Predicting measures of soil health using the microbiome and supervised machine learning

Roland C. Wilhelm, Harold M. van Es, Daniel H. Buckley [*]

School of Integrative Plant Science, Bradfield Hall, Cornell University, Ithaca, NY, 14853, USA

A B S T R A C T

Soil health encompasses a range of biological, chemical, and physical soil properties that sustain the commercial and ecological value of agroecosystems. Monitoring soil health requires a comprehensive set of diagnostics that can be cost-prohibitive for routine analyses. The soil microbiome provides a rich source of information about soil properties, which can be assayed in a high-throughput, cost-effective way. We evaluated the accuracy of random forest (RF) and support vector machine (SVM) regression and classification models in predicting 12 measures of soil health, tillage status, and soil texture from 16S rRNA gene amplicon data with an operationally relevant sample set. We validated the efficacy of the best performing models against independent datasets and also tested best practices for processing microbiome data for use in machine learning. Soil health metrics could be predicted from microbiome data with the best models achieving a Kappa value of ~0.65, for categorical assessments, and a R^2 value of ~0.8, for numerical scores. Biological health ratings were better predicted than chemical or physical ratings. Validation with independent datasets revealed that models had general predictive value for soil properties, including yield. The ecological profiles of several taxa important for model accuracy matched the observed relationships with soil health, including *Pyrinomonadaceae, Nitrososphaeraceae*, and Candidatus *Udeaobacter*. Models trained at the highest taxonomic resolution proved most accurate, with losses in accuracy resulting from rarefying, sparsity filtering, and aggregating at higher taxonomic ranks. Our study provides the groundwork for developing scalable technology to use microbiome-based diagnostics for the assessment of soil health.

1. Introduction

Managing soil health is designed to promote environmental sustainability by improving the long-term fertility of cropland, reducing agricultural inputs, conserving biodiversity, and mitigating air and water pollution (Doran, 2002; Lehmann et al., 2020). Soil health is measured using a range of biological, chemical and physical properties that correspond with the commercial and ecological value of agroecosystems (Rinot et al., 2019; Stewart et al., 2018). Health status is benchmarked with a comprehensive set of metrics and monitored over time to assess the impacts of management (Moebius-Clune et al., 2017). The need for routine soil health monitoring can be cost-prohibitive for many farm managers seeking to improve stewardship practices. Recent advances in automation and DNA sequencing technology have dramatically reduced the cost of assaying the composition of microbial communities (the 'microbiome'), which might serve as an integrated measure of soil health. The soil microbiome contains information about the biological, chemical and physical status of soil which could be leveraged by machine learning (ML) to predict conventional metrics of soil health.

Microbial communities can be sensitive indicators of environmental change and dysbiosis, like pollution (Qin et al., 2020; Rocca et al., 2019; Werner et al., 2011) and human disease (Liu et al., 2020; Statnikov et al., 2013; Sze and Schloss, 2018). In agroecosystems, the soil microbiome is a rich source of information on soil properties affected by land-use and agricultural management practices which has been used to differentiate organic and conventional agriculture (Francioli et al., 2016; Hartmann et al., 2015; Pershina et al., 2015), nitrogen fertilization regime (Wessén et al., 2011; Zhalnina et al., 2013), and tillage practices (Degrune et al., 2017). Soil microorganisms are also active participants in processes that underlie soil health, like soil aggregate formation (Lehmann et al., 2017), disease and weed suppression (Cha et al., 2016; Liu et al., 2019; Mendes et al., 2011; Trognitz et al., 2016), and moisture retention and erosion control (Chamizo et al., 2018; Peng and Bruns, 2018; Rodríguez-Caballero et al., 2012; Zheng et al., 2018). However, variation in the soil microbiome due to the spatial and temporal effects of climate and geology is often far greater than the effects of management practices

Abbreviations

ACE	autoclaved citrate extractable
ANOVA	analysis of variance
ASV	amplicon sequence variant
CASH	Cornell assessment of soil health
CSS	cumulative sum scaling
K	potassium
ML	machine learning
P	extractable phosphorus
PCR	polymerase chain reaction
PERMANOVA	permutational multivariate analysis of variance
RF	random forest
rRNA	ribosomal RNA
SVM	support vector machine
t-SNE	t-distributed stochastic neighbor embedding

(Delgado-Baquerizo et al., 2017; Hartmann et al., 2015; Wilhelm et al., 2017). Thus, the promise of evaluating soil health using microbiome data has been constrained by the shear diversity and variation of soil microbial communities. This challenge might be overcome with ML approaches that are sensitive to underlying community structures that reflect soil properties relevant to soil monitoring.

Supervised ML can be used to predict system state characteristics from microbiome data. For example, microbiome-based ML has been successful in predicting human health status (Poore et al., 2020; Top-çuoğlu et al., 2020) and has shown promise in predicting crop productivity (Chang et al., 2017), soil carbon flux (Thompson et al., 2019), and physicochemical properties (Hermans et al., 2020). The efficacy of microbiome-based ML, however, varies by study size, data type, and biological system (Knights et al., 2011; Sze et al., 2019), and since soil microbiomes are far more diverse and heterogeneous than human microbiomes, it remains to be seen whether soil microbiome composition can inform soil health status. One of the major challenges in ML-based modeling of microbiome data is a tendency to overfit, since the number of features in the model (taxonomic units or genes) typically far exceeds the number of samples. The capacity of ML algorithms to capture variation in multi-dimensional community data might also be a liability due to inherent compositional biases and challenges in normalizing sampling efforts (Chen et al., 2018; McKnight et al., 2019; Morton et al., 2019). There are methods to address these challenges, such as regularization (Pasolli et al., 2016; Topçuoğlu et al., 2020) or aggregating features by broader classes (Statnikov et al., 2013; Zhou and Gallins, 2019), but microbiome data does not always have predictive value (Sze et al., 2019). The utility and inherent challenges of using microbiome-based ML to predict measures of soil health have yet to be examined.

Our study evaluated the use of microbiome-based ML to predict the physical, chemical, and biological measures of soil health encompassed in the Comprehensive Assessment of Soil Health (CASH) framework (Moebius-Clune et al., 2017). CASH is a diagnostic tool based on a suite of 12 soil health metrics ('health metrics') proven to be sensitive to the effects of agricultural practices like tillage intensity (Nunes et al., 2018) and organic versus conventional management (van Es and Karlen, 2019) and land-use practices across regions of the USA (Fine et al., 2017) and the globe (Bhadha et al., 2018; Frost et al., 2019; Gholoubi et al., 2018; Rekik et al., 2018; Williams et al., 2020). We collected 16S rRNA gene data from farmland soils in a manner consistent with operational soil health testing in the USA to evaluate the performance of regression- and classification-based predictions of health metrics using random forest (RF) and L2 regularized support vector machine (SVM) models. Building upon methods described by Topçuoğlu et al. (2020), we used ML to evaluate the importance of regularization, normalization, and the complexity of feature sets. We then compared the accuracy of predictions for health metrics and the features (*i.e.*, bacterial taxa) contributing most to ML model accuracy, providing a view of the underlying ecological basis of predictions. Finally, we validated the predictive performance of models against independent microbiome datasets, which either used the CASH framework ('Musgrave farm study'), or other related soil properties ('Pastureland study') (Lanzén et al., 2015). Our study aims to lay the groundwork for developing microbiome-based technology to perform soil health diagnostics to promote routine monitoring and the adoption of soil health practices.

2. Methods

2.1. Soil sampling and health diagnostics

A total of 949 soil samples were collected from farmlands across the USA and Canada as part of a national soil health characterization initiative by the U.S. Department of Agriculture Natural Resources Conservation Service (Fig. 1A). Soil sampling was performed according to the CASH protocol (Moebius-Clune et al., 2017). In brief, a 15 cm soil core is taken, homogenized and shipped on ice to the Cornell Soil Health Laboratory (Ithaca, NY), where they were air dried and generally analyzed within a period of 0.5–2 months. Over short periods, changes in microbiome composition due to air drying and storage are minor and comparable to the effects of freezing (Clark and Hirsch, 2008; Ivanova et al., 2017; Lauber et al., 2010; Tatangelo et al., 2014; Tzeneva et al., 2009). All samples were subjected to a full characterization of soil health (Fig. 1B), which includes measures of biological (organic matter content, respiration, 'autoclaved citrate extractable' (ACE) protein, and active carbon [also known as 'permanganate-oxidizable carbon']), chemical (pH, extractable phosphorus and potassium, and minor elements), and physical status (aggregate stability, available water capacity, and surface and subsurface hardness) (Schindelbeck et al., 2016). The 'minor elements' rating is an aggregate score based on magnesium, manganese, iron, and zinc concentrations. Each metric is designed to resolve a specific dimension of soil health related to management practices (see Supplementary Information; SI). Raw data from each metric is transformed using a scoring function to create a normalized soil health rating (Moebius-Clune et al., 2017). A total soil health rating ('health rating') is calculated based on the unweighted mean of all twelve ratings. Each metric is also assigned to a soil health category ('health category') upon which management recommendations are based. The following categories are assigned based on numerical ratings: (0–20): "very low," (20–40): "low," (40,60]: "medium," [60,80]: "high," and (80,100]: "very high" (Moebius-Clune et al., 2017). In addition to health metrics, ML models were created to classify total DNA yield (as a proxy for microbial biomass), the intensity of tilling ("no till"; "shallow till" $\leq$ 6 inch depth, and "deep till" $\geq$ 6 inch depth), and soil texture class, based on ratings of sand, silt and clay (details in SI).

