019 (green) and 2020 (blue). Corg, organic carbon; MIC, microbial biomass; PCA, principal component analysis; Ptot, total phosphorous; resp, microbial respiration; WHC, water holding capacity.

(Supporting Information: Figure S1). These fields were managed by the farmers according to Swiss standards of conventional agriculture (Richner & Sinaj, 2017). We collected soil samples in spring before fertilization and measured a broad panel of a total of 39 physical, chemical and biological properties, including texture, pH, macro- and micronutrients, microbial biomass and microbial respiration (Supporting Information: Table S1). We classified the texture of the soils according to the Food and Agriculture Organization and found that most soils were either clay loam (17 field soils) or loam (29 field soils), while the remaining 13 field soils were categorized in other and less abundant classes (Supporting Information: Figure S6). We correlated all the measured properties with each other and found several groups of highly correlated measurements (Supporting Information: Figure S7); for example, different extraction methods of magnesium, potassium or phosphorus as well as soil fertility measures correlated with each other. To avoid overfitting, the number of variables was further reduced by retaining one measurement for each type of analysis (Supporting Information: Figure S7). We performed a PCA with this reduced set of 18 soil properties to examine the physical, chemical and biological diversity of the sampled soils. The first axis of the PCA, which accounted for 34.5% of the variance, separated the soil according to microbial biomass, and other variables strongly correlated with this variable, such as organic carbon and water holding capacity (Figure 2). The second axis of the PCA, which explained 16.0% of the variance, separated the soils according to the concentration of several macro- and micronutrients, with the main driver for the second axis being the concentration of phosphorous.

#### 3.2 | Fungal communities

From the exact same soil samples, we characterized the fungal communities by amplifying the full-length ITS region with the PCR

primers ITS1F and ITS4 and sequencing the amplicons with long-read sequencing technology. Consistent with our previous results (Bodenhausen et al., 2019), we found that this primer pair predominantly amplified DNA from Ascomycota, Basidiomycota and Mortierellomycota (Supporting Information: Figure S8) and detected only a small amount of Glomeromycota (<5%). Substantial (>5%) proportions of other phyla were found for Zoopagomycota (sites F32, F33, F36, F38, F39 and F42), Chytridiomycota (sites F05, F14, F27, F39, F40, F44 and F50), and Olpidiomycota in one field soil (F12). Inspecting the most abundant members of the fungal community, we noticed that the three most abundant fungal sequences were present in all field soils, while the prevalence pattern of less abundant community members varied more between the different soils (Figure 3). The most abundant fungal sequence (OTU1), accounting for an average of 18% of the community, belonged to a *Mortierella* species (Table 1). Other consistently abundant sequences corresponded to *Plectosphaerellaceae* (OTU2) and an *Orbiliaceae* (OTU3). A few field soils contained uniquely abundant fungi such as *Fusicolla septimanifiniscentiae* (OTU8), *Gibberella intricans* (OTU11), *Fusarium solani* (OTU19), all three of them Ascomycota, and *Solicoccozyma aerea* (OTU09), a Basidiomycota. Grouping the soil fungal communities using hierarchical clustering identified four major clusters, which mainly reflected combinations of the three most abundant fungi. OTU1 was moderately abundant in the communities of cluster C1, OTU1 and OTU2 were both very abundant in C2, and OTU1 and OTU3 had different ratios in clusters C3 and C4 (Figure 3).

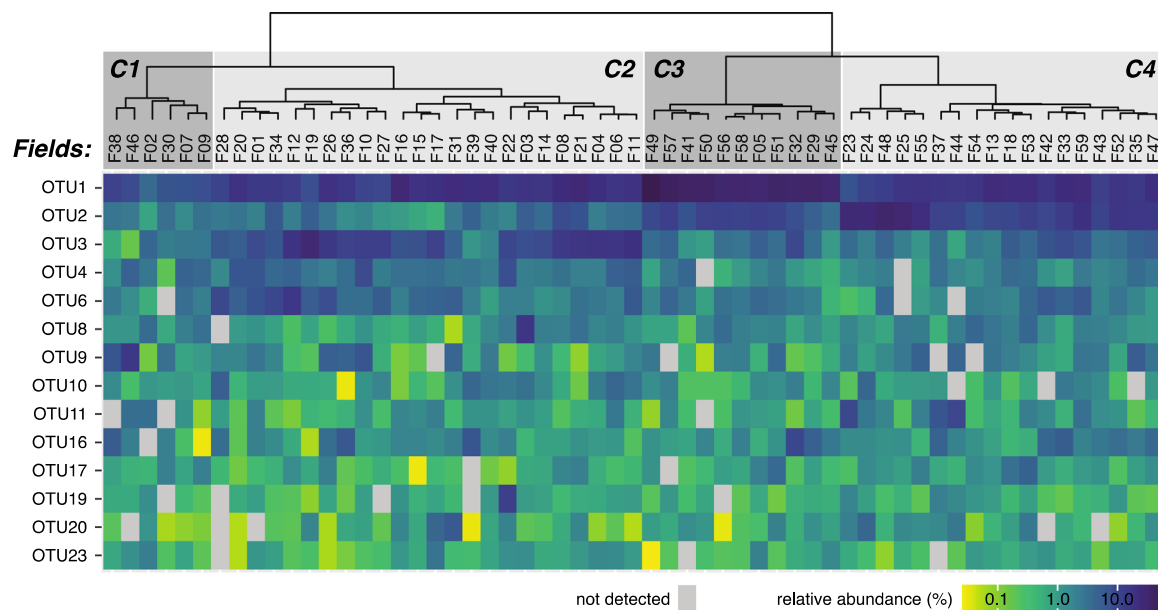
Analogous to the analysis of soil properties, we performed a PCA to examine the fungal diversity in the sampled soils. We noticed that the samples clustered by year (Supporting Information: Figure S9A), reflecting the fact that a separate sequencing library

