THESIS

MACHINE LEARNING METHODS TO DISCOVER PATTERNS IN MICROBIALLY DRIVEN SOIL CARBON SEQUESTRATION

Submitted by

Jaron Thompson

Department of Chemical and Biological Engineering

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2020

Master's Committee:

Advisor: Brian Munsky

Jessica Metcalf
Joshua Chan

ABSTRACT

MACHINE LEARNING METHODS TO DISCOVER PATTERNS IN MICROBIALLY DRIVEN SOIL CARBON SEQUESTRATION

Understanding how microbiomes function is a major area of research, as microorganisms play a significant role in environments spanning nearly every corner of the earth. Recent advances in DNA sequencing technology have made it possible to profile microbial communities, yet noise and sparsity in microbiome data makes it difficult to identify consistent patterns in microbial community behavior. In this thesis, we apply a host of machine learning methods to elucidate the role of the soil microbiome in mediating soil carbon sequestration. We demonstrate that broad characteristics of the soil microbiome such as richness and biomass can be used to forecast abundance of dissolved organic carbon (DOC) in soil. We also show that feature selection analysis using a host of machine learning and standard statistical techniques identifies a consensus set of significant taxa that predict DOC abundance. Finally, we demonstrate how these feature selection techniques can be used to explore more advanced probabilistic models that assign accurate estimates of prediction confidence. The methods proposed in this thesis could be used to design optimized microbial communities that combat climate change by promoting increased levels of carbon storage in soil.

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Dr. Brian Munsky, for his continuous support from the onset of this project. I extend a huge thanks to Dr. John Dunbar and the members of the Terrestrial Microbial Carbon Cycling Science Focus Area at Los Alamos National Laboratory for providing additional guidance and for designing and performing the experiments that make this research possible. Dr. John Dunbar also co-advised my summer employment as a graduate student intern at Los Alamos National Laboratory. I would like to thank my committee members Dr. Jessica Metcalf and Dr. Joshua Chan, the Walter Scott Jr. College of Engineering, and the Chemical and Biological Engineering Department at Colorado State University.

DEDICATION

I would like to dedicate this thesis to my parents.

TABLE OF CONTENTS

Chapter 1

Introduction

Current levels of atmospheric carbon significantly exceed the greatest recorded levels dating back 800,000 years [1,2]. The alarming increase of greenhouse gases and global temperatures motivates the need for strategies to combat climate change. A potential strategy to mitigate increasing levels of atmospheric carbon dioxide (CO_2) is carbon sequestration, which is the storage of CO_2 in a stable carbon sink. A particularly promising carbon sink is soil, which contains more carbon than vegetation and atmospheric carbon combined. Currently, the amount of carbon dioxide that is emitted from soil is more than ten times greater than that from fossil fuel emissions, suggesting that a small change towards carbon storage in soil could significantly reduce atmospheric carbon levels [3,4].

Terrestrial carbon sequestration occurs when decaying organic matter (e.g. plant litter) is converted into soluble forms of carbon that remain stable in deeper soil layers. Shifting the conversion process of decaying organic matter to produce soluble carbon reduces the flux of CO_2 from soil and increases levels of dissolved organic carbon (DOC) stabilized by the soil matrix. While it is known that the soil microbiome plays a significant role in plant litter decomposition [5], discovering the mechanisms by which bacteria and fungi influence the accumulation of DOC remains a major area of research [3]. By understanding the mechanisms that underpin microbially mediated carbon sequestration, microbiomes could be optimized to shift soil carbon cycling towards increased carbon storage.

The emergence of next generation sequencing technology has made it possible to taxonomically profile microbial communities, which provides a means of linking microbial abundance with macroscopic environmental traits, such as levels of DOC. Using sequencing data, machine learning and statistical modeling offers a powerful tool to discover patterns that reveal how the microbiome influences or is influenced by its environment. However, constructing models using sequencing data is difficult owing to high dimensionality, noise, and sparsity. Furthermore, machine learning

models are often limited in that they provide little insight into *how* the model makes decisions, and predictions are rarely assigned a metric of prediction uncertainty.

This thesis explores a host of machine learning approaches to examine how the soil microbiome drives carbon storage in terrestrial ecosystems. We showed that initial community level characteristics of the soil microbiome, such as richness and total biomass, are predictive of downstream levels of DOC. To model the relationship of specific taxa with DOC, we applied random forest regression, neural network regression, and indicator species analysis for feature selection and prediction tasks, using a framework we call RFINN (Random Forest Indicator species analysis Neural Network). Using RFINN to identify a reduced set of taxa, we showed that probabilistic models are a powerful tool for constructing easily interpreted models that reliably estimate prediction uncertainty. Furthermore, the developed models accurately predicted levels of DOC across different types of decomposing plant litter.

Machine learning analyses were applied to predict DOC outcomes in soil microcosms using profiled microbiomes (16S rRNA gene profiles) as features. Experimental data was acquired after inoculating sterilized litter samples with a spectrum of decomposer microbial communities sampled from nine states in the Western United States. Carbon flow (DOC and cumulative carbon dioxide) was measured after a six week decomposition period, and microbial communities were profiled at the beginning and end of the experiment. To understand the relationship between community level traits of the soil microbiome and carbon sequestration, features such as richness, diversity, and biomass were used to predict levels of DOC in soil microcosms using a logistic regression model. The ability of the model to predict emergent levels of DOC from initial microbial community traits corroborates the paradigm that the soil microbiome directly influences carbon flow in soil.

In contrast to models that use community level microbial traits, we examined how abundance of OTUs (operational taxonomic units) can be used as features in machine learning models to predict abundance of DOC. With feature importance ranking using random forest regression, neural network regression, and indicator species analysis, we identified a consensus set of OTUs that were

highly ranked among all three approaches. This approach, called RFINN, significantly reduced the total number of OTUs identified using 16s rRNA gene sequencing. Using this reduced feature set, model prediction performance did not significantly change using random forest regression, and was even improved using the neural network regression model. Furthermore, feature selection is an important step towards understanding which microbial taxa are most significantly linked with carbon flow in soil.

The majority of machine learning approaches that have been applied to analyze the microbiome typically do not assign a metric of uncertainty to predictions and typically provide little insight into the mechanisms that dictate such predictions [6, 7]. Using a reduced set of bacterial genera determined by RFINN, we applied Bayesian networks to make probabilistic predictions of DOC abundance. Bayesian networks are graphical structures composed of nodes and edges that describe the joint probability distribution of model features and targets [8]. In addition to standard cross-validation with held-out testing data, we investigated the transferability of the model using samples from independent litter decomposition experiments. We found that a Bayesian network model trained using samples from a pine litter decomposition study accurately predicted abundance of DOC using samples from an oak litter decomposition experiment, but failed to accurately estimate DOC from grass litter decomposition samples. Despite insignificant prediction accuracy when applied to grass litter samples, model predictions were markedly less confident, with over half of the samples discarded due to uncertainty. Taken together, uncertainty analysis and cross-validation showed that the Bayesian network model correctly assigned greater prediction uncertainty to samples that proved more difficult to accurately predict.

Machine learning approaches applied to microbiome data from litter decomposition studies reveal that levels of DOC can accurately be estimated from microbial community traits. With the ability to identify a small subset of taxa whose abundances are linked with DOC, machine learning approaches provide valuable insight towards understanding how microbial communities might influence carbon storage in soil. Cross-validation of machine learning models using data from independent litter decomposition studies suggests that the relationship between soil microbiomes

and DOC may be conserved across different litter types. With insight from feature selection and uncertainty analysis, the machine learning models explored in this thesis provide valuable insight for potentially engineering microbial communities to combat climate change by optimizing carbon sequestration in soil.

Chapter 2

Soil bacterial and fungal richness forecast patterns of early pine litter decomposition [1]

2.1 Overview

Discovering widespread microbial processes that drive unexpected variation in carbon cycling may improve modeling and management of soil carbon [9–11]. A first step is to identify community features linked to carbon cycle variation. We addressed this challenge using an epidemiological approach with 206 soil communities decomposing Ponderosa pine litter in 618 microcosms. Carbon flow from litter decomposition was measured over a six-week incubation. Cumulative CO_2 efflux varied 2-fold among microcosms and dissolved organic carbon (DOC) varied 5-fold, demonstrating large functional variation despite constant environmental conditions where strong selection is expected. "High" and "Low" DOC functional states were delineated and the communities in each cohort were taxonomically profiled. A logistic model including total biomass, fungal richness, and bacterial richness measured in the original soils or in the final microcosm communities predicted the functional states with 72 ($P<0.05$) and 80 ($P<0.001$) percent accuracy, respectively. The strongest predictors of the DOC functional state were biomass and either fungal richness (in the original soils) or bacterial richness (in the final microcosm communities). Successful forecasting of functional patterns after lengthy community succession in a new environment reveals

---

[1] Michaeline Albright[a], Renee Johansen[a], Jaron Thompson[b,*], Deanna Lopez[a], La Verne Gallegos-Graves[a], Marie E Kroeger[a], Andreas Runde[a], Rebecca Mueller[c], Alex Washburne[d], Brian Munsky[b,e], Thomas Yoshida[d], John Dunbar[a]

a Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, United States
b Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO 80523, USA
c Center for Biofilm Engineering, Montana State University, Bozeman, MT 59717, United States
d Department of Microbiology and Immunology, Montana State University, Bozeman, MT 59717, United States
e Keck Scholar, School of Biomedical Engineering, Colorado State University, Fort Collins, CO 80523, USA

***** I designed and implemented the logistic regression model used to predict high or low levels of dissolved organic carbon from microbial community traits such as biomass and richness. I also performed the statistical analysis used to examine the significance of model accuracy.

strong historical contingencies. Forecasting future community function is a key advance beyond correlation of functional variance with end-state community features. The importance of taxon richness - the same feature linked to carbon fate in gut microbiome studies - underscores the need for increased understanding of biotic mechanisms that shape richness in microbial communities, independent of physicochemical conditions.

2.2 Introduction

Modeling existing soil carbon stocks is a starting point to predict future feedbacks to climate [12]. Accurate modeling of current carbon stocks remains a challenge as indicated by large unexplained variance, weak spatial correlation at the global scale, and deviation of entire habitat types [13–15]. Many factors may contribute to these discrepancies, but an emerging view posits a strong role for microbial composition [13, 16–18] because microbial communities are not always functionally equivalent [19]. Different microbial community "types" can occur within a habitat type, contributing substantial variation to ecosystem function [20–22]. A community type is defined as a discernable compositional cluster in a multi-dimensional landscape of compositional possibilities [22]. The existence of alternative soil community types that vary in function under the same environmental conditions has been postulated [18], including communities with functional extremes analogous to stable dysbiosis in the human gut [23]. Such communities in nature would create variation in carbon cycling that is unexplained in conventional models.