2.2. DNA extraction and amplicon sequencing

DNA was extracted from soil using the 'DNeasy PowerSoil HTP 96 Kit', as per the manufacturer's instructions (QIAGEN, Germantown, MD) with a bead beating treatment of 2.5 min at 5.5 m s^{-1} (Bio Spec Products, Santa Clara, CA). Soil DNA extracts were quantified using the Quant-iT™ PicoGreen™ dsDNA Assay Kit (Thermo Fisher Scientific, Inc., Waltham, MA) and measured with a FilterMax F5 micro-plate reader (Molecular Devices, San Jose, CA). Bacterial community composition was determined by sequencing the V4 region of the 16S rRNA gene amplified from soil DNA extracts by polymerase chain reaction (PCR) using dual-indexed barcoded 515f/806r primers as described by (Kozich et al., 2013). PCR was performed with 2 ng of DNA template, in triplicate, according methods described in (Zwetsloot et al., 2020) and pooled, purified and normalized to a standard concentration using the SequalPrep normalization kit (Invitrogen, CA, USA). Amplicon
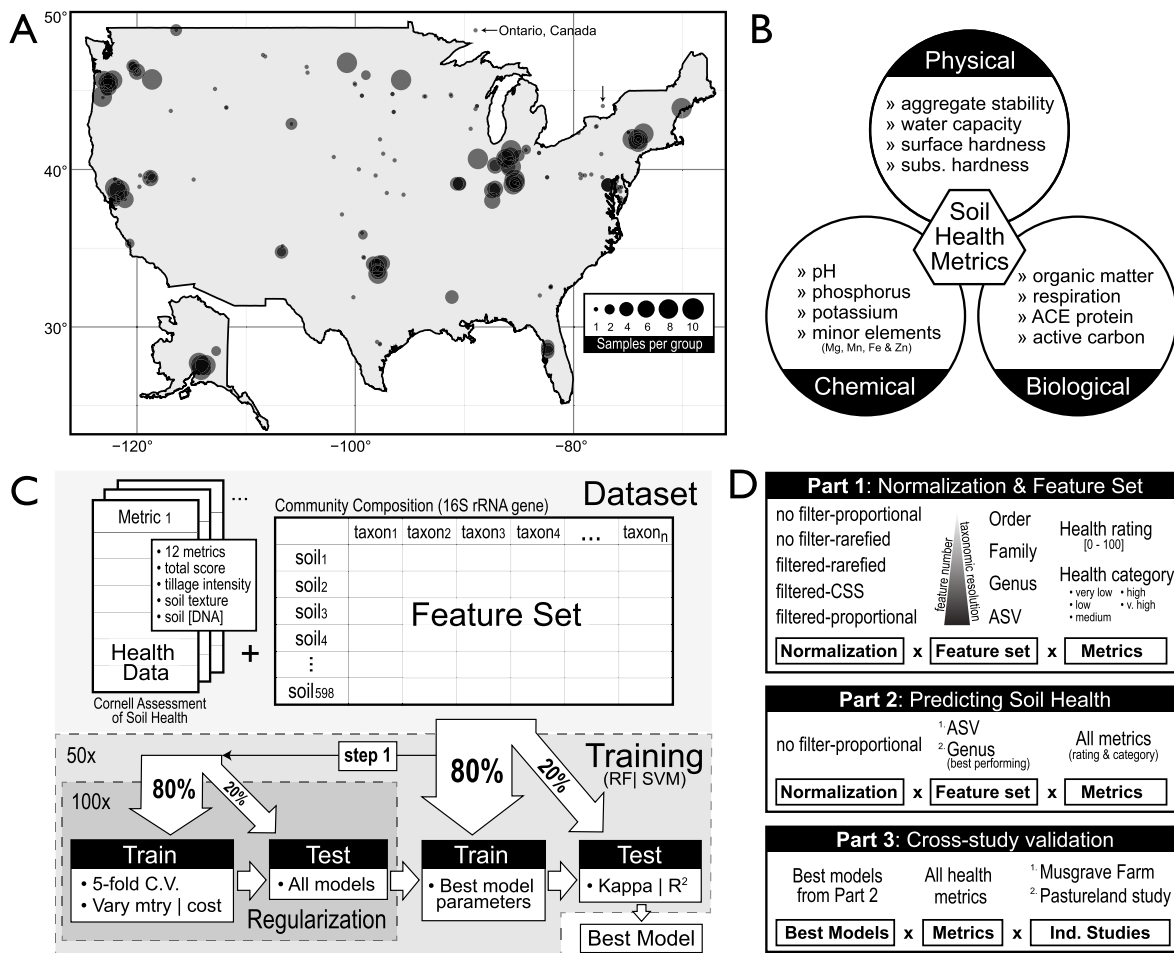
Fig. 1. An overview of the main components of our study. In (A), the geographical location of farmland soils used for soil health and microbiome analyses. In (B), the twelve soil health metrics predicted by machine learning models based upon soil bacterial community composition. In (C), a schematic outlining the machine learning workflow used to create classification and regression models for soil health metrics using different microbiome feature sets. In brief, data was first split into training (80%) and testing sets (20%). Training data then used to select the best parameters through cross validation and regularization. In (D), the tests performed in the three parts of our study: (1) identify the optimal processing of microbiome data, (2) identify which soil health metrics can be predicted from the microbiome and (3) to validate that the best models contain information that can predict soil health in independent datasets.

libraries were multiplexed and sequenced on a total of five lanes of Illumina MiSeq (2 × 250 paired-end) with a spike-in of 8% PhiX at the Biotechnology Resource Centre (Cornell University, Ithaca, NY, USA). The raw sequencing data was archived at the National Centre for Biotechnology Information (BioProject accession: PRJEB35975).

2.3. Bacterial community composition

A total of 753 samples were successfully sequenced, originating from 162 unique geographical regions (defined by a 1 km radius) and 191 unique field sites when subdivided by agricultural management practices ('geogroup'). The dataset was subsampled to cap the number of samples from a geogroup to 10, reducing the final dataset to 598 samples (mean = 3.1 sample per field site, median = 1). The samples that failed to sequence either yielded insufficient DNA during extraction or failed during PCR amplification and were two times more likely to be assigned to the low health category (Fisher's Exact; $p = 0.04$). Amplicon libraries were processed using *QIIME2* (v. 2020.2) (Bolyen et al., 2019), with dependencies on *DADA2* (Callahan et al., 2016), to assign sequences to amplicon sequence variants (ASVs). Taxonomic classification was performed using the *QIIME2* 'q2-feature-classifier' trained on the *Silva* database (nr_v132) (Quast et al., 2013). All ASVs detected in no-template controls were removed from analyses. Normalization was performed to standardize the sampling effort (i.e., sequencing depth)

among samples based on rarefying (sub-sampling to a common sequencing depth), proportioning (division by total sequencing depth) or cumulative sum scaling ('CSS', a form of proportioning based on quantiles). Sparsity filtering was performed to remove sparse ASVs, occurring at low frequency (fewer than 10 samples), and rare ASVs, occurring at low relative abundance (<0.01% of average read depth). Five dataset types were generated based on the filtering and read-depth normalization method applied: (i) unfiltered (rarefied), (ii) unfiltered (proportional), (iii) filtered (rarefied), (iv) filtered (CSS) and (v) filtered (proportional). Our sequencing efforts produced an average and minimum sequencing depths of 25,300 and 19,000 quality-processed reads per library, respectively. Alpha-diversity metrics (Shannon diversity index and Pielou's evenness) were calculated from filtered rarified libraries using the 'plot_richness' function from the *phyloseq* package (McMurdie and Holmes, 2013). Weighted Unifrac distances were calculated using the 'UniFrac' function in *phyloseq* (Lozupone et al., 2011) and used to assess beta-diversity using PERMANOVA and covariance in community composition and health ratings (using Bray-Curtis dissimilarity) with the Mantel test ('mantel') from the *vegan* package (Oksanen et al., 2015). The compositional similarity of bacterial communities was visualized using t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) with the *Rtsne* package (Krijthe, 2015). All multiple pairwise comparisons were performed using the Kruskal-Wallis test with the 'kruskalmc' function from

pgirmess (Giraudoux et al., 2018).

2.4. Supervised machine learning

Our ML approach generally followed the ML workflow proposed by (Topçuoğlu et al., 2020) to which we added the capacity to perform multi-factor classification- and regression-based modeling (Fig. 1C). All analyses were performed in *R* (v. 3.6.0) (Core Team, 2020) using the *caret* package (Kuhn, 2008). Two ML algorithms were used: random forest (RF), which routinely performs best on microbiome data (Thompson et al., 2019; Zhou and Gallins, 2019), and a L2 regularized SVM with a linear kernel (method 'svmLinear3' in *caret*), chosen for speed and efficacy (Topçuoğlu et al., 2020). Both algorithms are capable of performing classification- and regression-based modeling, which were used to predict health category and rating, respectively. Regularization was based on five-fold cross-validation of training data combined with a grid search to select the best hyperparameter settings for SVM (cost and loss) and RF (mtry). Predictive performance was evaluated using the kappa statistic (for classification models) and the coefficient of determination (R^2) of the linear regression of observed versus predicted data (for regression models). Training and validation were performed on an 80:20% split of the sample set which was repeated 50 and 25 times to account for lucky/unlucky splits for SVM and RF, respectively. Fewer splits were run for RF due to the extensive training time. Each feature (ASV or taxon) was scaled between 0 and 1 with the 'preProcess' function from *caret*. Model accuracy was tested using samples excluded by the random 20% split. Models were cross-validated using microbiome and soil health data from an independent CASH-based study of tillage and cover crop use ('Musgrave Farm study' data; n = 21) and from a study on mountain pastureland management ('Pastureland study' data; n = 198; BioProject: PRJEB9654) that measured soil properties related to the health metrics used in CASH (Lanzén et al., 2015) (data provided in Table S1).

The relative importance of features (i.e., taxa or ASVs) in the top performing models was determined using a leave-one-out approach, where the degree of decrease in prediction accuracy served as a measure of importance. The change in accuracy for each individual feature was assessed by excluding it from microbiome data input into existing models as previously described (Topçuoğlu et al., 2020). Each feature was evaluated once per model, and the change in accuracy was averaged across models trained on different sample splits (n_{SVM} = 50 and n_{RF} = 25). Features that caused a decrease in $\geq$0.2% in model accuracy were deemed 'important features.' Highly correlated features were evaluated as a single feature set to reduce computational efforts during importance testing. Correlated features were identified using Spearman's rank correlation with the 'rcorr' function from *Hmisc* (v. 4.2) (Harrell and Dupont, 2015) and grouped as a set if ρ $\geq$ 0.9. All analyses can be reproduced using the data and scripts provided in the Supplementary Data Package, and ML models can be obtained from the Open Science Foundation archive under the DOI: 10.17605/OSF.IO/6FC9B.

3. Results

Soil health and microbiome data were collected in a continent-scale survey of North American farmland soil to evaluate the predictive accuracy of ML classification and regression models trained on the composition of bacterial communities (a 16S rRNA gene survey). The Kappa statistic was used to compare classification accuracy because samples were not uniformly distributed across health categories (Table S2), since ratings for each health metric differed in their distributions (Fig. S1; Table S3). We first evaluated the impact of common pre-processing steps for microbiome data on the performance of SVM and RF classification and regression models (Fig. 1D). We evaluated the effects that read-depth normalization (rarefaction, cumulative sum scaling or proportioning), sparsity filtering (filtered vs. no filtering), and taxonomic resolution (aggregation of sequence count data at different taxonomic ranks) have on model performance. The best performing configurations were used in subsequent analyses.