was prepared for each year. We used partial PCA conditioned with 'year/library' to control for this effect (Supporting Information: Figure S9B) for the subsequent combinatory analysis of soil properties and fungal community data.

### 3.3 | Modelling soil properties and fungal communities

We would like to predict community composition based on combined soil properties. Therefore, we used RDA to model the relationship of the set of reduced soil properties with fungal community composition. RDA begins with a regression of the community data with the predictors which is the environmental data followed by a PCA ordination. To identify the soil properties that best explained the community composition, we performed forward variable selection in RDA while partialling out the library effect. We identified the following 10 variables (Supporting Information: Table S4 for associated F statistics and *p* values): pH, potassium, sand, microbial biomass, boron, mineral nitrogen, microbial respiration, manganese, clay and total phosphorus. This model has an adjusted  $R^2$  of 0.1430, which means that 14.3% of the variance is explained by these 10 soil variables; furthermore, we confirmed the significance of the relationship (global test:  $F = 2.05$ ,  $p = 0.001$ ).

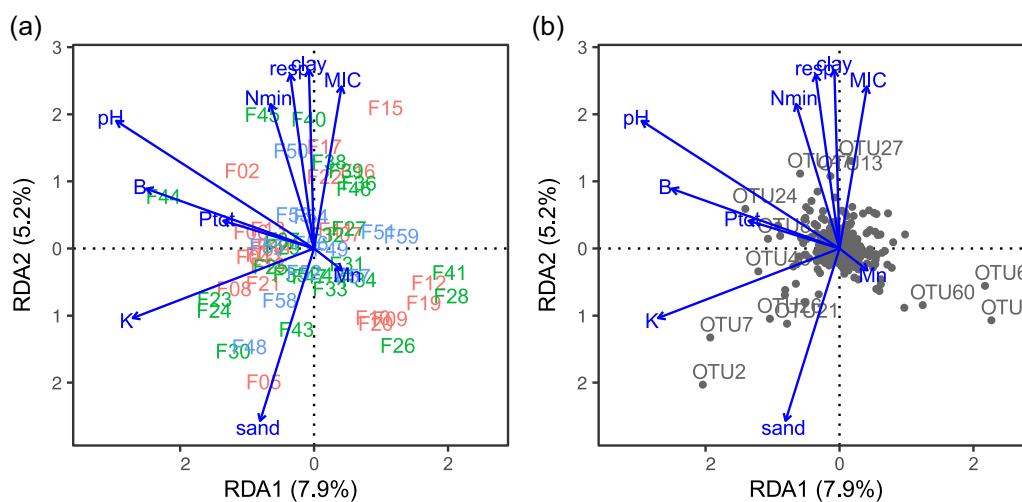
For visualization, we report the results of the RDA with two biplots, with the samples (data as site scores) on the left and as OTUs (data as species scores) on the right and the soil properties in both plots (Figure 4). Arrow length and direction visualize the contribution and relationship of the soil properties to the fungal community data. Potassium and pH were the soil properties that explained most of the



**FIGURE 3** Abundant community members were defined as the OTUs which were present in 90% of the samples. OTUs are ordered by mean abundance. Data were log<sub>10</sub> transformed for the heatmap. Zero counts are represented with a grey square. Hierarchical clustering (with ward.D method) was used to group the fields using Bray–Curtis dissimilarities. Four clusters, C1–C4, are highlighted with grey boxes.

**TABLE 1** Taxonomy and mean relative abundance of the abundant community members.

OTU	Phylum	Order	Genus and species	Classifier	Mean relative abundance (%)
OTU1	Mortierellomycota	Mortierellales	<i>Mortierella minutissima</i>	RDP	18.34
OTU2	Ascomycota	Glomerellales	NA	IDTAXA	7.59
OTU3	Ascomycota	Orbiliiales	NA	RDP	6.00
OTU4	Ascomycota	Hypocreales	NA	IDTAXA	2.45
OTU6	Mortierellomycota	Mortierellales	<i>Mortierella exigua</i>	RDP	3.06
OTU8	Ascomycota	Hypocreales	<i>Fusicolla septimanifiniscentiae</i>	RDP	1.64
OTU9	Basidiomycota	Filobasidiales	<i>Solicoccozyma aerea</i>	IDTAXA	1.92
OTU10	Ascomycota	NA	NA	RDP	1.19
OTU11	Ascomycota	Hypocreales	<i>Gibberella intricans</i>	RDP	1.29
OTU16	Zoopagomycota	Zoopagales	<i>Syncephalis</i> sp	IDTAXA	1.86
OTU17	Ascomycota	Helotiales	NA	IDTAXA	0.87
OTU19	Ascomycota	Hypocreales	<i>Fusarium solani</i>	RDP	0.79
OTU20	Ascomycota	Glomerellales	<i>Plectosphaerella niemeijerum</i>	IDTAXA	0.91
OTU22	Ascomycota	Thelebolales	<i>Pseudeurotium bakeri</i>	IDTAXA	0.75
OTU23	Ascomycota	Hypocreales	NA	IDTAXA	0.65