The specific features of microbial community composition that may drive substantial variation in soil carbon cycling are unknown [9]. Features that have been explored theoretically include ratios of fungi versus bacteria [24], active versus dormant populations [13], and oligotrophs versus copiotrophs [10]. However, experimental validation lags [25]. Microbial diversity has been proposed as a driver but continues to be intensely debated [25, 26] with conflicting experimental evidence against [27–30] and for a link [31–34]. In recent studies supporting a diversity-decomposition relationship, a single community was manipulated in each case by extreme dilution (e.g. undiluted versus 10-5 fold [32, 33] or by size-fractionation of a soil [34] to create diversity

gradients, but these gradients are difficult to imagine under natural scenarios. Examining diverse microbial communities in nature that foster different carbon cycling patterns under the same environmental conditions is a useful alternative to discover community features relevant to carbon cycling.

Toward this end, we used an epidemiological approach wherein a large population of plant litter decomposer communities in laboratory microcosms was screened to delineate cohorts with contrasting functional states. Although surface leaf litter decomposition is only one component of soil carbon cycling, it accounts for about half of the CO_2 efflux in temperate deciduous forests annually [35]. Plant litter decomposition is generally viewed as a two-stage process comprising an initial fast phase dominated by weedy microbial taxa, and a subsequent slow phase driven by taxa better equipped to deconstruct lignocellulose [36, 37]. The early phase of litter decomposition is of particular interest because carbon flow during rapid microbial growth on labile plant carbon is now understood to play an important role in the formation of soil organic matter [36, 38].

To acquire a spectrum of decomposer communities on Ponderosa pine leaf litter, 206 soil samples were collected from nine states in the Western U.S. (Figure 2.1) as source material for the dispersal of microbial communities onto leaf litter in 618 microcosms. We measured carbon flow during the early phase of plant litter decomposition by quantifying dissolved organic carbon (DOC) and cumulative carbon dioxide (CO_2) from a six-week decomposition period. Functional states were delineated as "high" versus "low" DOC. We hypothesized that the composition of the original soils would exert legacy effects that constrain succession in each microcosm, and therefore community features in the original soils would be linked to functional outcomes in the microcosms.

2.3 Materials and methods

2.3.1 Initial soil collection for microbial inoculum

Soil samples were collected from 206 locations throughout the southwestern United States between February and April, 2015 (Figure 2.1). The goal of this study was not to relate functional outcomes to detailed characteristics of the environments from which the soils were col-

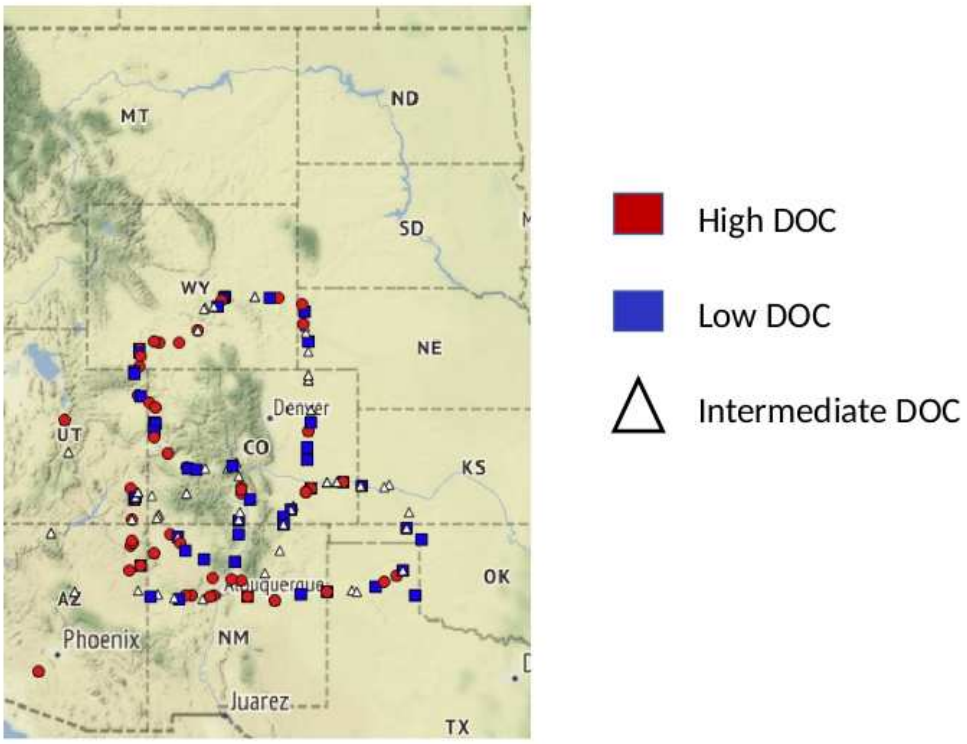

Figure 2.1: Map of sampling points in the western United States. Points are colored by the dissolved organic carbon (DOC) output generated by the microbial community extracted from the soil sample (mean of three replicates), following 44 days of pine litter decomposition.

lected. Therefore, a randomized collection scheme was not used, as this would have substantially increased the cost and logistical burden of sample collection without benefit. Samples were typically collected at locations approximately 80 km apart, at least 15 meters from roadways, from the top 3cm of the soil surface after removal of surface litter. In the collection region, ecosystems routinely have patchy ground cover with exposed soil and little, if any, litter layer at the soil surface. Samples were collected in sterile 50-ml screw-cap tubes, and immediately stored on ice. The location of each sample was recorded by GPS (coordinates available upon request) and photographed to facilitate description of the major ecosystem types from which samples were obtained.

2.3.2 Microcosm construction and CO_2 sampling

Microcosms consisted of 125ml serum bottles containing approximately 5g of sand and 0.12g of Ponderosa pine leaf litter, which had been ground in a Wiley Mill (Thomas Scientific, Swedesboro, NJ, USA). The microcosms were sterilized by autoclaving three times for 1 hour each, with at least an 8-hour resting interval between each autoclave cycle. Microbial communities were extracted from soil samples (n=206) by suspending one gram of soil in 9ml of phosphate-buffered saline (PBS), then creating a 1000-fold dilution in PBS amended with NH4NO3 at 4.8mg/ml. We used a high nitrogen background to represent the atmospheric deposition of nitrogen that has already occurred and will continue to increase in natural ecosystems [39]. Three microcosms per soil sample each received 1.3 mls of inoculum, pipetted directly onto a 0.02g aliquot of pine litter (n=618 microcosms). These microcosms were then sealed with Teflon-lined crimp caps and incubated at 25 °C in the dark for 14 days to equilibrate the communities. Negative control microcosms, used to confirm the efficacy of sterilization, received the same quantities of PBS and NH4NO3, but no microbial communities. The headspace in each microcosm was evacuated using a vacuum pump on days 3 and 7, and replaced with sterile-filtered air. On day 14, an additional 0.1g aliquot of litter sterilized by three rounds of autoclaving was added to each microcosm, and microcosms were re-sealed. The microcosms were incubated at 25 °C in the dark for a further 30 days. On days 2, 5, 9, 16, 23 and 30, CO_2 was measured by gas chromatography using an Agilent

Technologies 490 Micro GC (Santa Clara, CA, USA). After each measurement, the headspace air was evacuated with a vacuum pump and replaced with sterile-filtered air.

2.3.3 Dissolved organic carbon (DOC) and litter community sampling

After the 44-day (total) incubation, microcosms were destructively sampled to measure DOC and community composition. DOC extractions were performed using a rapid, gentle procedure to avoid measurement artifacts arising from microbial growth or microbial cell disruption. For each microcosm, 5ml of sterile deionized water was added, swirled manually for 30 seconds, then transferred to two 2-ml tubes. The tubes were centrifuged 4 minutes at 16,400xg. The supernatants were combined and sterilized by filtration through a 0.2 μm filter. The concentration of DOC in each sample was measured on an OI Analytical model 1010 wet oxidation TOC analyzer (Xylem Inc., Rye Brook, NJ, USA), calibrated daily. Following DOC sampling, material (sand and litter) from each microcosm was frozen at -80 °C for DNA extraction.

2.3.4 Bacterial and fungal community taxonomic profiling

Samples for community profiling were down-selected based on the mean DOC quantity of each set of three replicate microcosms at day 44. The profiled samples represented the two tails of the DOC frequency distribution. Ribosomal RNA gene profiles were obtained for original soil samples (n=128) and their corresponding replicate microcosms (n=384). DNA extractions were performed with a DNeasy PowerSoil 96-well plate DNA extraction kit (Qiagen, Hilden, Germany). The standard protocol was used with the following two exceptions: 1) 0.3 grams of material was used per extraction; 2) bead beating was conducted using a Spex Certiprep 2000 Geno/Grinder (Spex SamplePrep, Metuchen, NJ, USA) for three minutes at 1900 strokes/minute. DNA samples were quantified with an Invitrogen Quant-iTTM dsDNA Assay Kit (Thermo Fisher Scientific, Eugene, OR, USA) on a BioTek Synergy HI Hybrid Reader (Winooski, VT, USA). PCR templates were prepared by diluting an aliquot of each DNA stock in sterile water to 1ng/μl. The bacterial (and archaeal) 16S rRNA gene (V3-V4 region) was amplified using primers 515f-R806 [40]. The fungal 28S rRNA gene (D2 hypervariable region) was amplified using the LR22R primer [41] and the

reverse LR3 primer [42]; this target sequence is amenable to phylogenetic tree construction and provides genus-level resolution equivalent to that provided by internal transcribed spacer sequences [43].

A two-step amplification procedure was used based on Mueller et al. [41], with Phusion Hot Start II High Fidelity DNA polymerase (Thermo Fisher Scientific, Vilnius, Lithuania). In the first PCR, barcoded amplicons were produced over 22 cycles using gene primers flanked by 6nt barcodes that jointly provided a unique 12-mer barcode for each sample [44]. Cycling conditions were 30 s at 98 °C, 22 cycles of (98 °C for 15 s, 60 °C for 30 s, 72 °C for 30 s), and a final extension step of 72 °C for 5min. The second PCR extended Illumina adapter sequences on the amplicons over 10 cycles. Cycling conditions were 30 s at 98 °C, 10 cycles of (98 °C for 15 s, 65 °C for 30 s, 72 °C for 30 s), and a final extension step of 72 °C for 5min. Amplicons were cleaned using a MoBio UltraClean PCR clean-up kit (Carlsbad, CA, USA), quantified using the same procedure as for the extracted DNA, and then pooled at a concentration of 10ng each. The pooled samples were further cleaned and concentrated using the Mobio UltraClean PCR clean-up kit. All clean ups were undertaken as per the manufacturer's instructions with the following modifications: binding buffer was reduced from 5X to 3X sample volume and DNA was eluted in 50 μl Elution Buffer. DNA quality of the amplicon pool was assessed with a bioanalyzer, concentration was verified by qPCR, and sequencing was performed on an Illumina MiSeq with paired-end 250 bp chemistry at Los Alamos National Laboratory.