3.1. Effects of normalization, sparsity filtering, and taxonomic resolution on prediction accuracy

Read-depth normalization and aggregation by taxonomic rank had significant effects on the accuracy of model predictions (Fig. 2). Normalizing by rarefying to an even depth significantly lowered the accuracy of both classification and regression models (ANOVA; Table S4). Models trained at lower taxonomic resolution were also significantly less accurate (Fig. 2; Table S4). However, normalizing by proportioning counts by sequencing depth (*i.e.*, proportional or cumulative sum scaling) did not significantly affect model accuracy for either ML algorithm or feature set (Table S5). In general, models trained on smaller feature sets were less accurate (Fig. 3AB), corresponding with a loss in taxonomic information and the total numbers of features due to rarefying and/or sparsity filtering (see statistical interactions in Table S4). Model accuracy did not significantly differ among models using larger feature sets (unfiltered, proportioned genus or ASV). ASV-based models performed best, though genus-based models performed better than corresponding ASV-based models in a sizeable number of instances (~17%). For these models, regression accuracy was significantly lower during cross-validation than in testing stages (Table S6), suggesting a degree of overfitting, which was not observed in classification models.

The efficacy and performance of the ML algorithms differed by task with SVM outperforming RF in classifying health categories while RF surpassed SVM in regression-based prediction of ratings (Fig. 2). The loss in accuracy at lower taxonomic resolution was more pronounced for SVM (ANOVA; $F_{category}$ = 342; F_{rating} = 122) than RF (ANOVA; $F_{category}$ = 40; F_{rating} = 41; Table S4). There were no significant differences in accuracy between cross-validation and testing stages (i.e., overfitting) between RF and SVM models. The variability in prediction accuracy was high for both SVM and RF due to inherent differences in the samples chosen (*i.e.*, 'split') for training and validation (Fig. 3C). Training time was significantly different between algorithms, completing in an order of minutes for SVM versus days for RF (Fig. 3D). The computational time to train SVM or RF models was correlated with feature set size (Pearson's r = 0.95; $p < 0.001$).

3.2. Predicting measures of soil health

The performance of microbiome-based ML models was evaluated for 12 health metrics, as well as tillage intensity, soil texture, and soil DNA yield (Fig. 1B). Based on the results described above, we used microbiome data resolved at the ASV or genus rank and normalized by proportion without sparsity filtering. SVM was exclusively used for classification due to its consistently better performance and lower computational demands relative to RF-based classification (Figs. 3D and 2), while both algorithms were used in regression-based modeling.

All models performed well at predicting health categories and ratings with the best models achieving a Kappa value of ~0.65 and a R^2 value of ~0.8, respectively (Fig. 4; full data in Fig. S2), especially for biological metrics. All models achieved their highest kappa or R^2 value for the target metric or soil property used in training (Table S7). Models exhibited a degree of accuracy for non-targeted metrics, particularly among biological metrics where five models achieved an accuracy of R^2 $\geq$ 0.4 for other non-targeted biological metrics (Table S7). Models for ACE protein had especially high off-target accuracy for active carbon (R^2 = 0.66) and total health rating (R^2 = 0.57). Classification-based models produced higher prediction accuracy for health categories ($\bar{x}_{kappa}$+0.14) than if categories were assigned post hoc from regression-based ML predictions (Fig. S3A). Post-hoc predictions were most reliable for categories in the middle of the soil health spectrum (Fig. S3B). Regression models predicted a narrower range of rating values than
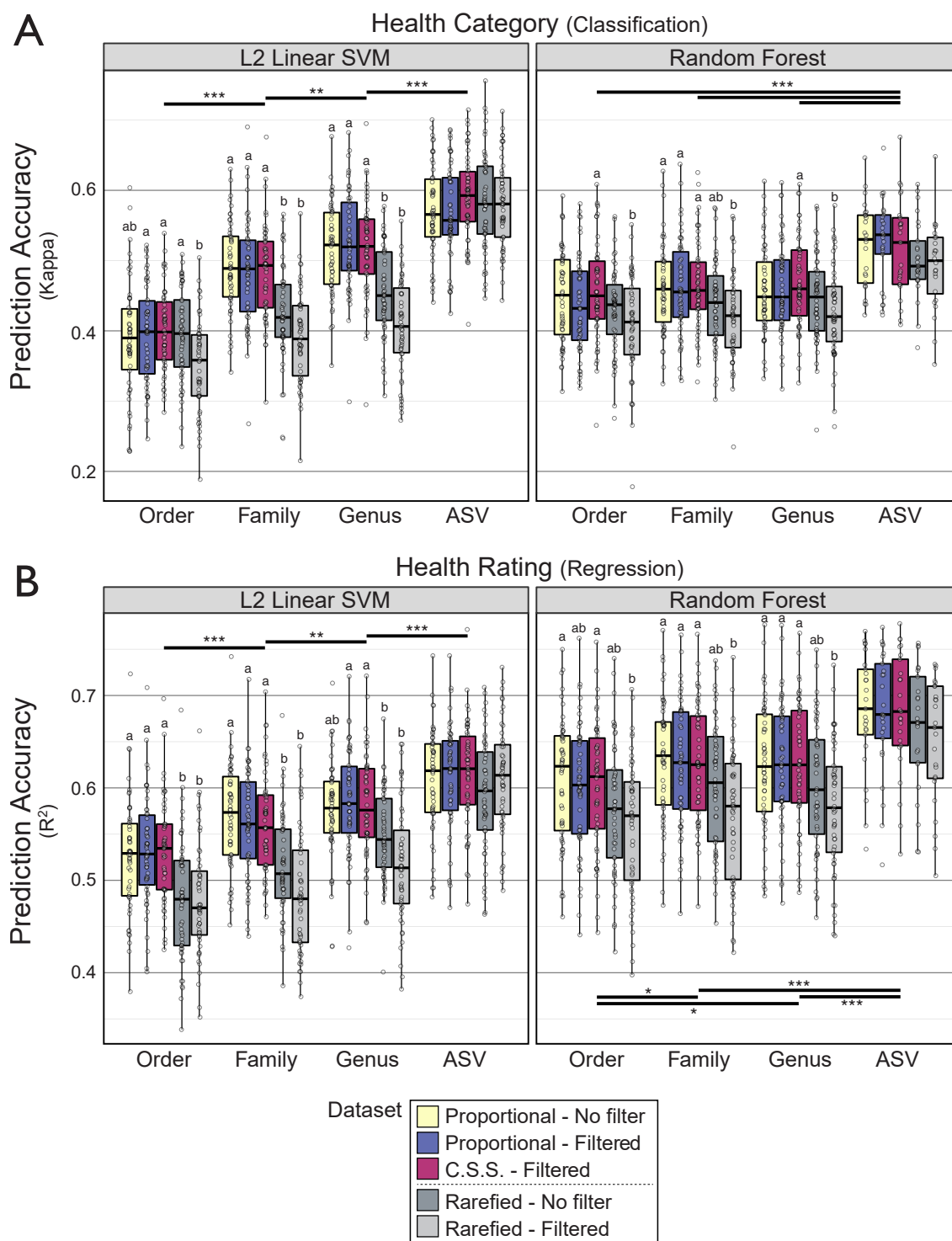
Fig. 2. The accuracy of support vector machine (SVM) and random forest (RF) models in predicting overall health category (A) or health rating (B) differed by the type of read-depth normalization, filtering, and taxonomic resolution of feature sets used in training. Models trained on data normalized by proportion of total read depth and using ASV-based count data, the most phylogenetically resolved data, were most accurate. SVM models outperformed RF in classification of health category, while RF performed better in regression. Each point corresponds to a model trained using a unique split of microbiome data in training/testing. Significant differences among taxonomic ranks are denoted with bars with asterisk representing: * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$). Pairwise significant differences among datasets within each rank are denoted by lettering ($p < 0.05$).

observed, with SVM capturing a broader range (82%) than RF (74%; Table S8). There was no correlation between accuracy and observed range (Pearson's r = 0.1; $p = 0.59$) or predicted range (Pearson's r = 0.08; $p = 0.67$).

3.3. Identifying the most predictive members of the soil microbiome

The ASV-based features contributing most to model performance were identified based on the change in prediction accuracy when each individual ASV was excluded. The exclusion of any ASV tended to cause

A — Feature Set Size

Normalization	Sparsity	Order	Family	Genus	ASV
Proportional	no filtering	370	642	1060	6798
Proportional	filtering	278	484	827	6699
CSS	filtering	278	484	827	6699
Rarefied	no filtering	301	522	856	5126
Rarefied	filtering	222	385	633	4000

B — Average Accuracy (R^2) in Health Rating Prediction

Normalization	Sparsity	Order	Family	Genus	ASV
Proportional	no filtering	0.526	0.569	0.576	0.61
Proportional	filtering	0.531	0.562	0.585	0.61
CSS	filtering	0.531	0.561	0.58	0.62
Rarefied	no filtering	0.479	0.516	0.55	0.59
Rarefied	filtering	0.48	0.486	0.517	0.61

C — Model Variance (std. deviation)

	SVM	RF
Kappa	0.065	0.063
R^2	0.056	0.065

D — Training Time (hrs)

	SVM	RF
ASV	0.34	450
Genus	0.1	46
Family	0.08	18
Order	0.05	11

Fig. 3. Preliminary results on model performance used to determine the best approach for preparing microbiome data for machine learning. In (A), feature set size was greatest at the highest taxonomic resolution when filtering was not applied and when proportion-based normalization methods were used, as opposed to rarefying. In (B), models trained on larger feature sets produced higher accuracy (R^2), as evident in the prediction of health ratings trends in SVM-regression. In (A) and (B), cells have been colored green by increasing value, illustrating the correspondence between feature set size and accuracy. In (C), prediction accuracy varied by upwards of 10% based on differences in the splitting of training and testing data with classification-based SVM and regression-based RF models exhibiting the greatest variability. In (D), the training time of RF and SVM algorithms differed substantively with the former taking upwards of two weeks to complete with the largest feature sets.
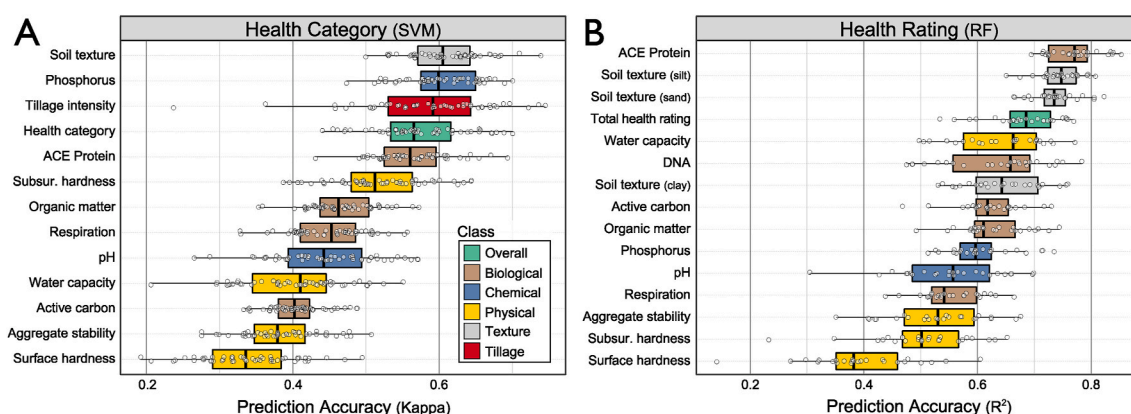