**FIGURE 4** Redundancy analysis (RDA) biplots reporting the fungal community and soil chemical data. The fungal community data are reported based on site (a) or species scores (i.e., OTUs, b) together with the chemical soil properties. Only OTUs with absolute values of scores >1 are labelled. MIC, microbial biomass; Nmin, mineral nitrogen; Ptot, total phosphorous; resp, microbial respiration.

community structure (longest arrows), while manganese had only a weak contribution (shortest arrow). The angles between the response variables (fungal community) and explanatory variables (soil properties) or between variables themselves reflect their relationships (Borcard et al., 2018). The first RDA axis separated the fungal communities mainly according to the key nutrients phosphorous, potassium and boron, with manganese pointing in the opposite

direction. The second axis separated the fungal communities due to clay, Nmin, microbial respiration and microbial biomass (all measures of soil fertility) and was opposite to sand. Hence, the second RDA axis represents a soil fertility gradient across the sampled field sites. The contribution of pH to the ordination is particular, as its arrow is roughly at a 45° angle to both RDA axes. The fields F02 and F44 are close to the arrowhead of pH and perpendicular to the arrow for sand

(Figure 4a), which indicates that these soils are characterized by relatively high pH and low sand contents (Supporting Information: Table S1). Analogously, the RDA ordination can be interpreted for the relationships of the fungal species with the soil properties. For instance, OTU2 is close to the arrowheads of sand and potassium (Figure 4b) and is therefore expected to be abundant in potassium-rich sandy soils. By contrast, OTU5 and OTU6 are located opposite the arrow for pH, indicating that they are abundant in low-pH soils. These examples of interpretations of the RDA can be confirmed with single linear regression (Supporting Information: Figure S10).

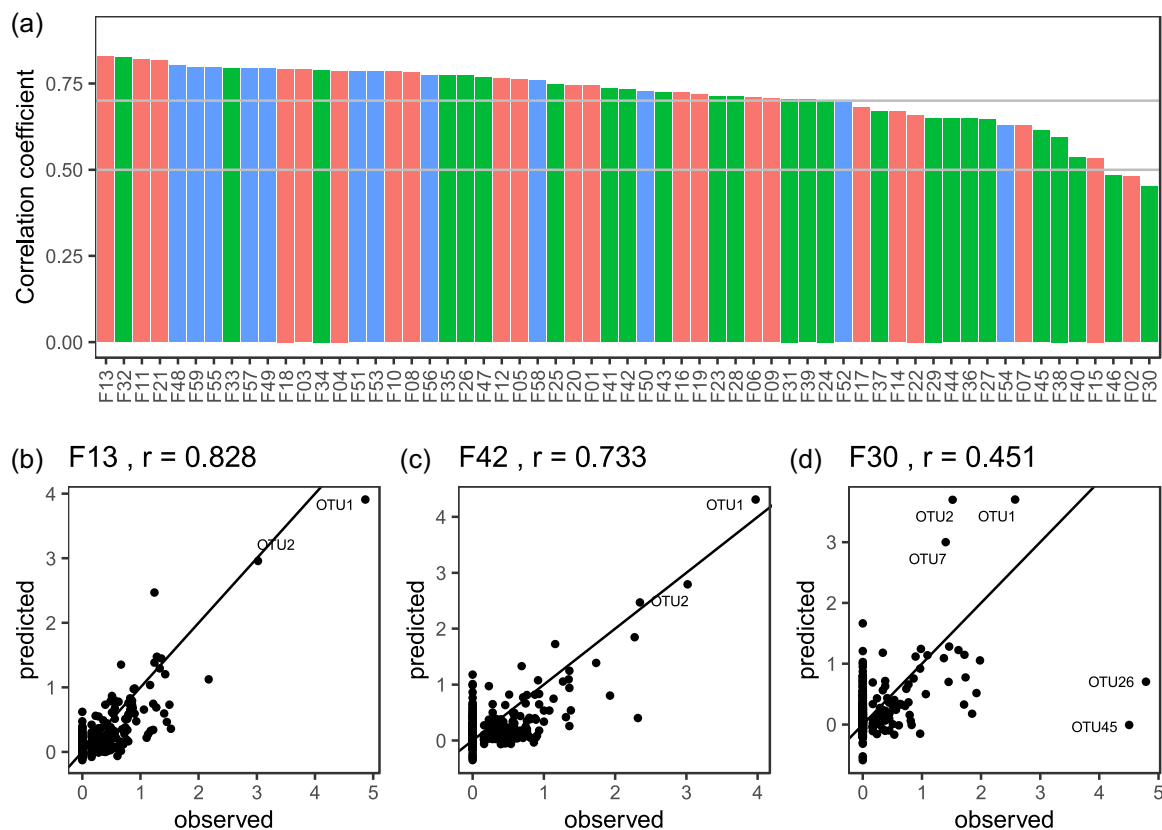
### 3.4 | Predicting fungal communities from soil properties

Next, we tested the suitability of RDA to predict fungal community composition using leave-one-out validation (Supporting Information: Figure S4). At each iteration, we performed forward variable selection to identify the model that best explained the community composition. Supporting Information: Figure S11B lists for each site the retained soil properties of these models. Three variables (pH, K and microbial respiration) were selected for all the iterations. The ten variables from the full model (Figure 4) were selected by most of the

iterations. In addition, Corg, WHC and Ca were selected in 1, 2 or 3 iterations.