Bacterial and fungal sequences were merged with PEAR v 9.6 [45], quality filtered to remove sequences with 1% or more low-quality (q20) bases, and demultiplexed using QIIME [46] allowing no mismatches to the barcode or primer sequence. Further processing was undertaken with UPARSE [47]. Sequences with an error rate greater than 0.5 were removed, remaining sequences were dereplicated, OTU clustering was performed at 97%, and putative chimeras were identified de novo using UCHIME [48]. Bacterial and fungal OTUs were classified via the Ribosomal Database Project (RDP) classifier [49]. OTUs that were not classified as bacteria or fungi with 100% confidence were removed from the dataset. Bacterial OTUs with less than 80% classification confidence

at the phylum level were also removed. The omitted data accounted for less than 5% of the total. Of the 128 source soil samples that yielded high or low DOC outcomes in microcosms, 123 of the samples passed sequence quality control and 1,481,601 and 1,741,698 total sequences were obtained for bacteria and fungi respectively. The sequences represented 5595 bacterial OTUs (an average of 409 detected per soil) and 2270 fungal OTUs (an average of 112 detected per soil). From the day-44 microcosms samples representing the high and low DOC groups, a total of 9,576,525 sequences from 349 of 384 microcosms that passed quality control were obtained for bacteria and 13,124,107 sequences from 377 microcosms were obtained for fungi. These represented 2,527 bacterial OTUs (an average of 275 detected per microcosm) and 753 fungal OTUs (an average of 47 detected per microcosm).

Sequence data were deposited in the NCBI Sequence Read Archive (PRJNA515766 for the source soils and PRJNA478595 for the day-44 microcosm samples). All other data including OTU tables are available upon request.

2.3.5 Total biomass, fungal, and bacterial abundance

The DNA quantity extracted from each sample was used as a proxy for biomass. Fungal and bacterial abundance were separately estimated by quantitative PCR (qPCR) using 18S rRNA gene primers nu-SSU- 1196F and nu-SSU-1536R for fungi [50] and 16S rRNA gene primers EUB 338 [51] and EUB 518 [52] for bacteria as described by Castro et al. [53]. Assays were performed with the Biorad iQ SyBr Green Supermix on a BioRad CFX Connect Real-Time System (BioRad, Hercules, CA). DNA templates were normalized to 1.0 ng/μl. Six-point calibration standards were created by serial dilution of linearized plasmid DNA containing a cloned Phoma 18S rRNA gene fragment (for fungi) or genomic DNA from Burkholderia thailandensis E264, ATCC 70038 (for bacteria). Melt curves were generated for every run to detect potential false positives.

2.3.6 DOC binding assay

The fraction of DOC able to bind to mineral surfaces was measured for one DOC sample replicate from each of the high DOC (n = 64) and each of the low DOC (n = 64) day-44 com-

munities. Aluminum oxide was used as a representative mineral for DOC binding [54]. For each sample, 0.5 ml of DOC was added to 1 ml of sterile water (3X dilution factor) and 0.3 grams of aluminum oxide (Al2O3). Samples were mixed by inversion with a Thermolyne rocker (Barnstead/Thermolyne, Dubuque, IA, USA) for 30 minutes and then centrifuged at 16,100 x g for 5 minutes. Supernatant was transferred to a new tube and stored at -20 °C until DOC quantification on a TOC analyzer. The percentage of bound DOC was calculated as 100% x (DOCpost-binding x dilutionFactor)/DOCpre-binding.

2.3.7 Statistical analyses

Community composition analyses were performed with rarefied data unless otherwise stated. For original soil samples, bacterial communities were rarefied to 1095 sequences per sample, and fungal communities were rarefied to 1385 sequences per sample. For day-44 microcosms, bacterial communities were rarefied to 1023 sequences and fungal communities were rarefied to 2032 sequences. Bacterial and fungal richness and diversity (Shannon-Wiener index) were compared across high and low DOC groups for both original soil and day-44 microcosm samples using one-way ANOVAs. Bray-Curtis dissimilarity matrices for bacterial and fungal communities were computed using log-transformed data for bacteria and for fungi in the R library vegan v 2.4-3 [55]. A permutational multivariate analysis of variance (PERMANOVA; [50]) was performed to assess whether the community composition of high and low DOC groups differed (vegan v 2.4-3; [56]). The individual microcosms (day 44) were treated as independent samples in all statistical analyses because the replicates diverged substantially in community composition by the conclusion of the experiment and were therefore considered biologically distinct. Compositional analyses were also run on each set of replicates (set A, set B, set C) independently to confirm that conclusions were consistent irrespective of how replicates were treated. To further compare community composition between high and low DOC groups, OTU sequences were grouped phylogenetically at the Family level for bacteria and Order level for fungi to assess differential abundance of individual taxa. This analysis was performed for the original soils and day-44 microcosm samples. Family-level

comparisons were not made for fungi due to low classification confidence levels. For fungal orders and bacterial families, OTUs were only used that could be phylogenetically assigned with at least a 70% confidence level from the RDP Classifier. Due to the sparse data for rare taxa, further statistical analysis of individual taxa was restricted to the most abundant bacterial families and fungal orders that comprised on average at least 1% of the sequences of either the high or low DOC groups. Differences in taxon abundance in high versus low DOC groups were compared by t-tests. Correlations between various community features versus DOC quantity were measured with Pearson's (univariate) or Mantel (multivariate) tests. For Mantel tests with the original soils, the average day-44 DOC values among each set of three replicate microcosms were used to generate a Euclidean distance matrix for comparison with bacterial and fungal community Bray-Curtis matrices (ecodist package; [57]). For day-44 microcosm community samples, DOC values from all microcosms were used to create the distance matrix. Univariate features included fungal abundance (qPCR), bacterial abundance (qPCR), fungal:bacterial ratios, total biomass (measured as total extracted DNA), OTU richness, and Shannon diversity. All statistical analyses were performed using R v3.3.3 [58].

2.3.8 Prediction models for DOC abundance

To predict DOC abundance, a logistic regression model was developed using seven day-0 (original soil) community features as variables: total biomass, fungal richness, bacterial richness, fungal diversity, bacterial diversity, fungal abundance, and bacterial abundance. The total data set was partitioned into 1000 unique permutations of training and testing data with 30% of samples reserved for testing. Training data and testing data were partitioned such that the balance of high DOC and low DOC labels in each set was equivalent. Variables in the training data were standardized to be zero mean with unit variance, and variables in the testing data were similarly scaled using training data statistics. SCIKIT-LEARN's [59] Logistic Regression model was used to fit to the training data using an 'L2' penalty to reduce unnecessary features. After fitting the model to training data, feature selection was applied using SCIKIT-LEARN's SelectFromModel function to reject features

with regression coefficients less than a threshold value of 1e-5. Feature importance was assessed with the Wald statistic, defined as the logistic regression coefficient divided by the standard error of the coefficient [60]. Based on the Wald statistics (Table 2.3), three features (biomass, fungal richness, and bacterial richness) were down-selected as the most important variables to predict DOC abundance. A second logistic regression model with the reduced set of features measured at day-44 was applied to 1000 permutations of training and testing data as described above (Table 2.4).

A logistic regression model with the reduced set of features (measured in the original soils or in the day-44 microcosms) was applied to training and testing data as described above. To assess model significance, the average McFadden pseudo R2 value and the log likelihood ratio P value were calculated. Prediction accuracy is defined as the number of true positives and true negatives divided by the total number of samples in the testing dataset. The distribution of prediction accuracy over 1000 permutations of training and testing data was compared to a null model that always assigned the most prevalent label in the testing set. We used the z-test for the equality of two proportions to evaluate the significance of comparing the proportion of correctly labeled samples using the logistic regression model compared to that with the null model, and we report the median P-value over the 1000 permutations of the training and testing data. Code for data pre-processing, logistic regression and statistical analysis is available online at https://github.com/MunskyGroup/Albright_et_al_2019.

2.4 Results

2.4.1 Microbial-driven variation in respiration, DOC quantity, and DOC quality was large.

Over the 6-week decomposition period, respired CO_2 varied approximately 2-fold between 160 and 345 mg/g of litter, and DOC varied 5-fold, with between 3 and 18 mg/g of litter (Figure 2.2B). The CO_2 and DOC outputs from decomposition were negatively correlated ($R2 = 0.16$, $P = <0.001$). The DOC frequency distribution was used to delineate two contrasting functional states: high versus low DOC. These functional cohorts were balanced by requiring each group to contain 192

samples (i.e., all 3 replicate communities derived from 64 source soils). The high and low DOC groups varied not only in DOC abundance but also in DOC composition, as indicated by a mineral-binding assay. The fraction of DOC binding to aluminum oxide — a common soil mineral that binds organic carbon [54] — ranged from 16.9% - 55.8% among the subset of DOC samples tested. Communities with high quantities of DOC had, on average, DOC with significantly greater potential for mineral-binding (two-tailed t-test, t_{122} = 2.8, P = 0.006; Figure 2.2B).

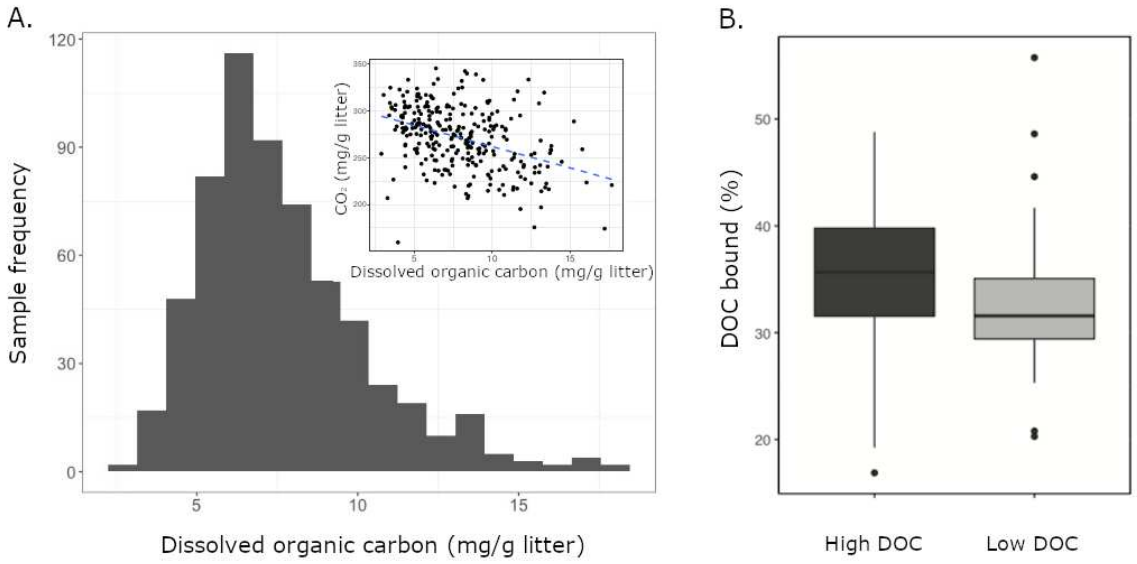

Figure 2.2: (A) Dissolved organic carbon (DOC) abundance among 611 microcosms after 44 days of pine litter decomposition. Inset panel - inverse correlation between CO_2 and DOC. (B) Proportion of dissolved organic carbon (DOC) that binds to aluminum oxide. DOC was obtained from microcosms after 44 days of pine litter decomposition. A greater proportion of DOC binds from high DOC than low DOC samples (P = 0.006, n = 128). DOC was produced by microbial communities during 44 days of decomposition of pine litter in microcosms.