Fig. 4. A ranking of the accuracy of ASV-based models for predicting health categories (A) and health ratings (B), as well as soil texture, tillage and DNA yields. Classification results from SVM are displayed in (A) and regression results from RF in (B). For SVM-based regression results, consult Fig. S2. Models for potassium and minor elements had low accuracy (not shown) and the latter were excluded from downstream analyses due to limited utility and predictive range (Table S8).

a greater decrease in the accuracy of classification (0.04% on average) than regression (0.003%). In RF-based regression models, the exclusion of individual ASVs caused only a slight decrease in accuracy (0.01%) for most ASVs examined in this way. Exclusion of some ASVs produced a much larger decrease in accuracy ($\geq$0.2%), and these taxa (n = 1320; 19% of total ASVs) were deemed 'important features' (Table S9). ASVs designated as important features were present in a greater proportion of samples in our collection ($\bar{x}$ = 13%; max = 97%; min = 1.7%) versus all other ASVs ($\bar{x}$ = 4.5%; max = 72%; min = 1.7%). Biological health metrics shared the greatest overlap in important features (21%), with fewer common among physical metrics (12%) and none shared among chemical metrics. Important features for predicting the overall health rating were also commonly identified as important in models for ACE protein and active carbon (Fig. 5A). Soil texture and organic matter shared many important features, as did metrics related to microbial biomass (DNA and respiration) and aggregate stability. The most important ASV-based features for predicting health rating (Fig. 5B) exhibited distinct trends in relative abundance corresponding with the overall health rating (Fig. 5C).

3.4. Cross-study validation

The performance of models was cross-validated using microbiome and soil health data from two independent studies. In the first dataset, CASH and microbiome data (n = 21) were collected from field soils at Musgrave Farm (Aurora, NY) as part of on-going research on the effects of tillage and cover crop use on soil health (Jernigan et al., 2020). Approximately 20% of ASVs in the Musgrave soil bacterial community were also present in our North America-wide dataset (n = 1200/6800 ASVs). SVM and RF models were predictive of several health metrics (R^2 = 0.25–0.35), achieving the best accuracy for health rating (Fig. 6A). In contrast, models input with randomly permuted data had no accuracy (i. e., non-significant R^2 values, data not shown). A small set of the important features identified in the initial testing data were also important for the prediction accuracy of Musgrave Farm health ratings ($_{SVM}$ = 18 and $_{RF}$ = 70 ASVs; Fig. 6B) and other metrics (Table S10).

Next, we cross-validated models using microbiome data (n = 198) from a study linking mountain pastureland management practices to soil health (details in Table S1). This dataset contained information on soil pH, respiration, organic matter, potassium, penetrability, compaction, and yield (tons · ha^{-1} · year^{-1}) (Lanzén et al., 2015). We expected microbiome-based ML models trained on related CASH metrics to predict differences in these pastureland soil properties. A large proportion of ASVs were common between the pastureland study and the continent-wide data collection (2800/6800 ASVs; 41%). Models had a reasonable accuracy for predicting soil pH and yield (R^2 > 0.3), low
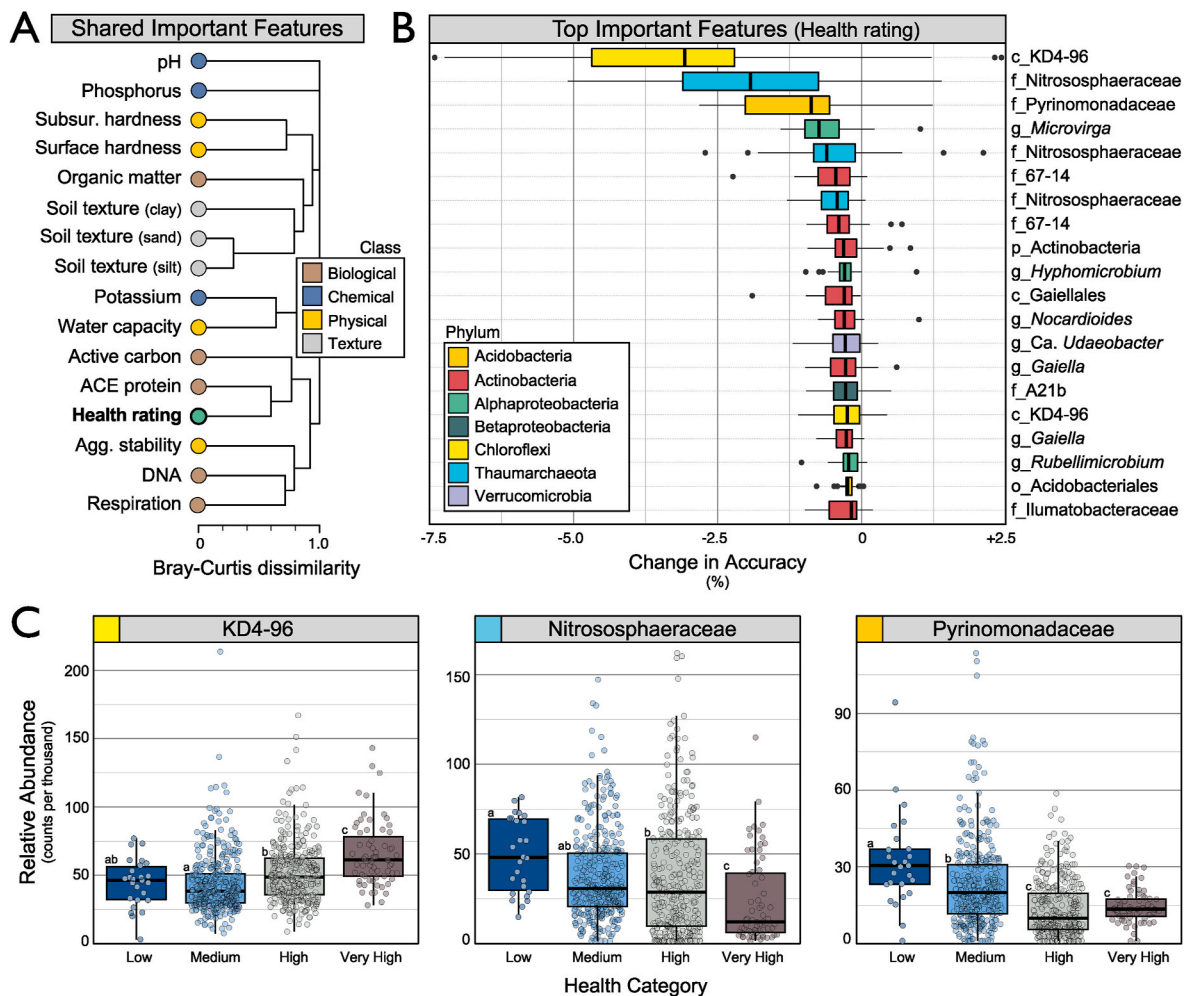
Fig. 5. Important ASVs were determined as those which caused a greater than 0.2% decrease in accuracy when excluded from test data. Important ASVs were shared between models of health metrics. In (A), these similarities are represented by a dendrogram with branch lengths corresponding to the Bray-Curtis dissimilarity distance calculated from presence/absence of important features. Leaves are colored by the class of health metric. In (B), a ranking of the top important ASVs in RF model-based predictions of total health rating. The lowest supported taxonomic classification ('g_': genus, 'f_': family, 'o_': order, and 'c_': class) for each ASV is provided along the y-axis and bars are colored by phylum. In (C), trends in the relative abundance of the three most important ASVs displayed in (B) across total health categories. Statistical differences were based on pairwise Kruskal-Wallis tests ($p < 0.05$).

accuracy for predicting respiration, organic matter, and potassium content ($R^2 < 0.2$), and no accuracy for predicting penetrability, compaction, or induced respiration (Table S11). SVM and RF models had a 0 and 5% false positive rate (i.e., $R^2 > 0$ and $p < 0.05$) based on randomly permuted data, respectively. The accuracy of models fell within a relatively narrow range (Fig. 6C) compared to Musgrave Farm predictions (Fig. 6A). Accordingly, we suspected that a relatively small set of shared features drove variation in model performance. We identified twenty-one important ASVs present in all models predictive of pastureland yield (Table S12) and predictions based solely on this 'minimum set' achieved a higher accuracy than when the full microbiome data was used (Fig. 6C). The minimum set were most important for predicting yield using models trained on biological metrics, with predictions based on models trained on health ratings exhibiting a clear correlation with observed yields (Fig. 6D). The total relative abundance of the taxa in the minimum set was correlated with yield data (Pearson's r = 0.45; $p < 0.001$) compared to randomly sampled taxa (average r = 0.16; perm = 9999), providing evidence for the basis of model accuracy.

3.5. Characterization of microbial diversity

We characterized the diversity, evenness, and composition of bacterial communities to provide context for interpreting the performance

of models. The soils used in our analyses were collected from farmland across the continental USA and Canada, representing a random set of samples routinely processed in soil health testing facilities. Community composition varied primarily due to geographical region (PERMANOVA; $R^2 = 0.59$; $p < 0.001$), tillage intensity ($R^2 = 0.026$), soil texture ($R^2 = 0.011$), and health category ($R^2 = 0.006$; Fig. 7; Table S13). Soil microbiomes associated with higher health categories shared more phylogenetic similarity than soil microbiomes associated with lower health categories (Fig. 7C). No significant differences were observed in Shannon diversity (Fig. 7D) or Pielou's evenness (not shown) among health categories. Among soil health classes, most variation in microbiome composition was explained by chemical properties (Mantel test; r = 0.28; $p < 0.001$), followed by biological properties (r = 0.16; $p < 0.001$), and physical properties (r = 0.09; $p = 0.002$). Individually, ACE protein, active carbon and available phosphorus ratings explained the greatest variation in community composition (PERMANOVA; $R^2 = 0.04$, 0.02 and 0.02, respectively; Table S13).