RDA can be used to predict the abundances of individual fungi (estimates of the community data). Using leave-one-out validation, we predicted the OTU abundances for each soil and compared them with the observed OTU abundance data (Figure 5). We calculated the RMSE to compare the quality of the models. The fields with highest RMSE (poor prediction) were F02, F15 and F30 (Supporting Information: Figure S11A). As an intuitive measure for 'goodness' of prediction, we utilized Pearson's  $r$  from correlations of observed versus predicted fungal abundances. For more than half of the soils, the model predicted well (Pearson  $r > 0.7$ ) the abundances of individual community members (Figure 5a). Predictions were fair (Pearson  $0.7 > r > 0.5$ ) for 15/59 fields while in only three fields (namely, F02, F30 and F46), the predicted abundances did not agree (Pearson's  $r < 0.5$ ) with the observed community data. Figure 5 provides examples of field soils where predictions were best (F13; Figure 5b), average (F42; Figure 5c), and worst (F30; Figure 5d). These panels also show that species which were absent in the observed communities were predicted to occur by the RDA.

To understand what explained poor predictability, we investigated the contributions of different technical and experimental factors of the data set. Prediction success depends weakly on soil



**FIGURE 5** Predicting species abundance with leave one-out-validation. (a) Pearson correlation coefficient for each site; horizontal grey lines show thresholds of 0.7 and 0.5. (b)–(d) Relationship of predicted and observed abundance of individual OTUs for each field: (b) best-predicted field, (c) median predicted field and (d) worst predicted field. Only OTUs with values  $> 2.5$  are labelled.



texture classes (Supporting Information: Figure S12A) and the 'year/library' effect (Supporting Information: Figure S12B). By contrast, prediction success is negatively correlated with the maximum Bray–Curtis dissimilarity (Supporting Information: Figure S12C), indicating that fungal communities of sites which have similar communities to other sites can be better predicted than sites with a unique fungal community (larger Bray–Curtis dissimilarity). Similarly, prediction success depends strongly on the group membership of the most abundant members of the community (Supporting Information: Figure S12D). All three soils of poor predictability belonged to cluster C1 (Figure 2), which is also the cluster with the lowest numbers of soils.

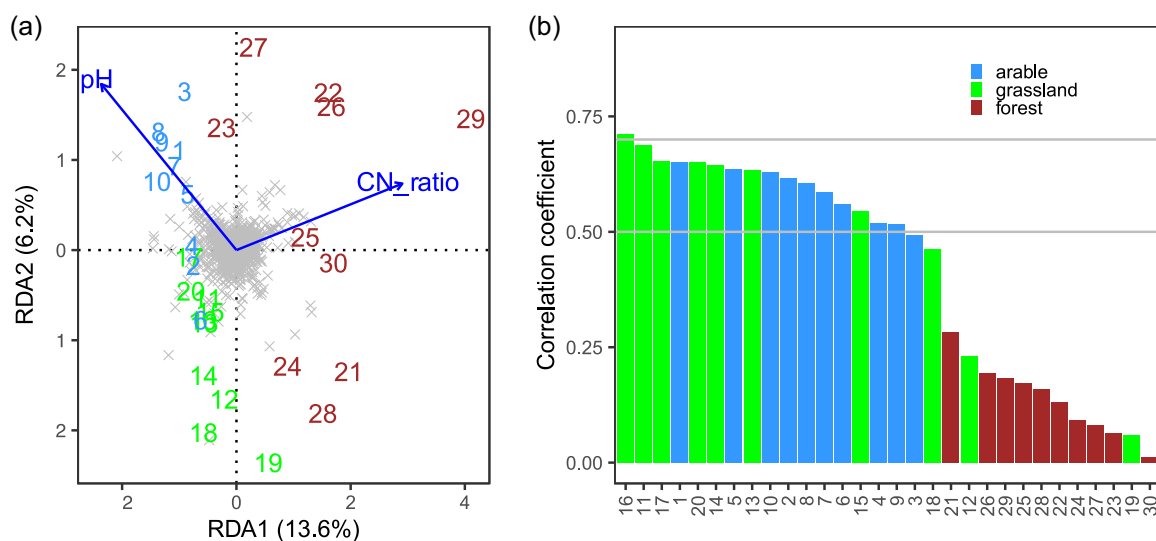
We validated our prediction approach with a publicly available data set of paired physicochemical soil properties and soil fungal profiles, which was collected for a soil monitoring network of 10 arable, 10 grassland and 10 forest sites (Gschwend et al., 2021). We analyzed this data using the approach described above (Supporting Information: Data set S1) and subsequently for RDA-based prediction. The RDA modelling the relationship of soil properties with fungal community composition revealed a clear clustering by land use and a particularly high variability among the forest sites (Figure 6a). A full model with soil C/N ratios and pH (Supporting Information: Table S6) was sufficient to explain differences in soil fungal communities with an adjusted  $R^2$  of 0.139 (significance confirmed with a global test:  $F = 3.34$ ,  $p = 0.001$ ). These soil properties were retained after forward variable selection in all leave-one-out iterations and resulted in predictions models of similar quality as with our data set (RMSE range: 0.15–0.34; Supporting Information: Data set S1). Again, we used Pearson correlation between observed and predicted abundances as a measure for 'goodness' of prediction. Land use strongly affected prediction ( $F = 30.01$ ,