2.4.2 Microbial communities with contrasting function were geographically intermingled.

The original soils were obtained from eight ecosystem types defined broadly by dominant and minor plant types or by agricultural land-use (Table 2.1). Ecosystem type significantly influenced the frequency of observing high or low DOC abundance (chi-squared test, 2 d.f., X2=17.89,

Table 2.1: Prevalence of DOC categories within ecosystem types

	Samples per DOC category		
Ecosystem type	**Low**	**Medium**	**High**
Grassland - shrub	23	44	50
Mixed	10	8	3
Juniper woodland - grass	10	3	2
Agricultural field - active	3	10	2
Agricultural field - fallow	4	6	3
Grassland - juniper	5	3	2
Pinon juniper woodland - grass	6	2	0
Pine forest	3	2	2

P<0.001) in the microcosm experiment. However, both functional outcomes occurred among source soils from all but one ecosystem type (Table 2.1), fulfilling the primary objective of acquiring diverse source communities that exhibit a similar functional pattern. Source soil samples yielding high versus low DOC communities in our study were also geographically intermingled (Figure 2.1) and co-occurred less than 30m apart at 14% of 49 geographic locations where two or more soil samples were collected from the same site.

2.4.3 Community features were linked to DOC abundance.

Community composition. Microbial community composition was a significant feature. The composition of microbial communities in the low versus high DOC groups differed significantly, both for the original soil communities and for the day-44 microcosm communities. (Figure 2.3). For the original soil communities and the day-44 microcosm communities, DOC was more strongly correlated with bacterial rather than fungal community composition (Mantel test; bacteria original soils r=0.26, P =0.001; bacteria day-44 microcosms r=0.28, P =0.001; fungi original soils r=0.19, P =0.001; fungi day-44 microcosms r=0.12, P =0.001). Microcosm bacterial communities at day-44 were slightly more correlated with DOC than original communities, while fungal communities showed the opposite trend.

Specific taxa. A suite of taxa were significantly linked to DOC abundance. In the original soil samples 17 of 31 bacterial families and in the day-44 microcosms 13 of 23 bacterial families

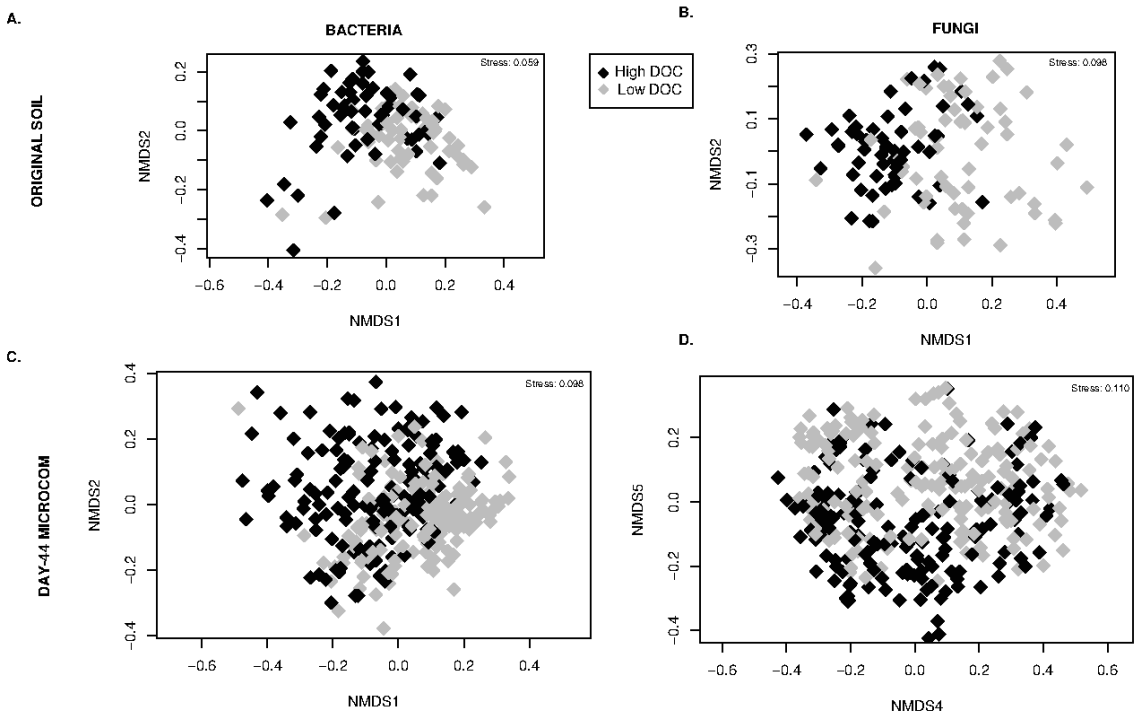

Figure 2.3: Relationship between microbial community composition and dissolved organic carbon (DOC) abundance. Non-metric multidimensional scaling ordinations performed on rarefied data for (A) bacterial communities in original soils, (B) fungal communities is original soils, (C) bacterial communities in day-44 microcosms (D) fungal communities in day-44 microcosms. Points are shaded by DOC cohort: high (black) and low (gray). The stress value is derived from six dimensions.

comprising on average at least 1% of the sequences were significantly different in relative abundance between high and low DOC groups (Figure 2.4A, Figure 2.4B). Among these families, only four (Methylobacteriaceae, Nocardioidaceae, Hyphomicrobiaceae, and Caulobacteraceae) showed consistent differences between DOC groups in both the original soils and the day-44 communities (Figure 2.4A, Figure 2.4B). Among the fungal orders comprising on average at least 1% of sequences, 6 of 18 orders in the original soils and 4 of 7 orders in day-44 microcosms were significantly different in relative abundance between high and low DOC groups (Figure 2.4B, Figure 2.4D). Eurotiales was the only fungal order that was significantly different (higher in the low DOC group) between DOC groups in both original and day-44 communities.

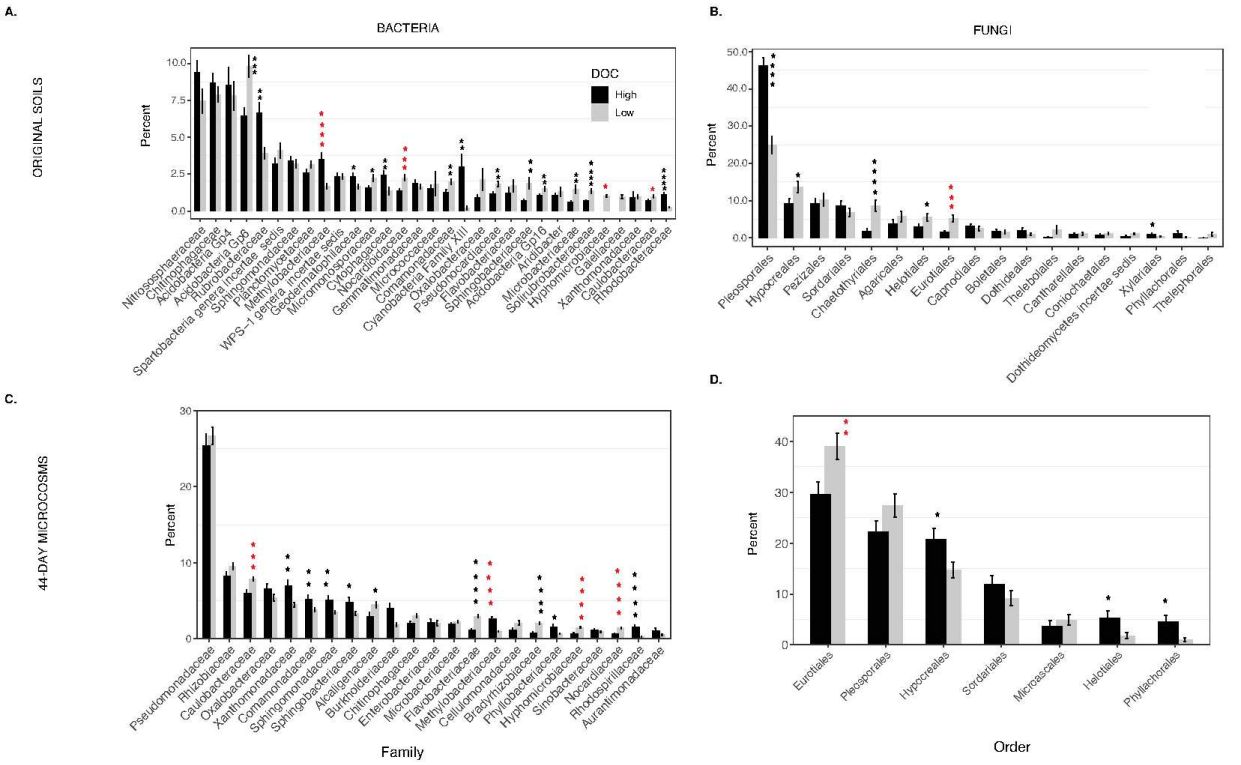

Figure 2.4: Operational taxonomic unit (OTU) richness in original soils for bacteria (A) and fungi (B). Shannon-Wiener diversity for bacteria (C) and fungi (D) in original soils. Operational taxonomic unit (OTU) richness in day-44 microcosms for bacteria (E) and fungi (F). Shannon-Wiener diversity for bacteria (G) and fungi (H) in day-44 microcosms. The data are expressed as the mean value $\pm$ SEM.

Biomass. Microbial biomass was a significant feature. Original soil communities in the high DOC group had, on average, 36% less biomass (measured as total extracted DNA) than those in

the low DOC group (two-tailed t-test, t_{126} = -3.87, P <0.001; Table 2.2, Figure 2.5). Similarly, day-44 microcoms communities in the high DOC group had 18% less biomass than those in the low DOC group (two-tailed t-test, t_{378} = 4.7, P <0.001; Table 2.2, Figure 2.5). Even so, DOC was only weakly correlated with biomass (Pearson correlation; original soils: r=-0.28, P =0.001; day-44: r = -0.22, P <0.001).