4. Discussion

Our study demonstrated that information derived from 16S rRNA gene sequencing ('microbiome data') can predict properties of soil health based on an operationally relevant sample set. ML models
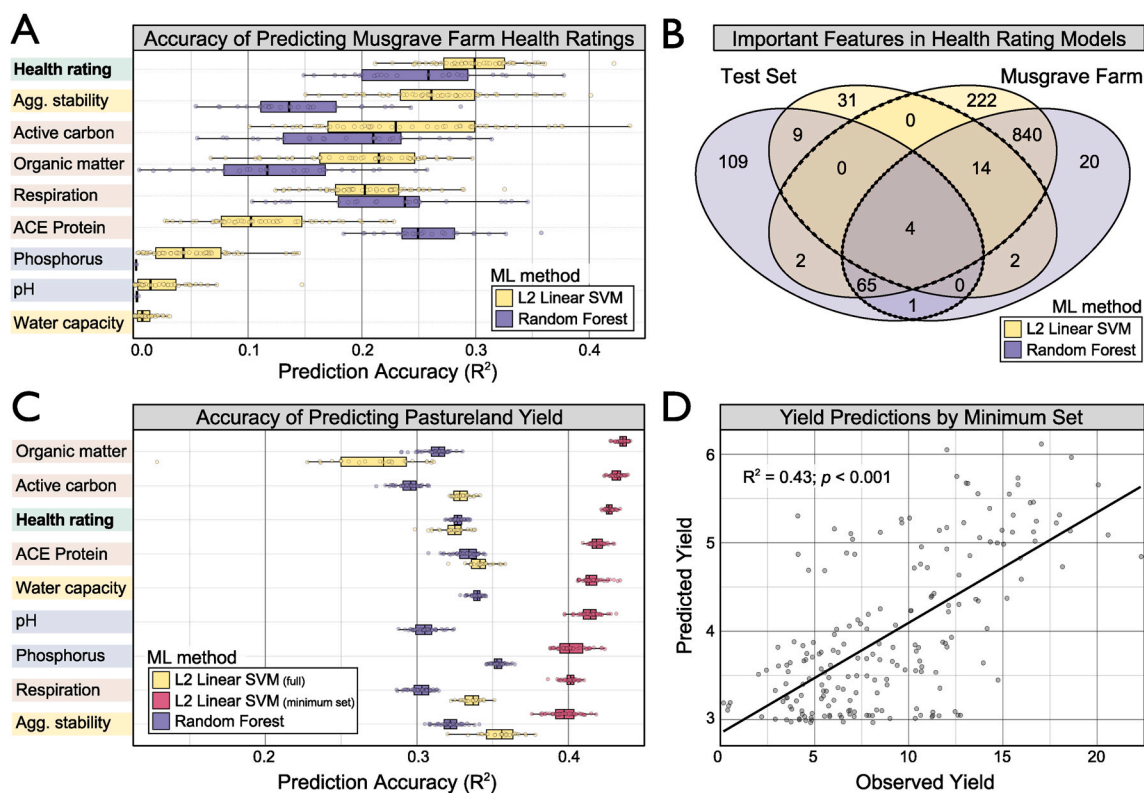
Fig. 6. The performance of ASV-based models was cross-validated using microbiome data from two independent studies: the Musgrave Farm study (A and B), which used identical measures of soil health metrics, and the Pastureland study (C and D), which used comparable soil properties (Lanzén et al., 2015). In (A), SVM and RF regression models were capable of predicting health ratings for Musgrave Farm soils with a significant, albeit low, degree of accuracy. The y-axis labels indicate which health metric was used in training the regression models. In (B), health rating was the most accurately predicted metric, yet few important features were shared between the test set and Musgrave farm set of either SVM or RF models. Line patterns highlight the overlap between SVM models (broad dotting) and between RF models (narrow dotting). In (C), SVM and RF regression models were capable of predicting pastureland yield (tons $\cdot$ ha^{-1} $\cdot$ year^{-1}) with improved accuracy when using a subset of 21 important features common to all models (i.e., the 'minimum set'). In (D), a scatter plot illustrating the relationship between predicted and observed pastureland yield when the minimum set was input to SVM models trained on soil health ratings.

achieved fair to good accuracy in predicting biological health metrics with a modest-sized training set that included diverse soils representing broad differences in geography, farming practice, and soil type. This success is all the more remarkable since the continental-scale microbiome data was highly variable, exhibiting 100 times more variation ($R^2_{PERMANOVA}$) due to geographic soil origin than due to soil health rating. Cross-validation revealed that models trained on a continent-wide dataset can predict trends for individual farms at the landscape scale even though prediction accuracy was low. We expect that the development of regional scale predictive models, or cropping-system specific models, might be valuable for improving prediction accuracy. Our results showed the promise of using microbiome-ML to obtain information about a breadth of relevant soil health conditions with a single measurement.

4.1. How well can the microbiome predict diverse properties of soil health?

The accuracy of model predictions varied largely by the individual health metric. Models were predictive across the full range of soil health classes with at least one metric among the best predicted in each of the physical (available water capacity), chemical (P), and biological (ACE protein) classes. However, models that predicted biological metrics generally fared better, while those for chemical and physical metrics fared worse, suggesting there are differences in the predictive capacity of the microbiome (Fig. 4; Fig. 6AC). Microbiome data could accurately predict certain physical properties, like soil texture and tillage intensity, but these are not health metrics per se and can be easily determined in

other ways. Yet, physical metrics like aggregate stability or hardness were more difficult to assess with microbiome data, at least using 16S rRNA gene-based approaches. Notably, models that predicted microbial biomass and activity (i.e., DNA yield and respiration) had a degree of accuracy for aggregate stability and shared many important features (Table S7; Fig. 5A). Aggregate stability is heavily influenced by microbial biomass and activity (Lehmann et al., 2017), suggesting the possibility of further refining underperforming models.

The accuracy of microbiome-based ML for predicting most biological metrics illustrates that the soil microbiome can serve as a common denominator for multiple health metrics. ACE protein rating was the most accurately predicted measure of soil health (mean $R^2 = 0.77$). Models for ACE protein were also predictive of active carbon and total health ratings ($R^2 = 0.66$ and 0.57, respectively) with all three metrics sharing many important features (i.e., the ASVs contributing most to model accuracy; Fig. 5A). ACE protein is a measure of the insoluble, refractory proteinaceous content of soil and can be predictive of nitrogen mineralization rates important for plant nutrition (Geisseler et al., 2019; Hurisso et al., 2018). Active carbon is similarly correlated with particulate forms of soil carbon (Weil et al., 2003). The common accuracy of models for these metrics may reflect similar adaptations in populations of bacterial decomposers that access insoluble forms of organic matter (Wilhelm et al., 2021) and indicates that overall soil health status is particularly evident in the microbiome associated with the quality of soil organic matter. This observation is consistent with the results from a comprehensive survey of health metrics (Norris et al., 2020) which found active carbon was the best single predictor of most soil health metrics (Liptzin et al., 2020). The common accuracy of biological
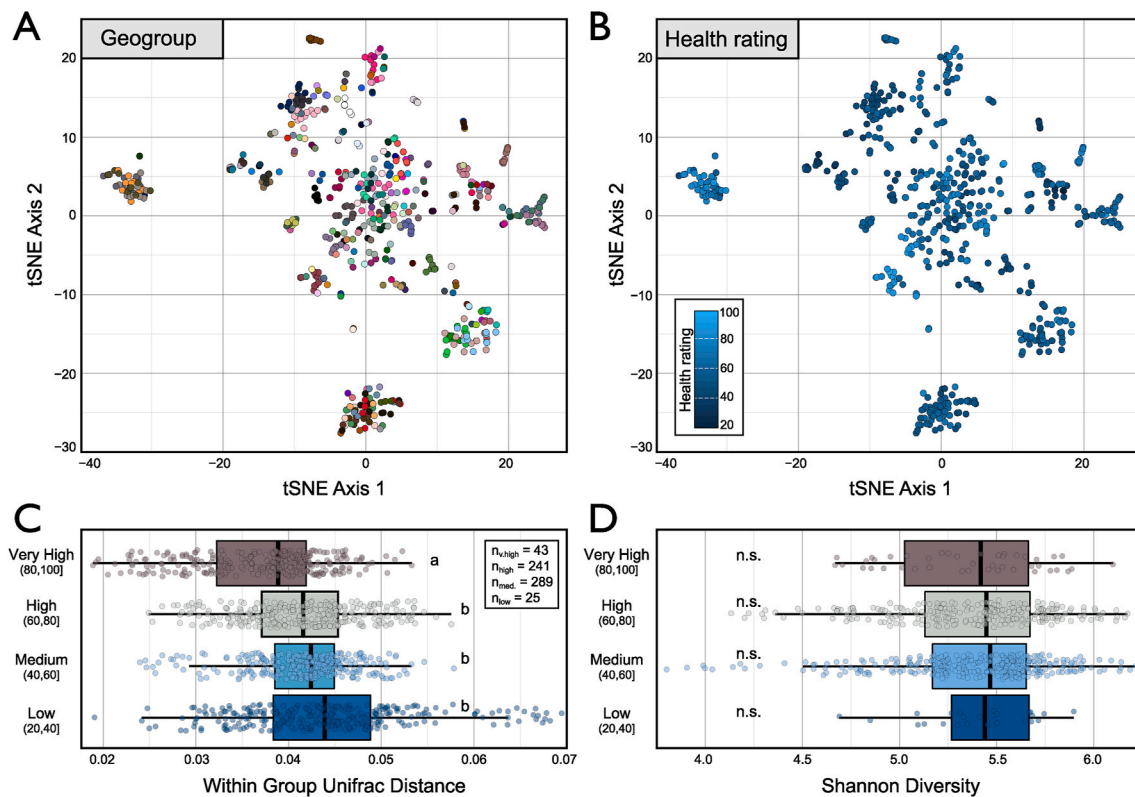
Fig. 7. Measures of alpha- and beta-diversity revealed the extent of heterogeneity in the soil microbiome data used in training ML models. Bacterial communities (n = 598) exhibited a relatively low degree of clustering when ordinated by a t-SNE transformation, with bacterial communities differing primarily by geographical location (A) and, to a lesser extent, health rating (B). In (C), the degree of phylogenetic similarity shared between microbiomes within health categories differed based on Unifrac distance. Microbiomes within the 'very high' category shared a greater phylogenetic similarity than others while those within the 'low' class were, on average, less phylogenetically similar than other categories. In (D), the alpha-diversity of microbiomes did not significantly differ among health categories according to the Shannon diversity index.

models underscores that certain health metrics may be redundant and draws attention to the underlying basis for the accuracy of microbiome-ML predictions.