$p < 0.001$ , Supporting Information: Figure S13B), with predictions for arable and most grassland sites working well (Pearson  $r > 0.7$ ) to fair (Pearson  $0.7 > r > 0.5$ ), while predictions failed for the forest sites (Pearson  $r < 0.5$ ; Figure 6b). Soil texture classes did not affect predictability, because forest soils are of many different soil textures (Supporting Information: Figure S13A). On the other hand, forest fungal communities are very different from each other (Supporting Information: Figure S13C) and do not have the same abundant taxa as arable and grassland sites (Supporting Information: Figure S13D). This validation experiment highlights that predictions generally work if closely related samples of similar physicochemical soil properties and similar fungal communities are present in the database during leave-one-out validation.

## 4 | DISCUSSION

### 4.1 | 10 factors in our model

Using RDA, we modelled the relationship between soil properties and fungal communities and identified 10 explanatory variables: pH, potassium, sand, microbial biomass, microbial respiration, Nmin, manganese, boron, clay, and total phosphorus (Figure 4). We found soil pH to be one of the strongest variables shaping the fungal community in the agricultural soils assessed in this study. Soil pH is known to be a major factor shaping bacterial communities at the global scale (Luber et al., 2009) and also at the field level (Rousk et al., 2010). By contrast, soil pH was not found to be the strongest factor for fungal communities either at the global scale (Bahram et al., 2018) or at more local scale, such as long-term field experiments (Rousk et al., 2010) or in a collection of abandoned



**FIGURE 6** Validation of our approach. We tested our approach with another data set of 30 sites of a long-term soil monitoring network with sites under three different land (arable, grassland and forest). (a) Redundancy analysis (RDA) triplot reporting the sites in colours, the OTUs with grey crosses, and arrows showing the soil properties, retained after forward variable selection. (b) Pearson correlation coefficient between predicted and observed fungal communities at each site; horizontal grey lines show thresholds of 0.7 and 0.5.

farmlands (Zhang et al., 2018). However, in our experiment, pH was one of the strongest factors, which could be because all our soils are used for the same purpose, as arable land, and managed by conventional farmers, and they do not differ very much in terms of other soil properties.

Soil pH can affect microbial communities either directly, for example, by modifying enzymatic activities, or indirectly, by changing the solubility of soil minerals. Indeed, we found the macronutrients potassium and phosphorous as well as the micronutrients manganese and boron to be important drivers of the fungal communities in Swiss arable fields. Nitrogen and phosphorous availability are well-known factors structuring microbial communities (Fierer & Jackson, 2006), but less is known about manganese and boron. Manganese is a micronutrient and plays a role in redox equilibrium and as a co-factor in enzymes and transcription factors (Kinskovski & Staats, 2022). On the other hand, boron is thought to be a nonessential element for fungi, but it can be toxic thus is often used as an antifungal agent (Estevez-Fregoso et al., 2021). Using a pot experiment, Vera and coworkers were able to show that boron concentration affected bacterial community composition (Vera et al., 2019); however, they found only weak evidence for the effect of boron on the fungal community, in contrast with our results with Swiss agricultural soils. Manganese application was also shown to alter both fungal and bacterial community composition in an incubation experiment (Jin et al., 2022). However, we found few publications reporting the effects of boron and manganese on the fungal communities of agricultural soils.

Soil texture, which is the proportion of sand, silt and clay, is a known factor shaping microbial communities (Fierer, 2017); nevertheless, microbial biomass and respiration are not often described as factors shaping microbial communities, perhaps because they are not often analyzed (Bünemann et al., 2018). However, we (and others) have found that they are both strongly correlated with soil organic carbon and water holding capacity, which are often measured and are known drivers of the microbial community (Fierer, 2017).

## 4.2 | Other (not measured) factors

The RDA model of this study, including 10 soil properties, explained only 14% of the variance in the fungal community; similarly, the RDA model of the publicly available data used for validation also explained only 14% of the variance. These seemingly low values are typical for microbiome studies. For example, in our previous microbiome studies of long-term field experiments, the management system explained 13.5% (Kundel et al., 2020) to 30% (Hartman et al., 2018) of the variance in fungal communities. However, we might have missed key soil properties or biological parameters that predict fungal abundance (e.g., fungal grazers). Some classical soil indicators, in particular physical properties, were missing from our list. For example, we did not measure bulk soil density because we expected that it would not differ much since all the farmers participating in our study ploughed their fields. Nevertheless, soil compaction has been shown to impact

microbes in forests (Hartmann et al., 2014) as well as in potato fields (Gattinger et al., 2002). Perhaps more importantly, we did not assess climatic variables even though precipitation has been shown to impact arbuscular mycorrhizal fungi (Davison et al., 2021) and rhizosphere communities (Mittelstrass et al., 2021). Finally, we did not measure pesticide residues, which were recently shown to impact microbial communities (Riedo et al., 2021). Of course, the more soil properties that could be measured, the better. Future work will need to show which additional soil parameters further enhance the predictions of soil fungal communities. For this proof of concept study, we started with the chemical soil properties that farmers often have at hand for decisions about agricultural management.