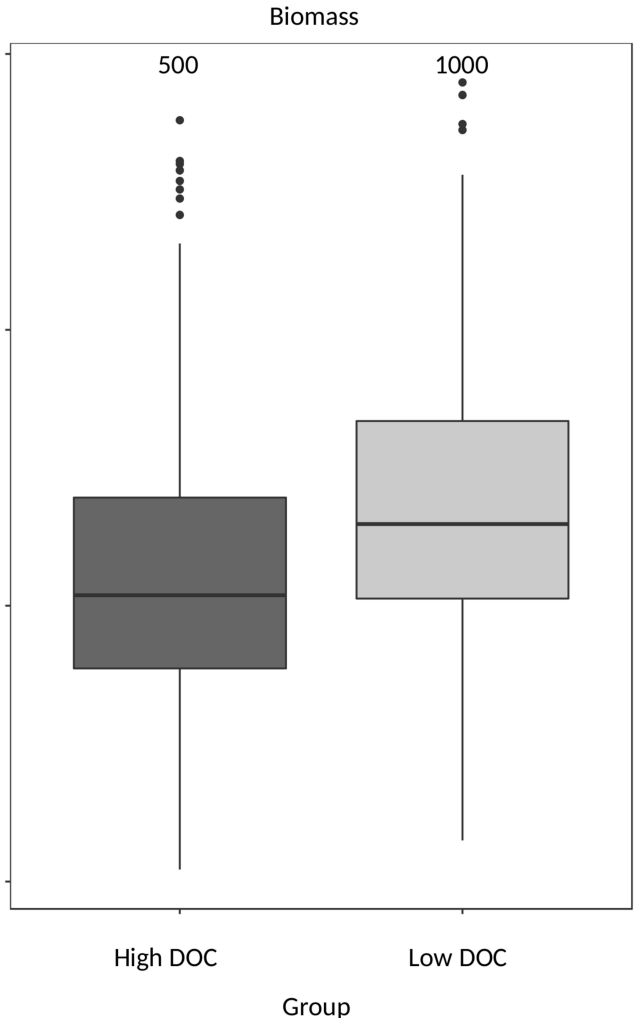

Figure 2.5: Biomass (as measured by DNA quantity) in high and low DOC microcosms. There is more biomass, on average, in low DOC microcosms than in high DOC microcosms (P = <0.001, n = 380).

Community richness and diversity. Microbial community richness and Shannon diversity were the most significant features linked to DOC (Table 2). Bacterial richness and diversity of the original soil and day-44 microcosm communities were significantly lower in the high compared to the

Table 2.2: Correlations between DOC quantity and community features

	Day 0				Day 44			
Feature	DOC correlation	P	% Diff of means	P	DOC correlation	P	% Diff of means	P
Bacteria Composition	0.26	**			0.38	**		
Biomass	-0.19	*	36	*	-0.10	NS	19	NS
Diversity	-0.27	**	34	*	-0.55	***	130	***
Richness	-.37	***	52	**	-0.64	***	143	***
Fungi Composition	0.19	**			0.12	**		
Biomass	-0.14	NS	23	NS	-0.14	NS	36	NS
Diversity	-0.30	***	33	NS	-0.02	NS	3	NS
Richness	-0.46	***	69	***	-0.08	NS	14	NS
Total Biomass	-0.28	**	60	***	-0.22	***	47	***
F:B	0.03	NS	-8	NS	0.03	NS	-7	NS

low DOC groups (two-tailed t-test; original soil: richness t_{114} =-3.2, p=0.002; diversity t_{112}=-2.39, P =0.02; 44 day microcosms: richness t_{306} =-13.74, P <0.001; diversity t_{288} =-10.39, P <0.001, Figure 2.6) . In both original soils and day-44 microcosms bacterial richness was negatively correlated with DOC quantity and was the community level trait most strongly linked to DOC in day-44 communities (Pearson correlation; original soils r = -0.39, P <0.001; day-44 microcosms r = -0.64, P <0.001). In the original soils fungal richness was also significantly lower in the high DOC group (two-tailed t-test; t_{116} =-4.0, P =0.0001) and negatively correlated with DOC quantity (Pearson correlation; r = -0.45, P <0.001). No differences in fungal richness were observed in the day-44 microcosms (two-tailed t-test, t_{343} =-1.32, P =0.187, Figure 2.6). Fungal diversity did not differ between high and low DOC groups in either original soil or day-44 microcosm samples (two-tailed t-test; original soil: t_{116}=-1.89, P =0.06; day-44 microcosm: t_{331}=-0.24, P =0.813, Figure 2.6).

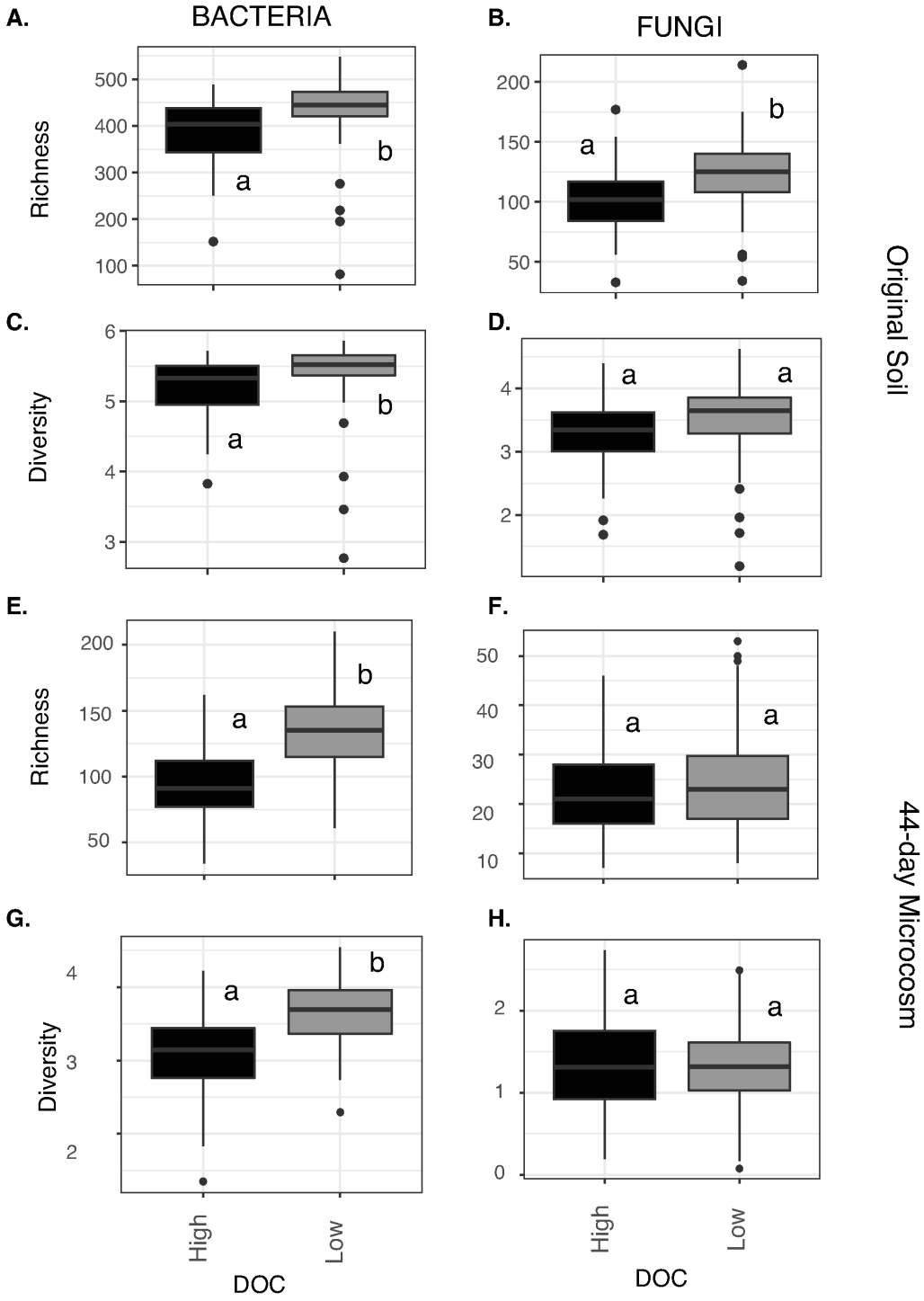

Figure 2.6: Diversity and richness of bacterial (A,C,E,G) and fungal (B,D,F,H) communities in high and low DOC microcosms.

2.4.4 Community features forecast high and low DOC outcomes

Logistic regression models predicted DOC abundance ("high" or "low") in the 44-day in microcosms significantly better than chance (z-test for a proportion, P< 0.05 using day-0 community features Figure 2.7A and P<.001 using day-44 community features Figure 2.7B). The average DOC prediction accuracy of the logistic model from 1000 permutations of training and test data was 0.72 and 0.80, when using feature values from the original soil communities and the final microcosm communities, respectively. In every permutation of training and testing data, the logistic regression model achieved greater prediction accuracy than the null model. In models using original soil community data, the feature importance (Wald statistic) of total biomass, fungal richness, and bacterial richness was -2.5, -1.5, and -1.0, respectively (Table 2.3). In contrast, the importance scores in models using day-44 microcosm community data were -4.5, -0.7, and -6.6 (Table 2.4).

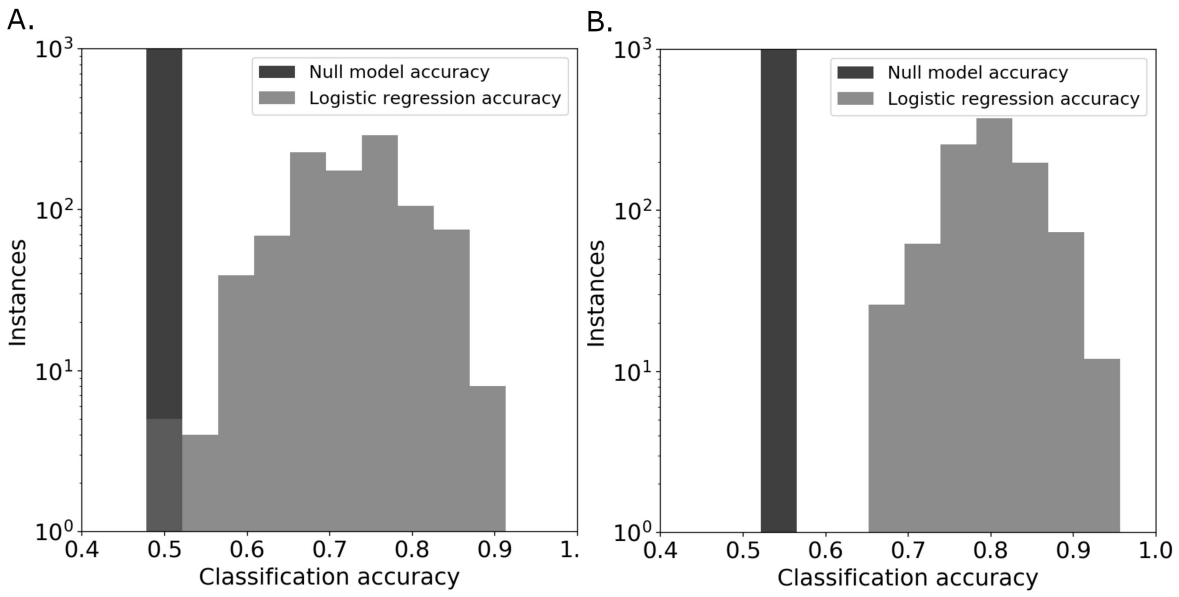

Figure 2.7: Logistic regression models for DOC. The null model consisted of automatic assignment of samples to the most prevalent DOC category that occurred in the test data set. (A) Distribution of prediction accuracy using day-0 community features. (B) Distribution of prediction accuracy using day-44 community features.