4.2. What properties of the microbiome underlie model accuracy?

We expected the microbiome to predict conventional measures of soil health based on the diverse and distinguishing traits of its members, including their varied metabolism and life-history strategies. The majority of ASVs important for model accuracy belonged to poorly characterized unclassified or candidate groups, limiting our ability to broadly infer the underlying ecological basis of models. Yet, among those most important for predicting total health rating were ASVs classified to groups linked to soil properties relevant to soil health. For example, several of the most important features were classified as ammonia-oxidizing *Nitrososphaeraceae* (Kerou and Schleper, 2015), which were indicative of low health rating (Fig. 5C), and which might be indicative of soils whose fertility is heavily dependent on mineral fertilizers (Dong et al., 2021; Zhalnina et al., 2013). In addition, other ASVs indicative of low health rating (Fig. 5C; Table S9) included those identified as *Pyrinomonadaceae*, a group of thermophilic Acidobacteria (Lee et al., 2015) found in semiarid soils having low organic matter content (Ivanova et al., 2020), and those identified as *Massilia*, a group associated with degraded agricultural soils (Zhang et al., 2020). While a full analysis of these relationships is beyond the scope of our study, these examples demonstrate the kinds of ecological information underlying model performance and show that an ML approach can help reveal relationships between specific taxa and soil health.

Models performed best when trained on microbiome data resolved at the ASV-level (Fig. 2). Yet, most ASVs (~90%) contributed minimally to

model performance. Those that had significant impacts on model performance (*i.e.*, important features) tended to be widespread, present in approximately three-fold more samples, and tended to be predictive of multiple health metrics, with ~41% impacting the accuracy of more than one metric. The non-specificity of important ASVs likely reflects both the interrelatedness of health metrics as well as the complex ecology of soil microbes, which might be indicators of multiple interrelated soil properties. An intriguing consequence of the shared predictive value of ASVs was the capability of models trained on CASH health metrics to predict yield information from an independent study of pastureland management (Fig. 6D). A relationship between CASH health metrics and long-term crop yields has previously been shown, with ACE protein and active carbon ratings the two strongest correlates (van Es and Karlen, 2019). Consistent with this finding, the microbiome-ML models predicting ACE protein and active carbon were also among the most accurate in predicting pastureland yield (Fig. 6C). Furthermore, correlations between members of the soil microbiome and crop yield was recently demonstrated for wheat (Yergeau et al., 2020), with several of the major indicator groups (Blastocatellales, *Gaiella* and Candidatus *Udeaobacter*) among the important features in our ML models, and in the original pastureland study (Lanzén et al., 2015). The existence of common microbial predictors of soil health and yield in cropland and pastureland systems is noteworthy given that natural grasslands are considered a benchmark of high soil health (Glover et al., 2010; Maharjan et al., 2020). We do not expect the relationships between microbiome, soil health and crop yield to be so simple or direct, but these results confirm that health metrics and the soil microbiome can reflect conditions where plants are more productive.

Model accuracy did not necessarily depend on the information obtained from the whole soil microbiome. During validation with

pastureland microbiome data, superior accuracy was achieved when models were provided with data from just 21 commonly important ASVs. Inputting fewer, more informative features may have helped reduce the impact of model overfitting to features whose association with soil health could vary – possibly due to differences between pastureland and farmland. However, models trained on larger feature sets had the highest accuracy with test data, where samples spanned a larger spatial scale (Fig. 3AB). In this case, training with a larger feature set may have increased the chances of capturing features important to a given soil type or region. Larger feature sets may provide redundancy needed to address variation in the occurrence/distribution of taxa. This redundancy was apparent in the limited overlap of ASVs identified as important for predicting health rating from Musgrave Farm versus our dataset (Fig. 6B). In our analyses, feature set size varied as a function of taxonomic resolution, sparsity filtering and normalization method. A more comprehensive account of the factors underlying the relationship between feature set size and accuracy remains to be determined.

4.3. What challenges remain?

Many of the challenges we identified in using microbiome-based ML to predict soil health metrics could be remedied with greater sequencing depth, greater taxonomic resolution, and a larger set of samples spanning the full range in health status. Accuracy varied greatly depending on the initial split of training and test set, reflecting the relatively small size of our collection (n = 598). The decrease in accuracy in models trained on rarefied data was due to a loss in total features, which could be avoided with greater sequencing effort. Regression models consistently predicted a narrower range of values than observed, leading to poor accuracy at either end of the soil health spectrum. Performance will improve with better representation of microbiome data from soils at the health extremes. For example, models predicting pH rating had surprisingly low accuracy (Fig. 4) given pH is a strong determinant of bacterial community structure (Rousk et al., 2010; Tripathi et al., 2018). The low accuracy was likely due to the narrow range in the pH rating of soils in our dataset, which were the narrowest of all metrics (Fig. S1), because farmland is rarely situated on soils with extreme pH. Furthermore, the fact that many soils with low health ratings failed entirely during sequencing exacerbates the challenge of achieving representation of microbiomes from low (n = 25) and very low (n = 0) health soils, which were also the most phylogenetically diverse (Fig. 7C). Once a broader representation of soil health status has been obtained, microbiome-ML would ideally use regression modeling to predict health ratings, which would subsequently be assigned to a health category for ease of interpretation by farm managers. At present, the post-hoc categorization of regression model predictions was less accurate than classification models.

5. Conclusions

Our study demonstrated that the information contained in a 16S rRNA gene-based survey of the soil microbiome can be used to predict soil health metrics currently used by farm managers. Our findings indicated that a microbiome-ML approach has high potential to supplement or replace biological health metrics or to serve as an integrated measure of biological health given the overlap in important features among biological models. Predictions were generally weaker for chemical and physical metrics, though targeted refinements may be possible for important physical metrics like aggregate stability. Meanwhile, the low cost and long history of chemical soil testing makes a microbiome-based substitute less important. We conclude that standard chemical tests and a microbiome-ML approach may serve as a low-cost measure for soil health while providing additional information about soil functioning, according to our growing understanding of the ecological and functional traits of members of the soil microbiome.

We are cautiously optimistic about the potential for this technology.

We expect accuracy will improve as the number of observations increases, and where models are trained on regional or management specific data. ML models trained on microbiome data tailored to geography, soil type and cropping system will likely perform better, given the effects of plant legacy (Schmid et al., 2021) and regional differences on the soil microbiome (Gschwend et al., 2021). The accuracy of our models was remarkable given the scale and geographic variation in our sample collection, which was intended to capture the typical breadth of samples processed at a soil health testing facility. Our study also highlighted the need to develop high-throughput methods for collecting microbiome data that overcome challenges in processing low health soils, which tended to have low DNA yields. Our results did not indicate a clear preference for which ML algorithms are best suited for predicting soil health from microbiome data. L2-linear SVM models outperformed RF in several important ways, including in classification-based predictions, cross-validation and training time, while RF models outperformed in regression-based predictions. These questions may be more thoroughly addressed by the adoption of microbiome and ML methods in the current development of soil health frameworks (Rinot et al., 2019; Stewart et al., 2018) and in digital agriculture (Kinoshita et al., 2012), either as a single measure or in combination with more established measures.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.soilbio.2021.108472.

References

Bhadha, J., Khatiwada, R., Galindo, S., Xu, N., Capasso, J., 2018. Evidence of soil health benefits of flooded rice compared to fallow practice. Sustainable Agriculture Research 7, 31. https://doi.org/10.5539/sar.v7n4p31.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K. Bin, Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G. I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-

Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nature Methods 13, 581–583. https://doi.org/10.1038/nmeth.3869.

Cha, J.Y., Han, S., Hong, H.J., Cho, H., Kim, D., Kwon, Y., Kwon, S.K., Crüsemann, M., Lee, Y.B., Kim, J.F., Giaever, G., Nislow, C., Moore, B.S., Thomashow, L.S., Weller, D. M., Kwak, Y.S., Crusemann, M., Bok Lee, Y., Kim, J.F., Giaever, G., Nislow, C., Moore, B.S., Thomashow, L.S., Weller, D.M., Kwak, Y.S., 2016. Microbial and biochemical basis of a *Fusarium* wilt-suppressive soil. The ISME Journal 10, 119–129. https://doi.org/10.1038/ismej.2015.95.

Chamizo, S., Mugnai, G., Rossi, F., Certini, G., De Philippis, R., 2018. Cyanobacteria inoculation improves soil stability and fertility on different textured soils: gaining insights for applicability in soil restoration. Frontiers in Environmental Science 6. https://doi.org/10.3389/fenvs.2018.00049.

Chang, H.X., Haudenshield, J.S., Bowen, C.R., Allen, R., Iii, W., Parnell, J.J., Hartman, G. L., 2017. Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity. Frontiers in Microbiology 8, 519. https://doi.org/10.3389/fmicb.2017.00519.

Chen, L., Reeve, J., Chen, L., Reeve, J., Zhang, L., Huang, S., 2018. GMPR: a robust normalization method for zero-inflated count data with application to microbiome sequencing data. PeerJ 6, e4600. https://doi.org/10.7717/peerj.4600.

Clark, I.M., Hirsch, P.R., 2008. Survival of bacterial DNA and culturable bacteria in archived soils from the Rothamsted Broadbalk experiment. Soil Biology and Biochemistry 40, 1090–1102. https://doi.org/10.1016/j.soilbio.2007.11.021.

Core Team, R., 2020. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Degrune, F., Theodorakopoulos, N., Colinet, G., Hiel, M.P., Hartmann, M., Bodson, B., Taminiau, B., Daube, G., Vandenbol, M., Hartmann, M., 2017. Temporal dynamics of soil microbial communities below the seedbed under two contrasting tillage regimes. Frontiers in Microbiology 8. https://doi.org/10.3389/fmicb.2017.01127.

Delgado-Baquerizo, M., Bissett, A., Eldridge, D.J., Maestre, F.T., He, J.Z., Wang, J.T., Hamonts, K., Liu, Y.R., Singh, B.K., Fierer, N., 2017. Palaeoclimate explains a unique proportion of the global variation in soil bacterial communities. Nature Ecology and Evolution 1, 1339–1347. https://doi.org/10.1038/s41559-017-0259-7.

Dong, W., Li, X., Wang, E., Liu, X., Wang, M., Song, A., Yin, H., Fan, F., 2021. Linking microbial taxa and the effect of mineral nitrogen forms on residue decomposition at the early stage in arable soil by DNA-qSIP. Geoderma 400. https://doi.org/10.1016/j.geoderma.2021.115127.

Doran, J.W., 2002. Soil health and global sustainability: translating science into practice. Agriculture, Ecosystems & Environment 88, 119–127. https://doi.org/10.1016/S0167-8809(01)00246-8.

Fine, A.K., Schindelbeck, R.R., van Es, H.M., Schindelbeck, R.R., 2017. Statistics, scoring functions, and regional analysis of a comprehensive soil health database. Soil Science Society of America Journal 81, 589. https://doi.org/10.2136/sssaj2016.09.0286.

Francioli, D., Schulz, E., Lentendu, G., Wubet, T., Buscot, F., Reitz, T., Nicol, G.W., 2016. Mineral vs. organic amendments: microbial community structure, activity and abundance of agriculturally-relevant microbes are driven by long-term fertilization. Strategies 7, 1446. https://doi.org/10.3389/fmicb.2016.01446.