## 4.3 | Successful prediction with RDA

We evaluated our model with leave-one-out validation and predicted abundances of OTUs (Figure 5). Prediction was successful for most fields in our study. This collection of soils is quite similar because the fields are all used for the same goal (growing maize) and managed in the same way (conventional agriculture), and they are located in a relatively small area (80 km around Zürich in Switzerland). Moreover, we found that the fungal communities share several similarities in terms of the abundant community members (Figure 2), which probably contributes to good predictability. A limitation of our study is that we used a relatively small number of different samples in comparison to other publications, such as (García-Jiménez et al., 2020). However, this was a deliberate decision as we chose these fields for on-farm experiments with arbuscular mycorrhizal fungi inoculation, which will be reported in a separate publication (Lutz et al., Submitted). The aim of that subsequent publication is to predict inoculation success using both soil properties and microbiome composition. In contrast, this present study represents a first step to predict the soil fungal microbiome based on soil properties.

Additionally, we validated our approach with a data set from a soil monitoring network. Compared to our field soil data set, the latter is more heterogeneous because it includes different types of land uses. We found that fungal communities from arable and grassland soils could also be predicted with RDA while prediction of forest sites was generally poor, which could be explained by the highly variable nature of the forest sites. The forest sites were more different from each other based on the soil physicochemical properties (Figure 6a) as well as the fungal communities (Supporting Information: Figure S13). Consequently, they did not have closely related sites (similar properties and communities) present in the database for prediction with leave-one-out validation. With both data sets, prediction was less successful with soils which were most different from each other, indicating that the number of similar soils per group influences the goodness of prediction. Overall, the predictability of fungal communities functioned best for field soils that were most similar to the soils used to train the RDA model, whereas 'solitary' sites with more unique soil properties and noncanonical OTU dominance patterns were more difficult to predict.

Careful inspection of the predicted fungal communities revealed the presence of taxa which were not measured in the observed data (Figure 5b,c,d; Supporting Information: Data set S1). These missing taxa were generally fungi of low abundance. The observation of such missing taxa in the predictions could be due to insufficient sequencing depth of the particular field sample or because the prediction model uses data from a similar soil where this fungus is present at low abundance. As a note of caution, while species prediction is reliable for abundant fungi which are present at many sites, it has some distortion for rare members of the community. We think that this limitation will be overcome with more intensive sequencing and increasing numbers of field sites that contribute to modelling.

Here we demonstrated that RDA is an effective technique for predicting microbiome composition with two different data sets. Alternatively, machine learning type of approaches could be used, too. One such method is deep learning, which was used to predict the rhizosphere microbiome of maize based on five factors: temperature, precipitation, plant age, maize line and maize variety (García-Jiménez et al., 2020). One advantage of RDA compared to machine learning is that it is an ordination technique allowing for visual representation of the model. In addition, RDA is computationally inexpensive and does not require extensive bioinformatic expertise. Finally, RDA relies on statistical methods rather than a black box approach and can correct for known experimental covariates like year, library or sampling difference. Finally, both types of approaches share a common limitation in their ability to extrapolate beyond the training data, as they rely on the data used to train them and may not produce reliable predictions for data that fall outside the training set.

## 5 | CONCLUSIONS

Previous biogeography studies have identified certain factors that influence the microbial community: for example, the well-documented relationship between pH and bacterial diversity (Fierer & Jackson, 2006), or the association between soil calcium and fungal diversity (Tedersoo et al., 2014). However, to the best of our knowledge, no studies have attempted to use those factors in a 'forward' biogeographical manner to predict microbial species abundance. Our study is a proof-of-concept that microbial community composition can be predicted with multivariate ordination. We show that a model with 10 soil properties is sufficient to accurately predict fungal community composition in arable fields. Through leave-one-out validation, we were able to successfully predict species abundance for most of the fields. We further showed that this prediction approach also works with another publicly available data set. These findings show that multivariate analysis can be used to predict microbial community composition from environmental data. Future research is needed to further test our approach with fields outside the study area. In addition, prediction based on ordination should be tested with different soil types, such as forest soils, or marine ecosystems.

## AUTHOR CONTRIBUTIONS

Natacha Bodenhausen analyzed the data with help from Jan Waelchli and wrote the manuscript with contributions from Klaus Schlaeppi and Marcel G. A. van der Heijden. Julia Hess collected the soil samples and coordinated the soil analysis. Alain Valzano prepared the first library for sequencing. Gabriel Deslandes-Hérolé prepared the second, third and fourth libraries for sequencing. Jan Waelchli performed the bioinformatic analysis. Reinhard Furrer contributed to the data analysis. All authors read and approved the final manuscript.