2.4.5 Logistic Regression using Day-0 Community Features

A logistic regression model used day-0 community features (total biomass, fungal richness, bacterial richness, fungal diversity, bacterial diversity, fungal abundance, and bacterial abundance) to predict day-44 DOC. To show the predictive power of the logistic regression model, the total data set was partitioned into 1000 unique permutations of training and testing data. Training data and testing data were partitioned such that the balance of high DOC and low DOC labels in each set was equivalent. Variables in the training data were standardized to be zero mean with unit variance, and variables in the testing data were similarly scaled using training data statistics. Scikit-learn's ([59]) LogisticRegression model was used to fit to the training data using an 'L2' penalty, which penalizes the squared magnitude of each regression coefficient. After fitting the model to each permutation of training data from day-0, features were selected using the default settings of Scikit-learn's SelectFromModel function, which rejects features with coefficients less than a threshold value of 1e-5. The logistic regression model was re-trained with the reduced feature set and applied to testing data. For each set of training data, the Wald statistic (defined as the regression coefficient divided by the standard error of the regression coefficient) was computed to provide a measure of the significance of model variables ([61]) and to assess feature importance. This process was performed for every permutation of training and testing data, with the Wald statistics and prediction accuracy stored after each permutation. The average regression coefficient, Wald statistic and p-value for each feature over 1000 permutations of training and testing are reported in Table 2.3. The overall average day-0 prediction accuracy was 0.724 ($P<0.05$), where accuracy is defined as the number of true positives and true negatives divided by the total number of samples in the testing dataset (Figure 2.7A). Statistical significance of model performance on testing data was computed using a z-test for the equality of two proportions to compare the proportion of correctly labeled samples using the logistic regression model to the proportion of correctly labeled samples using the null model.

Table 2.3: Average regression coefficients, Wald statistics, and P values of day-0 community features

Feature	Avg Coefficient	Avg Wald stat.	P values
TotalBiomass	-1.016	-2.474	0.013
fungAbundance	-0.008	-0.014	0.989
bactAbundance	0.000	0.002	0.999
FungRichness	-0.546	-1.532	0.126
FungDiversity	0.002	0.004	0.997
BactRichness	-0.804	-0.964	0.335
BactDiversity	0.702	0.759	0.448

Table 2.4: Average regression coefficient, Wald statistic, and p-values of day-44 community features

Factor	Avg Coefficient	Avg Wald stat.	P values
Total Biomass	-1.269	-4.482	$<.001$
Fungi Richness	-0.148	-0.698	0.485
Bact Richness	-2.245	-6.639	$<.001$

2.4.6 Logistic Regression using Day-44 Community Features

The top three most significant day-0 community features (i.e., those with Wald statistic p-values less than 0.4) were measured at day-44 and used to predict day-44 DOC using a logistic regression model. These data were partitioned into 1000 unique permutations of training and testing data as described above. Regression coefficients, Wald statistics, and p-values for each feature are summarized in Table 2.4. The average prediction accuracy on test data using day-44 features was 0.8 (P<0.001). The distribution of prediction accuracy over the 1000 permutations is shown in Figure 2.7B.

2.4.7 Logistic Regression Analysis Using Entire Data Set

As an added confirmation that our final regression models were significant, we combined all training and testing data and computed the McFadden pseudo R^2 value and log likelihood ratio P-values, similar to the conventional approach used by others (e.g., Maynard, 2018). These tests revealed a R^2 of 0.292 for the 7-feature Day 0 model (P=1.22E-07) and a R^2 values of 0.465 for the 3-feature Day 44 model (P=7.63E-39). Moreover, both analyses were in agreement for the most significant features: (Total BioMass, Fungal Richness, Bacterial Richness, Bacterial Diversity) for

Day 0 features, and (Bacterial Richness and Total Biomass) for Day 44 features. The pseudo R^2 values and LLR P values reflect the goodness of fit of the model to the entire data set. In contrast, the more modest P values reported in the main text represent the significance of model predictions on held-out testing data.

2.5 Discussion

Discovering microbial community features that drive large variation in soil carbon abundance independent of environmental conditions may improve soil carbon modeling and management. Up to 70-fold variation in CO_2 flux or litter mass loss has been observed in year-long field studies of litter decomposition, and abiotic variables failed to explain the majority of variance [62,63]. Given the magnitude of unexplained variation in field decomposition studies and in model predictions of soil organic carbon abundance [11, 13], deciphering the role of microbial community composition is a priority. In our study, we made two important findings: 1) we identified specific community-level features linked to DOC abundance, and 2) we showed the features have strong predictive power when measured before community succession and decomposition begin.

Holding the environment constant within laboratory microcosms while varying microbial community composition reveals an indisputable link between microbial community composition and decomposition outcomes. We built upon valuable prior work by reducing geochemistry as a confounding factor [64] and by using natural microbial source communities instead of isolate mixtures [65]. Moreover, we focused on community features driving DOC variation—a priority which has previously been neglected [27]. In our study, high versus low DOC cohorts differed significantly in microbial community composition (Figure 2.3 and Figure 2.4). The significant difference occurred among the native soil communities as well as among the decomposer communities that arose in the microcosms, demonstrating ecological succession and carbon flow in the laboratory microcosms were constrained by the historical state of the communities in soil. DOC abundance correlated more strongly with the initial (original soil) fungal community composition than the final fungal community composition (day-44 microcosms) while bacterial community composition

showed the opposite trend. Fungi are generally considered the main microbial drivers of plant litter decomposition due to their production of powerful enzymes for deconstruction of plant lignocellulose [66]. However, bacterial communities also contribute to decomposition outcomes [67]. Our results are consistent with the view that fungi are critical in launching major deconstruction of litter and driving the overall rate, while bacteria play an increasing role over time as secondary consumers shaping the quantity and quality of DOC that remains available for transport into soil.

The large range of variation in CO_2 and DOC outputs in our study combined with the general magnitude (c.a. 75 Pg) of natural CO_2 flux from soil microbial respiration [35, 68] supports the concept of steering soil microbial respiration to offset anthropogenic CO_2 emissions for climate change mitigation [69]. The true range in CO_2 flux that can arise from natural or manipulated microbial community variation within a natural ecosystem remains unknown. Variation in surface litter carbon flow may be counter-balanced in nature by compensatory processes over longer timescales [67] or in other components of the carbon cycle, such that an ecosystem will exhibit a fairly stable mean CO_2 flux. Nonetheless, our findings motivate further investigation of the potential to alter carbon flow over long time scales by manipulating microbial community composition.

The 5-fold range we observed in DOC abundance suggests a potential for microbial community control over soil carbon abundance. In natural systems, DOC from surface litter contributes substantially to soil carbon stocks [70]. When DOC from decomposing surface litter is transported to deeper layers, some of the carbon adsorbs to mineral surfaces [71, 72] enabling carbon stabilization over millennial timescales [73, 74]. Because the amount of carbon stored is related to the magnitude of DOC flux [70], microbial communities that yield larger quantities of DOC create a possibility for greater soil carbon storage.

Our results show that microbial community composition also alters DOC quality, which plays a role in soil carbon accumulation. Communities with high quantities of DOC had, on average, DOC with higher mineral binding potential. Enrichment of DOC with compounds that have greater affinity for mineral surfaces can increase carbon stabilization in soil [54]. Enrichment of DOC may occur through different mechanisms including (a) variable depletion of compounds released from

plant litter, (b) production of taxon-specific microbial by-products (e.g. polyphenolics produced by Actinobacteria) [75] and (c) release of taxon-specific residues from dead microbial cells such as melanin, chitin, B-glucans, or glycoproteins (e.g. glomalin) from fungi [76–78]. Combining the effects of DOC quantity and quality, we observed a 7-fold range in the quantity of carbon that could be readily-stabilized in soils. In a natural ecosystem, the realized quantity of carbon stored would depend on additional factors such as the magnitude of precipitation events for DOC transport to deep mineral layers [79], soil porosity [80], soil minerology and chemistry [81], and variation in the composition of subsurface microbial communities that control the extent of DOC decomposition during DOC transport through the soil [82].

Identifying specific community features that drive decomposition outcomes is a crucial advance beyond demonstrating a basic link between community composition and outcomes. Since regional and global carbon models cannot account for thousands of different microbial species' abundances, we focused on identifying broader traits that may predict DOC. Among the eight community-level features we examined, total biomass (extracted DNA), fungal richness, and bacterial richness were the most important features linked to DOC abundance (Table 2.2, Table 2.3). The predictive power of these features was robust, as indicated by the nearly equal performance of the set of features measured before or after six weeks of community succession. Simple logistic regression models with these three features predicted DOC outcomes ("high" versus "low") equal or better than regression models created with a random forest technique (data not shown). The reversal in the importance of fungal versus bacterial richness as DOC predictive features at the beginning versus end of the microcosm incubation again points to time-dependent roles of fungi and bacteria in the decomposition process that merit further investigation. The importance of initial fungal taxon richness suggests fungi create early priority effects that constrain the trajectory of decomposition and shape the assembly of bacterial communities that ultimately control DOC abundance and composition. The capacity to use easily measured community features to forecast the functional patterns of soil communities can simplify mapping the geographic distribution of a functional pattern that is driven by microbes, not the environment. To be climate relevant, an unex-

pected microbial functional pattern must be geographically prevalent to cause the mean behavior of an ecosystem to deviate from conventional soil carbon models. Our predictive DOC model is an encouraging first step towards such a capability. However, considerable validation is needed, including confirmation of prediction performance when applied to new soils and when applied to other litter types.

The strong correlation between lower bacterial richness and higher DOC abundance is a priority for further analysis. If bacterial richness proves to be a robust factor to predict DOC abundance among natural ecosystems, understanding the factors that control richness may reveal mechanisms that can be integrated in models, improving prediction for soil carbon stocks. Bacterial richness is known to vary at the landscape scale, declining with greater aridity [83, 84], and with lower pH [85]. However, richness that is strongly driven by environmental factors may be uninformative in soil carbon models because the empirically calibrated environmental variables in conventional models are likely to capture the linked functional consequences. Biotic interactions that affect species richness independent of the environment are more likely to create unexplained variance in soil carbon models. Biotic interactions that reduce richness and suppress function may include antibiotic production [86], predation [87], or bacteriophage activity [88].