Frost, P.S.D., van Es, H.M., Rossiter, D.G., Hobbs, P.R., Pingali, P.L., 2019. Soil health characterization in smallholder agricultural catchments in India. Applied Soil Ecology 138, 171–180. https://doi.org/10.1016/j.apsoil.2019.02.003.

Geisseler, D., Miller, K., Leinfelder-Miles, M., Wilson, R., 2019. Use of soil protein pools as indicators of soil nitrogen mineralization potential. Soil Science Society of America Journal 83, 1236–1243. https://doi.org/10.2136/sssaj2019.01.0012.

Gholoubi, A., Emami, H., Alizadeh, A., 2018. Soil quality change 50 years after forestland conversion to tea farming. Soil Research 56, 509–517. https://doi.org/10.1071/SR18007.

Giraudoux, P., Antonietti, J.-P., Beale, C., Pleydell, D., Treglia, M., 2018. Pgirmess: Spatial Analysis and Data Mining for Field Ecologists.

Glover, J.D., Culman, S.W., DuPont, S.T., Broussard, W., Young, L., Mangan, M.E., Mai, J. G., Crews, T.E., DeHaan, L.R., Buckley, D.H., Ferris, H., Turner, R.E., Reynolds, H.L., Wyse, D.L., 2010. Harvested perennial grasslands provide ecological benchmarks for agricultural sustainability. Agriculture, Ecosystems & Environment 137, 3–12. https://doi.org/10.1016/j.agee.2009.11.001.

Gschwend, F., Hartmann, M., Hug, A.S., Enkerli, J., Gubler, A., Frey, B., Meuli, R.G., Widmer, F., 2021. Long-term stability of soil bacterial and fungal community structures revealed in their abundant and rare fractions. Molecular Ecology 30, 4305–4320. https://doi.org/10.1111/mec.16036.

Harrell, F., Dupont, C., 2015. Hmisc: harrell miscellaneous. R package.

Hartmann, M., Frey, B., Mayer, J., Ma, P., Mäder, P., Widmer, F., Ma, P., 2015. Distinct soil microbial diversity under long-term organic and conventional farming. The ISME Journal 9, 1177–1194. https://doi.org/10.1038/ismej.2014.210.

Hermans, S.M., Buckley, H.L., Case, B.S., Curran-Cournane, F., Taylor, M., Lear, G., 2020. Using soil bacterial communities to predict physico-chemical variables and soil quality. Microbiome 8, 1–13. https://doi.org/10.1186/s40168-020-00858-1.

Hurisso, T.T., Moebius-Clune, D.J., Culman, S.W., Moebius-Clune, B.N., Thies, J.E., Es, H. M., 2018. Soil protein as a rapid soil health indicator of potentially available organic nitrogen. Agricultural & Environmental Letters 3, 180006. https://doi.org/10.2134/ael2018.02.0006.

Ivanova, A.A., Zhelezova, A.D., Chernov, T.I., Id, S.N.D., 2020. Linking ecology and systematics of Acidobacteria: distinct habitat preferences of the Acidobacteriia and

Blastocatellia in tundra soils. PLoS One 15, e0230157. https://doi.org/10.1371/journal.pone.0230157.

Ivanova, E.A., Korvigo, I.O., Aparin, B.F., Chirak, E.L., Pershina, V., Romaschenko, N.S., Provorov, N.A., Andronov, E.E., 2017. The preservation of microbial DNA in archived soils of various genetic types 1–14. https://doi.org/10.1371/journal.pon e.0173901.

Jernigan, A.B., Wickings, K., Mohler, C.L., Caldwell, B.A., Pelzer, C.J., Wayman, S., Ryan, M.R., 2020. Legacy effects of contrasting organic grain cropping systems on soil health indicators, soil invertebrates, weeds, and crop yield. Agricultural Systems 177, 102719. https://doi.org/10.1016/j.agsy.2019.102719.

Kerou, M., Schleper, C., 2015. Nitrososphaera. In: Bergey's Manual of Systematics of Archaea and Bacteria, pp. 1–10.

Kinoshita, R., Moebius-Clune, B.N., van Es, H.M., Hively, W.D., Bilgili, A.V., 2012. Strategies for soil quality assessment using visible and near-infrared reflectance spectroscopy in a Western Kenya chronosequence. Soil Science Society of America Journal 76. https://doi.org/10.2136/sssaj2011.0307.

Knights, D., Parfrey, L.W., Zaneveld, J., Lozupone, C., Knight, R., 2011. Human-associated microbial signatures: examining their predictive value. Cell Host & Microbe 10, 292–296. https://doi.org/10.1016/j.chom.2011.09.003.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., Schloss, P.D., 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the Miseq Illumina sequencing platform. Applied and Environmental Microbiology 79, 5112–5120. https://doi.org/10.1128/AEM.01043-13.

Krijthe, J.H., 2015. Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation.

Kuhn, M., 2008. Building predictive models in R using the caret package. Journal of Statistical Software 28, 1–26. https://doi.org/10.18637/jss.v028.i05.

Lanzén, A., Epelde, L., Garbisu, C., Anza, M., Martín-Sánchez, I., Blanco, F., Mijangos, I., 2015. The community structures of prokaryotes and fungi in mountain pasture soils are highly correlated and primarily influenced by pH. Frontiers in Microbiology 6, 1–15. https://doi.org/10.3389/fmicb.2015.01321.

Lauber, C.L., Zhou, N., Gordon, J.I., Knight, R., Fierer, N., 2010. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS Microbiology Letters 307, 80–86. https://doi.org/10.1111/j.1574-6968.2010.01965.x.

Lee, K.C.Y.Y., Morgan, X.C., Power, J.F., Dunfield, P.F., Huttenhower, C., Stott, M.B., 2015. Complete genome sequence of the thermophilic Acidobacteria, *Pyrinomonas methylaliphatogenes* type strain K22T. Standards in Genomic Sciences 10, 1–8. https://doi.org/10.1186/s40793-015-0099-5.

Lehmann, A., Zheng, W., Rillig, M.C., 2017. Soil biota contributions to soil aggregation. Nature Ecology & Evolution 1. https://doi.org/10.1038/s41559-017-0344-y.

Lehmann, J., Bossio, D.A., Kögel-Knabner, I., Rillig, M.C., 2020. The concept and future prospects of soil health. Nature Reviews Earth and Environment 1, 544–553. https://doi.org/10.1038/s43017-020-0080-8.

Liptzin, D., Bean, G.M., Cappellazzi, S., Cope, M., Greub, K.L.H., Norris, C.E., Rieke, E., Tracy, P.W., Morgan, C.L.S., Honeycutt, C.W., 2020. Effects of Climate, Intrinsic Soil Properties, and Management on Soil Health Indicators of Carbon Dynamics. AGU, pp. B114–B110004.

Liu, L., Huang, X., Zhao, J., Zhang, J., Cai, Z., Huang, X., Zhao, J., Zhang, J., 2019. Characterizing the key agents in a disease-suppressed soil managed by reductive soil disinfestation. Applied and Environmental Microbiology 85. https://doi.org/10.1128/aem.02992-18 e02992-18.

Liu, R., Gu, Y., Shen, M., Li, H., Zhang, K., Wang, Q., Wei, X., Zhang, H., Wu, D., Yu, K., Cai, W., Wang, G., Zhang, S., Sun, Q., Huang, P., Wang, Z., 2020. Predicting postmortem interval based on microbial community sequences and machine learning algorithms. Environmental Microbiology 1–19. https://doi.org/10.1111/1462-2920.15000, 00.

Lozupone, C., Lladser, M.E., Knights, D., Stombaugh, J., Knight, R., 2011. UniFrac: an effective distance metric for microbial community comparison. The ISME Journal 5, 169–172. https://doi.org/10.1038/ismej.2010.133.

Maharjan, B., Das, S., Acharya, B.S., 2020. Soil Health Gap: a concept to establish a benchmark for soil health management. Global Ecology and Conservation 23, e01116. https://doi.org/10.1016/j.gecco.2020.e01116.

McKnight, D.T., Huerlimann, R., Bower, D.S., Schwarzkopf, L., Alford, R.A., Zenger, K.R., 2019. Methods for normalizing microbiome data: an ecological perspective. Methods in Ecology and Evolution 10, 389–400. https://doi.org/10.1111/2041-210X.13115.

McMurdie, P.J., Holmes, S., 2013. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8. https://doi.org/10.1371/journal.pone.0061217.

Mendes, R., Kruijt, M., Bruijn, I. De, Dekkers, E., Voort, M. Van Der, Bakker, P.A.H.M., Raaijmakers, J.M., 2011. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science 322, 1097–1100. https://doi.org/10.1126/science.1203980.

Moebius-Clune, B.N., Moebius-Clune, D.J., Gugino, B.K., Idowu, O.J., Schindelbeck, R.R., Ristow, A.J., van Es, H.M., Thies, J.E., Shayler, H.A., McBride, M.B., Kurtz, K.S.M., Wolfe, D.W., Abawi, G.S., 2017. Comprehensive Assessment of Soil Health - the Cornell Framework Manual, third ed. Cornell University, Ithaca, NY.

Morton, J.T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L.S., Edlund, A., Zengler, K., Knight, R., 2019. Establishing microbial composition measurement standards with reference frames. Nature Communications 10. https://doi.org/10.1038/s41467-019-10656-5.

Norris, C.E., Bean, G. Mac, Cappellazzi, S.B., Cope, M., Greub, K.L.H., Liptzin, D., Rieke, E.L., Tracy, P.W., Morgan, C.L.S., Honeycutt, C.W., 2020. Introducing the North American project to evaluate soil health measurements. Agronomy Journal 112, 3195–3215. https://doi.org/10.1002/agj2.20234.

Nunes, M.R., van Es, H.M., Schindelbeck, R., Ristow, A.J., Ryan, M., 2018. No-till and cropping system diversification improve soil health and crop yield. Geoderma 328, 30–43. https://doi.org/10.1016/j.geoderma.2018.04.031.

Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G., Solymos, M., Stevens, H.H., Szoecs, E., Wagener, H., 2015. Vegan: community ecology package. R Package.

Pasolli, E., Truong, D.T., Malik, F., Waldron, L., Segata, N., 2016. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Computational Biology 12, 1–26. https://doi.org/10.1371/journal.pcbi.1004977.

Peng, X., Bruns, M.A., 2018. Development of a nitrogen-fixing cyanobacterial consortium for surface stabilization of agricultural soils. Journal of Applied Phycology 31. https://doi.org/10.1007/s10811-018-1597-9.