## ACKNOWLEDGEMENTS

We thank our technicians Andrea Bonvicini and Susanne Mueller from Agroscope for the analysis of the microbial respiration and biomass, Erich Szerencsits from Agroscope for the map of the field sites, Pamela Nicholson from the next-generation sequencing platform from the University of Bern for technical support and Jean-Claude Walsler from the genetic diversity centre at ETH for help with bioinformatics. We are grateful to Florian Gschwend and Janine Moll-Mielewicz from Agroscope for providing the data from the Swiss Soil Monitoring Network (NABO). This study was supported by the Gebert RUF Foundation (grant GRS-072/17 to NB, MvdH and KS) and core funding by the University of Basel to KS.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

The raw sequencing data are stored in the European Nucleotide Archive (<http://www.ebi.ac.uk/ena>) under project number PRJEB53587. The barcodes and primers of each sample are provided in the four mapping files on [github.com/PMI-Basel/Bodenhausen\\_et\\_al\\_Prediction\\_soil](https://github.com/PMI-Basel/Bodenhausen_et_al_Prediction_soil).

## ETHICS STATEMENT

The authors confirm that they adhered to the ethical policies of the Journal of Sustainable Agriculture and Environment.

## ORCID

Natacha Bodenhausen  <http://orcid.org/0000-0001-9680-1176>

Klaus Schlaeppi  <http://orcid.org/0000-0003-3620-0875>

## REFERENCES

- Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, et al. Structure and function of the global topsoil microbiome. *Nature*. 2018;560:233–7.
- Banerjee S, van der Heijden MGA. Soil microbiomes and one health. *Nat Rev Microbiol*. 2023;2:6–20.
- Bodenhausen N, Somerville V, Desirò A, Walsler J-C, Borghi L, van der Heijden MGA, et al. Petunia- and Arabidopsis-specific root microbiota responses to phosphate supplementation. *Phytobiomes J*. 2019;3:112–24.
- Borcard D, Gillet F, Legendre P. Numerical ecology with R. 2nd ed. Springer; 2018.