2.6 Summary

To improve climate predictions by including microbial processes in soil carbon models, climate-relevant microbial processes and simple traits that represent them must first be identified, as has been achieved with plant traits [10, 89]. Our study showed a strong influence of microbial community composition over decomposition outcomes in a constant environment, resulting in large differences in carbon flow from litter decomposition. It is reasonable to expect that microbial composition drives variation in every component of soil carbon cycling (e.g. surface litter decomposition, subsurface litter decomposition, plant productivity and carbon allocation). Our findings motivate investigation of this phenomenon in natural systems to assess its importance to climate feedbacks within and among existing ecosystems and its implications for managing soil carbon.

We identified a high-level trait, bacterial richness, linked to DOC abundance and known to be geographically patterned. Bacterial richness has also been linked to carbon fate in mammals where lower richness correlates with increased carbon storage in the host [90, 91]. Our findings raise the tantalizing possibility of discovering robust principles that underpin functional states in extremely diverse systems ranging from soils to animal guts.

2.7 Conflicts of interest

The authors declare that they have no conflict of interest.

Chapter 3

Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition.[2]

3.1 Overview

Microbial communities are ubiquitous and often influence macroscopic properties of the ecosystems they inhabit. However, deciphering the functional relationship between specific microbes and ecosystem properties is an ongoing challenge owing to the complexity of the communities. This challenge can be addressed, in part, by integrating the advances in DNA sequencing technology with computational approaches like machine learning. Although machine learning techniques have been applied to microbiome data, use of these techniques remains rare, and user-friendly platforms to implement such techniques are not widely available. We developed a tool that implements neural network and random forest models to perform regression and feature selection tasks on microbiome data. In this study, we applied the tool to analyze soil microbiome (16S rRNA gene profiles) and dissolved organic carbon (DOC) data from a 44-day plant litter decomposition experiment. The microbiome data includes 1709 total bacterial operational taxonomic units (OTU) from 300+ microcosms. Regression analysis of predicted and actual DOC for a held-out test set of 51 samples

---

[2] (2019) Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition. PLOS ONE 14(7): e0215502. https://doi.org/10.1371/journal.pone.0215502

Jaron Thompson[a,*], Renee Johansen[b], John Dunbar[b], Brian Munsky[a,c]

a Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO 80523, USA
b Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, United States
c Keck Scholar, School of Biomedical Engineering, Colorado State University, Fort Collins, CO 80523, USA

***** I was responsible for the formal analysis, investigation, methodology, software, validation, and writing of this work.

yield Pearson's correlation coefficients of .636 and .676 for neural network and random forest approaches, respectively. Important taxa identified by the machine learning techniques are compared to results from a standard tool (indicator species analysis) widely used by microbial ecologists. Of 1709 bacterial taxa, indicator species analysis identified 285 taxa as significant determinants of DOC concentration. Of the top 285 ranked features determined by machine learning methods, a subset of 86 taxa are common to all feature selection techniques. Using this subset of features, prediction results for random permutations of the data set are at least equally accurate compared to predictions determined using the entire feature set. Our results suggest that integration of multiple methods can aid identification of a robust subset of taxa within complex communities that may drive specific functional outcomes of interest.

3.2 Introduction

Microbial communities mediate essential functions in diverse ecosystems. While the microbiome controls many interesting macroscopic properties, elucidating the relationship between specific microbes and ecosystem functions remains a complex problem in ecology. Recent advances in DNA sequencing technology make it easy to acquire metagenomic data representing the taxonomic profile of bacteria and fungi in microbial communities. This opens the door to deciphering which components of the microbiome can drive changes in macroscopic properties. However, analysis of metagenomic microbial data poses several difficulties. The data are typically high dimensional (many taxa) with a small number of samples collected in each study. Additionally, sequencing results are noisy and yield sparse data sets [92].

Machine learning techniques provide a means to analyze high-dimensional data [93, 94] and could be used to elucidate relationships between microbial taxa (or other metagenomic features such as gene families or metabolic pathways) and environmental attributes. The random forest model is reportedly one of the most effective machine learning models for analyzing microbiome data; high classification accuracy has been demonstrated with a variety of 16S rRNA data sets for identification of body habitat, host, and disease states [6]. In another study, artificial neural net-

works were used to map complex relationships between microbial communities and environmental variables, enabling predictions of the abundance of microbial taxa across the English Channel, for example [95].

While most existing machine learning software packages focus on binary classification of microbial data sets [7, 96, 97], random forest and neural network models can also be used to identify the subset of microbial taxa whose relative abundances best predict a continuous target variable [98, 99]. The combination of random forest and neural network models can evaluate feature importance and reveal which microbial taxa are most positively or negatively correlated with target variables. To provide helpful perspective for microbial ecologists, we compare results from these machine learning techniques to indicator species analysis, a commonly used tool in ecology that is typically used for classification, though similar techniques have been adapted for regression problems [100]. We also show how our tool can be applied to study the effect of experimental sample size on model performance by evaluating prediction error over increasing subsets of training data. In this study, we apply the proposed random forest and neural network regression models to predict the abundance of dissolved organic carbon (DOC) from plant litter decomposition, where bacterial taxa abundances are treated as model features/variables. We use DOC and bacterial community data from a study that examined the role of soil microbial community composition in controlling carbon flow from plant litter decomposition [101]. Feature selection results determined by machine learning methods are compared to indicator analyses [102, 103] in which high and low DOC are used as classification category labels.

3.3 Materials and methods

Random forest and neural network regression models are examples of supervised machine learning algorithms. In contrast to unsupervised machine learning algorithms, these methods require a subset of the data called a *training* set to develop a mathematical relationship between *features* and *target* variables. A *feature* represents a model variable and the *target* is the variable the model predicts. For regression problems, the target variable is a continuous scalar, and for clas-

sification problems, the target is a discrete label. A *sample* is a single set of features paired with a target variable, which, in the context of the present case study, represents a bacterial community profile paired with DOC. To assess model performance, predicted target variables using features from a held-out set of *test* data are compared to known target variables. In this study, prediction performance is measured using Pearson's correlation coefficient, which quantifies the linear correlation between predicted and true target variables, and for which a value of one indicates a perfect positive linear correlation. In general, our regression model assumes that targets and features are related to one another by

$$y = \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{x}) + \varepsilon, \tag{3.1}$$

where $\boldsymbol{x} \in \mathbb{R}^M$ is a vector M features, $y \in \mathbb{R}$ is the corresponding true value of the target variable, $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{x})$ is some mathematical operation (or model) from $\mathbb{R}^M$ to $\mathbb{R}$, $\boldsymbol{\theta} \in \mathbb{R}^{N_{\boldsymbol{\theta}}}$ are model parameters, and ε is the prediction error.

We denote the set of M features with N samples as the $N \times M$ feature matrix $\boldsymbol{X} \in \mathbb{R}^{N \times M}$, which can be mapped to a vector of N target variables $\boldsymbol{y} \in \mathbb{R}^N$ according to

$$\boldsymbol{y} = \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{X}) + \boldsymbol{\varepsilon}, \tag{3.2}$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^N$ is the vector of prediction errors. While Eq. 3.2 describes the general regression problem common to most machine learning algorithms, the actual form of $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{X})$ varies according to the specific approach. We introduce a few of these machine learning approaches as follows.

3.3.1 Neural network regression model

A feed-forward neural network regression model applies a series of parameterized activation functions organized in layers to map features in a sample to a continuous target variable. Each layer of a feed-forward neural network is composed of a set of nodes which apply a nonlinear

transformation to the sum of the product of inputs from the previous layer and weight parameters plus an additional bias parameter. A stochastic gradient descent algorithm minimizes the cost function by adjusting model parameters (weights and bias values for each layer) via a process called error back-propagation, which updates model parameters in each layer based on the gradient of the cost function with respect to model parameters. The rate at which model parameters change during training can be adjusted by a learning rate hyper-parameter, and the cost function can be adjusted with a regularization hyper-parameter, which ensures that model parameters do not reach disproportionate values [93]. We built a feed-forward neural network regression model using THEANO [104] and PYTHON 3.7 with a randomized search algorithm for determining model hyper-parameters implemented with SCIKIT-LEARN [59]. As a default, the model includes a single hidden layer with 15 nodes with sigmoid activation functions and a single output layer with a linear activation function. A randomized hyper-parameter search uses the training data set to find the optimum hidden layer size, learning rate, and regularization coefficient. Our model applies the mean squared error between predicted and true values as a cost function for use with the training and validation analyses. Training the neural network model is an iterative process, where each iteration is called a training epoch. In each training epoch, the total set of training data is divided and trained over randomly chosen mini-batches. Once the cost function applied to the validation data set fails to decrease over a default of ten training epochs, training stops. For this study, the model was trained with 257 training samples and tested with a held-out set of test data with 51 samples. To assess the correlation between true DOC and predicted DOC for each sample, Pearson's correlation coefficient was computed for training and testing results.

3.3.2 Neural network feature selection

Methods for evaluating feature importance using a neural network model often focus on weights assigned to individual features after training of the model [105, 106]. Our proposed feature selection tool employs a similar approach, where the gradient of the model output with respect to weights associated with each feature is used to determine the feature importance vector. Each el-

ement of the feature importance vector corresponds to an individual feature, where the magnitude of each element is indicative of feature importance for predicting the target variable, and the sign indicates whether the feature has a positive or negative impact on the predicted variable.

For a feed-forward neural network model with M features as inputs that connect to J nodes in the first hidden layer, we can denote the $M \times J$ matrix of weights connecting each feature to each node as $\boldsymbol{\theta}^{In} \in \mathbb{R}^{M \times J}$, where $\boldsymbol{\theta}^{In}$ is a subset of the full parameter set $\boldsymbol{\theta}$. The gradient of the model output with respect to $\boldsymbol{\theta}^{In}$ provides the $M \times J$ feature importance matrix, $\boldsymbol{F}(\boldsymbol{\theta}, \boldsymbol{x}) \in \mathbb{R}^{M \times J}$, which we define as

$$F_{mj}(\boldsymbol{\theta}, \boldsymbol{x}) = \frac{\partial}{\partial \theta_{mj}^{In}} \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{x}). \tag{3.3}$$

Marginalizing the feature importance matrix over all nodes in the first hidden layer produces a M-dimensional vector, which we will call the feature importance vector $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{x})$, whose elements are

$$f_m(\boldsymbol{\theta}, \boldsymbol{x}) = \frac{1}{J} \sum_{j=1}^{J} F_{mj}(\boldsymbol{\theta}, \boldsymbol{x}). \tag{3.4}$$

After training the model, we determine the sensitivity of the model to each feature, denoted as the M-dimensional vector $\mathrm{s} \in \mathbb{R}^M$, by calculating the average value of the feature importance vector over the set of training data with K samples

$$\mathrm{s}_m =< f_m >= \frac{1}{K} \sum_{k=1}^{K} f_m(\boldsymbol{\theta}, \boldsymbol{x}_k). \tag{3.5}$$

To gain confidence in the importance assigned to features, feature importance is determined using a bootstrap method, which randomly samples 80% of the training data set over a default of 50 iterations. The average feature ranking values determined over all iterations represents the most confident set of ranked features.