Pershina, E., Valkonen, J., Kurki, P., Ivanova, E., Chirak, E., 2015. Comparative analysis of prokaryotic communities associated with organic and conventional farming systems. PLoS One 10, e0145072. https://doi.org/10.1371/journal.pone.0145072.

Poore, G.D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolek, T., Janssen, S., Metcalf, J., Song, S.J., Kanbar, J., Miller-Montgomery, S., Heaton, R., Mckay, R., Patel, S.P., Swafford, A.D., Knight, R., 2020. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature 579, 567–574. https://doi.org/10.1038/s41586-020-2095-1.

Qin, N., Liang, P., Wu, C., Wang, G., Xu, Q., Xiong, X., Wang, T., Zolfo, M., Segata, N., Qin, H., Knight, R., Gilbert, J.A., Zhu, T.F., 2020. Longitudinal survey of microbiome associated with particulate matter in a megacity. Genome Biology 21, 1–11. https://doi.org/10.1186/s13059-020-01964-x.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41, 590–596. https://doi.org/10.1093/nar/gks1219.

Rekik, F., Van, Es H., Hernandez-aguilera, J.N., Gómez, M.I., 2018. Regional soil health assessment for coffee farms on andosols in Colombia. Geoderma 14, e00176. https://doi.org/10.1016/j.geodrs.2018.e00176.

Rinot, O., Levy, G.J., Steinberger, Y., Svoray, T., Eshel, G., 2019. Soil health assessment: a critical review of current methodologies and a proposed new approach. The Science of the Total Environment 648, 1484–1491. https://doi.org/10.1016/j.scitotenv.2018.08.259.

Rocca, J.D., Simonin, M., Blaszczak, J.R., Ernakovich, J.G., Gibbons, S.M., Midani, F.S., Washburne, A.D., 2019. The Microbiome Stress Project: toward a global meta-analysis of environmental stressors and their effects on microbial communities. Frontiers in Microbiology 9, 3272. https://doi.org/10.3389/fmicb.2018.03272.

Rodríguez-Caballero, E., Cantón, Y., Chamizo, S., Afana, A., Solé-Benet, A., 2012. Effects of biological soil crusts on surface roughness and implications for runoff and erosion. Geomorphology 145–146, 81–89. https://doi.org/10.1016/j.geomorph.2011.12.042.

Rousk, J., Bååth, E., Brookes, P.C., Lauber, C.L., Lozupone, C., Caporaso, J.G., Knight, R., Fierer, N., 2010. Soil bacterial and fungal communities across a pH gradient in an arable soil. The ISME Journal 4, 1340–1351. https://doi.org/10.1038/ismej.2010.58.

Schindelbeck, R.R., Moebius-Clune, B.N., Moebius-Clune, D.J., Kurtz, K.S.M., van Es, H.M., 2016. Cornell University Comprehensive Assessment of Soil Health Laboratory Standard Operating Procedures.

Schmid, M.W., van Moorsel, S.J., Hahl, T., De Luca, E., De Deyn, G.B., Wagg, C., Niklaus, P.A., Schmid, B., 2021. Effects of plant community history, soil legacy and plant diversity on soil microbial communities. Journal of Ecology 109, 3007–3023. https://doi.org/10.1111/1365-2745.13714.

Statnikov, A., Alekseyenko, A.V., Li, Z., Henaff, M., Perez-Perez, G.I., Blaser, M.J., Aliferis, C.F., 2013. Microbiomic signatures of psoriasis: feasibility and methodology comparison. Scientific Reports 3, 1–7. https://doi.org/10.1038/srep02620.

Stewart, R.D., Jian, J., Gyawali, A.J., Thomason, W.E., Badgley, B.D., Reiter, M.S., Strickland, M.S., 2018. What we talk about when we talk about soil health. Agricultural & Environmental Letters 3, 5–9. https://doi.org/10.2134/ael2018.06.0033.

Sze, M.A., Schloss, P.D., 2018. Leveraging existing 16S rRNA gene surveys to identify reproducible biomarkers in individuals with colorectal tumors. mBio 9, 1–16. https://doi.org/10.1128/mBio.00630-18.

Sze, M.A., Topçuoğlu, B.D., Lesniak, N.A., Ruffin, M.T., Schloss, P.D., 2019. Fecal short-chain fatty acids are not predictive of colonic tumor status and cannot be predicted based on bacterial community structure. mBio 10, 1–7. https://doi.org/10.1128/mBio.01454-19.

Tatangelo, V., Franzetti, A., Gandolfi, I., Bestetti, G., Ambrosini, R., 2014. Effect of preservation method on the assessment of bacterial community structure in soil and water samples. FEMS Microbiology Letters 356, 32–38. https://doi.org/10.1111/1574-6968.12475.

Thompson, J., Johansen, R., Dunbar, J., Munsky, B., 2019. Machine learning to predict microbial community functions: an analysis of dissolved organic carbon from litter decomposition. PLoS One 14, 1–16. https://doi.org/10.1371/journal.pone.0215502.

Topçuoğlu, B.D., Lesniak, N.A., Ruffin, M.T., Wiens, J., Patrick, D., Arbor, A., Science, C., State, P., Medical, H., Topçuoğlu, B.D., Lesniak, N.A., Ruffin, M.T., Wiens, J., Schloss, P.D., Iv, T.R., Wiens, J., 2020. A framework for effective application of machine learning to microbiome-based classification problems. mBio 11, 1–13. https://doi.org/10.1101/816090.

Tripathi, B.M., Stegen, J.C., Kim, M., Dong, K., Adams, J.M., Kyung, Y., Adams, J.M., 2018. Soil pH mediates the balance between stochastic and deterministic assembly of bacteria. The ISME Journal 1072–1083. https://doi.org/10.1038/s41396-018-0082-4.

Trognitz, F., Hackl, E., Widhalm, S., Sessitsch, A., 2016. The role of plant-microbiome interactions in weed establishment and control. FEMS Microbiology Ecology 92, 1–15. https://doi.org/10.1093/femsec/fiw138.

Tzeneva, V.A., Salles, J.F., Naumova, N., de Vos, W.M., Kuikman, P.J., Dolfing, J., Smidt, H., 2009. Effect of soil sample preservation, compared to the effect of other environmental variables, on bacterial and eukaryotic diversity. Research in Microbiology 160, 89–98. https://doi.org/10.1016/j.resmic.2008.12.001.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605.

van Es, H.M., Karlen, D.L., 2019. Reanalysis validates soil health indicator sensitivity and correlation with long-term crop yields. Soil Science Society of America Journal 83, 721–732. https://doi.org/10.2136/sssaj2018.09.0338.

Weil, R.R., Islam, K.R., Stine, M.A., Gruver, J.B., Samson-Liebig, S.E., 2003. Estimating active carbon for soil quality assessment: a simplified method for laboratory and field use. American Journal of Alternative Agriculture 18, 3–17. https://doi.org/10.1079/AJAA2003003.

Werner, J.J., Knights, D., Garcia, M.L., Scalfone, N.B., Smith, S., Yarasheski, K., Cummings, T.A., Beers, A.R., Knight, R., Angenent, L.T., 2011. Bacterial community structures are unique and resilient in full-scale bioenergy systems. Proceedings of the National Academy of Sciences 108, 4158–4163. https://doi.org/10.1073/pnas.1015676108.

Wessén, E., Söderström, M., Stenberg, M., Bru, D., Hellman, M., Welsh, A., Thomsen, F., Klemedtson, L., Philippot, L., Hallin, S., 2011. Spatial distribution of ammonia-oxidizing bacteria and archaea across a 44-hectare farm related to ecosystem functioning. The ISME Journal 5, 1213–1225. https://doi.org/10.1038/ismej.2010.206.

Wilhelm, R.C., Cardenas, E., Maas, K.R., Leung, H., McNeil, L., Berch, S., Chapman, W., Hope, G., Kranabetter, J.M., Dubé, S., Busse, M., Fleming, R., Hazlett, P., Webster, K.L., Morris, D., Scott, D.A., Mohn, W.W., 2017. Biogeography and organic matter removal shape long-term effects of timber harvesting on forest soil microbial communities. The ISME Journal 11, 2552–2568. https://doi.org/10.1038/ismej.2017.109.

Wilhelm, R.C., Pepe-Ranney, C., Weisenhorn, P., Lipton, M., Buckley, D.H., 2021. Competitive exclusion and metabolic dependency among microorganisms structure the cellulose economy of an agricultural soil. mBio 12, 1–19. https://doi.org/10.1128/mBio.03099-20.

Williams, H., Colombi, T., Keller, T., 2020. The influence of soil management on soil health: an on-farm study in southern Sweden. Geoderma 360. https://doi.org/10.1016/j.geoderma.2019.114010.

Yergeau, E., Quiza, L., Tremblay, J., 2020. Microbial indicators are better predictors of wheat yield and quality than N fertilization. FEMS Microbiology Ecology 96, 1–13. https://doi.org/10.1093/femsec/fiz205.

Zhalnina, K., De Quadros, P.D., Gano, K.A., Davis-Richardson, A., Fagen, J.R., Brown, C.T., Giongo, A., Drew, J.C., Sayavedra-Soto, L.A., Arp, D.J., Camargo, F.A.O., Daroub, S.H., Clark, I.M., McGrath, S.P., Hirsch, P.R., Triplett, E.W., 2013. Ca. *Nitrososphaera* and *Bradyrhizobium* are inversely correlated and related to agricultural practices in long-term field experiments. Frontiers in Microbiology 4, 1–13. https://doi.org/10.3389/fmicb.2013.00104.

Zhang, Z., Han, X., Yan, J., Zou, W., Wang, E., Lu, X., Chen, X., 2020. Keystone microbiomes revealed by 14 years of field restoration of the degraded agricultural soil under distinct vegetation scenarios. Frontiers in Microbiology 11, 1–13. https://doi.org/10.3389/fmicb.2020.01915.

Zheng, W., Zeng, S., Bais, H., Lamanna, J.M., 2018. Plant growth-promoting rhizobacteria (PGPR) reduce evaporation and increase soil water retention. Water Resources Research 54, 3673–3687. https://doi.org/10.1029/2018WR022656.

Zhou, Y.H., Gallins, P., 2019. A review and tutorial of machine learning methods for microbiome host trait prediction. Frontiers in Genetics 10, 1–14. https://doi.org/10.3389/fgene.2019.00579.

Zwetsloot, M.J., Muñoz Ucros, J., Wickings, K., Wilhelm, R.C., Sparks, J.P., Buckley, D.H., Bauerle, T.L., 2020. Prevalent root-derived phenolics drive shifts in microbial community composition and prime decomposition in forest soil. Soil Biology and Biochemistry 145, 530–541. https://doi.org/10.1016/j.soilbio.2020.107797.