- Borcard D, Legendre P, Drapeau P. Partialling out the spatial component of ecological variation. *Ecology*. 1992;73:1045–55.
- Bünemann EK, Bongiorno G, Bai Z, Creamer RE, De Deyn G, de Goede R, et al. Soil quality—A critical review. *Soil Biol Biochem*. 2018;120:105–25.
- Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods*. 2016;13:581–3.
- Capblancq T, Forester BR. Redundancy analysis: a Swiss Army Knife for landscape genomics. *Methods Ecol Evol*. 2021;12:2298–309.
- Chang H-X, Haudenschild JS, Bowen CR, Hartman GL. Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity. *Front Microbiol*. 2017;8:425.
- Chen Y, Xu T, Fu W, Hu Y, Hu H, You L, et al. Soil organic carbon and total nitrogen predict large-scale distribution of soil fungal communities in temperate and alpine shrub ecosystems. *Eur J Soil Biol*. 2021;102:103270.
- Chu H, Gao G-F, Ma Y, Fan K, Delgado-Baquerizo M. Soil microbial biogeography in a changing world: recent advances and future perspectives. *mSystems*. 2020;5:e00803–19.
- Davison J, Moora M, Semchenko M, Adenan SB, Ahmed T, Akhmetzhanova AA, et al. Temperature and pH define the realised niche space of arbuscular mycorrhizal fungi. *New Phytol*. 2021;231:763–76.
- Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, et al. A global atlas of the dominant bacteria found in soil. *Science*. 2018;359:320–5.
- Estevez-Fregoso E, Farfán-García ED, García-Coronel IH, Martínez-Herrera E, Alatorre A, Scorei RI, et al. Effects of boron-containing compounds in the fungal kingdom. *J Trace Elem Med Biol*. 2021;65:126714.
- FAL R. FAW. Méthodes de référence des stations fédérales de recherches agronomiques. Switzerland: Agroscope; 2004. <https://www.agroscope.admin.ch/agroscope/fr/home/themes/environnement-ressources/monitoring-analyse/methodes-references.html>
- Fierer N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Microbiol*. 2017;15:579–90.
- Fierer N, Jackson RB. The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci U S A*. 2006;103:626–31.
- García-Jiménez B, Muñoz J, Cabello S, Medina J, Wilkinson MD. Predicting microbiomes through a deep latent space. *Bioinformatics*. 2020;37:1444–51.
- Gardes M, Bruns TD. ITS primers with enhanced specificity for basidiomycetes-application to the identification of mycorrhizae and rusts. *Mol Ecol*. 1993;2:113–8.
- Gattinger A, Ruser R, Schlotter M, Munch JC. Microbial community structure varies in different soil zones of a potato field. *J Plant Nutr Soil Sci*. 2002;165:421–8.
- Gschwend F, Hartmann M, Hug A-S, Enkerli J, Gubler A, Frey B, et al. Long-term stability of soil bacterial and fungal community structures revealed in their abundant and rare fractions. *Mol Ecol*. 2021;30:4305–20.
- Hartman K, van der Heijden MGA, Wittwer RA, Banerjee S, Walser J-C, Schlaeppli K. Cropping practices manipulate abundance patterns of root and soil microbiome members paving the way to smart farming. *Microbiome*. 2018;6:14.
- Hartmann M, Frey B, Mayer J, Mäder P, Widmer F. Distinct soil microbial diversity under long-term organic and conventional farming. *ISME J*. 2015;9:1177–94.
- Hartmann M, Niklaus PA, Zimmermann S, Schmutz S, Kremer J, Abarenkov K, et al. Resistance and resilience of the forest soil microbiome to logging-associated compaction. *ISME J*. 2014;8:226–44.
- Hermans SM, Buckley HL, Case BS, Curran-Cournane F, Taylor M, Lear G. Using soil bacterial communities to predict physico-chemical variables and soil quality. *Microbiome*. 2020;8:79.
- Jin M, Chen X, Gao M, Sun R, Tian D, Xiong Q, et al. Manganese promoted wheat straw decomposition by regulating microbial communities and enzyme activities. *J Appl Microbiol*. 2022;132:1079–90.
- Karimi B, Terrat S, Dequiedt S, Saby NPA, Horrigue W, Lelièvre M, et al. Biogeography of soil bacteria and archaea across France. *Sci Adv*. 2018;4:eaat1808.
- Kassambara A. ggcorrplot: Visualization of a Correlation Matrix using “ggplot2.” 2022.
- Kinskovski UP, Staats CC. Manganese and fungal pathogens: metabolism and potential association with virulence. *Fungal Biol Rev*. 2022;42:69–73.
- Kundel D, Bodenhausen N, Jørgensen HB, Truu J, Birkhofer K, Hedlund K, et al. Effects of simulated drought on biological soil quality, microbial diversity and yields under long-term conventional and organic agriculture. *FEMS Microbiol Ecol*. 2020;96:fiaa205.
- Lauber CL, Hamady M, Knight R, Fierer N. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol*. 2009;75:5111–20.
- Legendre P, Gallagher ED. Ecologically meaningful transformations for ordination of species data. *Oecologia*. 2001;129:271–80.
- Lutz BN, Hess J, Valzano-Held A, Waelchli J, Deslandes-Hérolde G, et al. Successful prediction of crop yield increases after microbiome engineering with mycorrhizal fungi. Submitted.
- Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, et al. Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol*. 2006;4:102–12.
- McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*. 2013;8:e61217.
- Mittelstrass J, Sperone FG, Horton MW. Using transects to disentangle the environmental drivers of plant-microbiome assembly. *Plant Cell Environ*. 2021;44:3745–55.
- Moeys J. Functions for Soil Texture Plot, Classification and Transformation. 2018. Functions for Soil Texture Plot, Classification and Transformation. <https://cran.r-project.org/web/packages/soiltexture/index.html>
- Nilsson RH, Larsson K-H, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, et al. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res*. 2019;47:D259–64.
- Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. 2022. <https://cran.r-project.org/web/packages/vegan/index.html>
- Richner W, Sinaj S. Grundlagen für die Düngung landwirtschaftlicher Kulturen in der Schweiz. *Agrarforschung Schweiz*. 2017. Available online: <https://www.agroscope.admin.ch/agroscope/de/home/themen/pflanzenbau/ackerbau/Pflanzenernaehrung/grud.html>
- Riedo J, Wettstein FE, Rösch A, Herzog C, Banerjee S, Büchi L, et al. Widespread occurrence of pesticides in organically managed agricultural soils the ghost of a conventional agricultural past? *Environ Sci Technol*. 2021;55:2919–28.
- Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD, et al. Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J*. 2007;1:283–90.
- Rousk J, Bååth E, Brookes PC, Lauber CL, Lozupone C, Caporaso JG, et al. Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J*. 2010;4:1340–51.
- Schlaeppli K, Bulgarelli D. The plant microbiome at work. *Mol Plant-Microbe Interact*. 2015;28:212–7.
- Team R. RStudio: Integrated Development Environment for R. 2016.
- Tedersoo L, Bahram M, Pölme S, Kõljalg U, Yorou NS, Wijesundera R, et al. Global diversity and geography of soil fungi. *Science*. 2014;346:1256688.
- Vera A, Moreno JL, García C, Morais D, Bastida F. Boron in soil: the impacts on the biomass, composition and activity of the soil microbial community. *Sci Total Environ*. 2019;685:564–73.
- Wagg C, Schlaeppli K, Banerjee S, Kuramae EE, van der Heijden MGA. Fungal-bacterial diversity and microbiome complexity predict ecosystem functioning. *Nat Commun*. 2019;10:4841.

- Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol.* 2007;73:5261–7.
- Weiss SJ, Xu Z, Amir A, Peddada S, Bittinger K, González A, et al. Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. *PeerJ PrePrints.* 2015e1408.
- White TJ, Bruns T, Lee S, Taylor J. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR protocols.* Academic Press, Inc.; 1990. p. 315–22.
- Wickham H. *ggplot2: elegant graphics for data analysis.* New York: Springer-Verlag; 2016.
- Zhang K, Cheng X, Shu X, Liu Y, Zhang Q. Linking soil bacterial and fungal communities to vegetation succession following agricultural abandonment. *Plant Soil.* 2018;431:19–36.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Bodenhausen N, Hess J, Valzano A, Deslandes-Hérolde G, Waelchli J, Furrer R, et al. Predicting soil fungal communities from chemical and physical properties. *J Sustain Agric Environ.* 2023;1–13.  
<https://doi.org/10.1002/sae.12055>