3.3.3 Random forest regression model

Decision tree based machine learning methods map features to target variables by splitting the set of possible target variables based on the values of individual features [93,107]. An *internal node* is a point at which the value of a feature determines a split in the set of possible target variables, and the nodes that follow an internal node are called *leaf nodes* [93]. The random forest method constructs a set of decision trees constructed from randomly selected subsets of the feature space and computes the model output by averaging the predictions from individual decision trees [108]. Using the random forest regressor from SCIKIT-LEARN [59], a random forest regression model is instantiated with a mean squared cost function, two samples required to split an internal node, and one sample required to be at a leaf node as the default. Hyper-parameters for the model include the number of samples required to split an internal node, the number of samples required to be at a leaf node, and the number of features to consider in each decision tree. These hyper-parameters can be optimized with the training set using SCIKIT-LEARN's randomized search algorithm. During training, the random forest regressor model fits an ensemble of 1000 decision trees trained on randomly selected sub-samples of the data set. All random forest results from this study use identical training and testing data to allow direct comparison to the neural network model.

3.3.4 Random forest feature selection

The random forest regressor made available by SCIKIT-LEARN [59] returns an array of feature importance values of length equal to the array of input features. Decision tree algorithms, such as random forest, assess feature importance by examining how well a feature (often referred to as variable in literature [107]) can split the potential output labels. In other words, a highly significant feature provides the greatest reduction of potential labels for a given sample. Additionally, feature importance is determined as part of the boot-strap method used for assembling random decision trees, where feature importance is greater for variables that result in greater prediction performance when included in the decision trees [107]. To gain confidence in the rank assigned to features, feature ranking is determined using a bootstrap method that randomly samples 80% of the training

data set over a default of 50 iterations. The highest average feature ranking values determined over all iterations represent the most confident ranked features.

3.3.5 Indicator species analysis for feature selection

Indicator species analysis [102, 103] is used for comparison with the feature selection results determined by the above machine learning methods. Indicator species (hereafter we use 'taxa', not 'species', for accuracy) are defined as the features that are most indicative of changes in DOC across different samples. To determine indicator taxa, a correlation value is calculated for each feature as the product of *specificity* and *fidelity* for a particular taxon in association with either high or low DOC samples [102]. Specificity measures how much a taxon associates with a single label (e.g., high or low DOC), and fidelity measures how frequently a taxon associates with that label. Specificity would be maximized if a taxon were only present in sites with a particular label, and fidelity would be maximized if a taxon were present at all sites associated with a particular label. A confidence score is assigned to each feature using a boot-strap algorithm that compares the correlation value for each feature determined using correct labels with correlations determined using randomly assigned labels. If the correlation statistic between features and site labels determined using random labels is not consistently lower than the correlation statistic using correct labels, then the confidence score for that feature-site correlation is low. Only taxa with at least a 95% confidence (features with correlation values greater than 95% of correlations determined with random labels) are considered in this study. Indicator taxa analysis was implemented in Python 3.7 with the methods described in Dufrene and Legendre, 1997 [102].

3.3.6 Data acquisition and data pre-processing

Microbiome data (16S rRNA gene profiles) were obtained from a prior study of pine needle litter decomposition in laboratory microcosms [101] (supporting information S1 Dataset). In brief, the microbial community in each of 206 soil samples was suspended in water, inoculated into three replicate microcosms containing sterile sand and pine litter, and incubated 44 days at 25C. At 44 days, the amount of DOC in the microcosms was measured, DNA was extracted from a

subset of microcosms, and 16S rRNA gene amplicons were sequenced on an Illumina MiSeq. Because the composition of bacterial communities among replicate microcosms diverged over the 44-day incubation period, the replicates were treated as independent samples. For machine learning analysis, however, the training and testing data were prohibited from sharing replicate samples to ensure independence between training and testing data sets (supporting information S2 Dataset, S3 Dataset). The bacterial community profiles from 308 samples were rarefied to 1023 sequences, which yielded a matrix with a total of 1709 bacterial taxa. By default, our tool standardizes features such that each feature is zero mean with unit variance over the training data set. The test data is similarly scaled but only using the sample statistics determined from the training data set.

3.4 Results

Our feed forward neural network regression model was trained with 257 community samples to predict level of DOC (Figure 3.1A). Our model was tested with a held out set of 51 test samples which yielded a Pearson's correlation coefficient of .636 between true and predicted DOC (Figure 3.1B) and a mean squared error of .565. The random forest regression model was trained and tested with identical sets of data used with the neural network model. Test results using the random forest regression model yielded a Pearson's correlation coefficient of .676 (Figure 3.1D) and a mean squared error of .516. A scatter plot of the prediction error using the neural network model versus the prediction error with identical test samples using the the random forest model are positively correlated with a Pearson's correlation coefficient of 0.743 (Figure 3.1E) .

To illustrate the degree of agreement of feature importance for predicting DOC between random forest, neural network, and indicator species approaches, Figure 3.4A shows a Venn diagram comparing feature selections. Feature selection was performed on the same training set used to produce Figure 3.1. Out of a feature set with 1709 taxa, 285 taxa were significant indicator taxa. Of the top 285 ranked features from the machine learning methods, 112 bacterial taxa were shared between random forest and neural network feature selections, and of these, 86 bacterial taxa overlapped with the set of indicator taxa. To further investigate agreement of feature importance be-

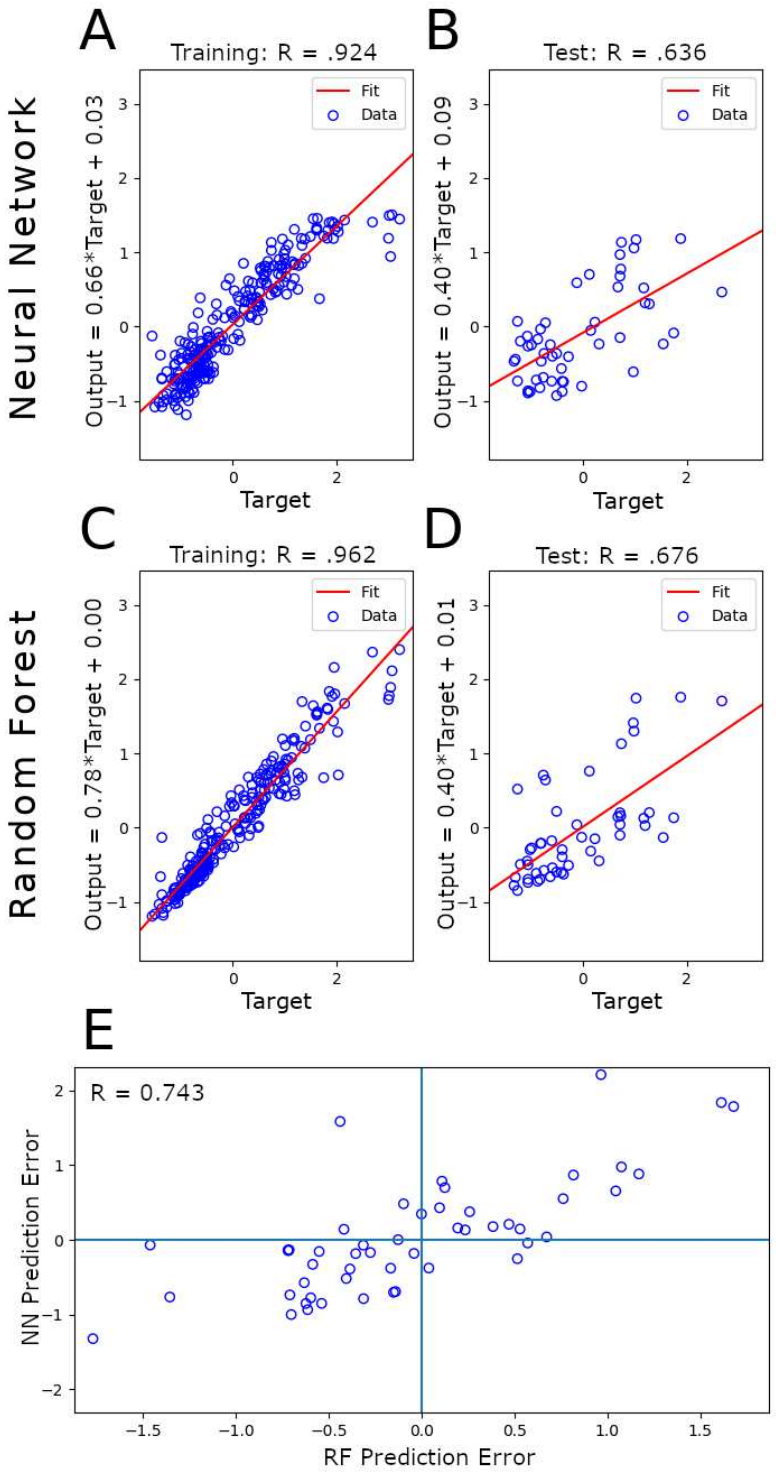

Figure 3.1: DOC prediction with neural network and random forest regression models. (A) Scatter plot of fitted DOC versus true DOC from training data samples (n=257) using neural network model. (B) Scatter plot of predicted DOC versus true DOC from test data samples (n=51) using neural network model. (C-D) Same as above but using random forest model. Training and testing data are identical for both methods. (E) A scatter plot of the prediction errors using the neural network model versus the prediction errors with identical test samples using the random forest model.

tween methods, Figure 3.4B shows how the shared set of ranked features determined by the neural network, random forest, and indicator taxa analysis varies as a function of feature rank. To investigate the significance of our feature selection results, we compared the number of features in the consensus set to the number of shared features that would occur if features were selected from three randomly organized sets. We applied a Monte Carlo approach that sampled features from three randomly organized sets of 1709 features and counted the number of features that were commonly selected in a pair of sets or within the intersection of all three sets. We plotted the mean and 99% confidence interval from 1,000 simulations as a function of the number of sampled features (a separate plot with just the Monte Carlo simulation curve is shown in Figure 3.2). The number of features in the consensus set is consistently greater than the number of shared features expected from random sampling, suggesting that each feature selection approach exploited similar, non-random trends in the data. Figure 3.4A,B show that feature importance determined by the neural network has greater agreement with indicator taxa compared to feature importance determined by random forest.

Indicator species analysis not only provides a feature importance metric, but also identifies which features are correlated with different labels, such as high DOC samples or low DOC samples. Feature importance determined by the neural network can be interpreted in the same way, where positive feature importance values imply a direct relationship with DOC, and negative values imply an inverse relationship. All 180 features shared by the neural network and the indicator species methods exhibit the same feature-label correlations. Figure 3.3 shows how prediction performance of the neural network and random forest models change as the number of features included in the model increases from a minimum of 10 features to a maximum of 86 features. The order in which features were included in each subsequent prediction corresponds to the rank determined by each feature selection method, such that the highest ranked features were included first. Both models reach close to peak prediction performance with only 86 features.

One might expect that the most informative features for DOC prediction would be those with highest or lowest abundances within communities. To examine this expectation, Figure 3.5A shows

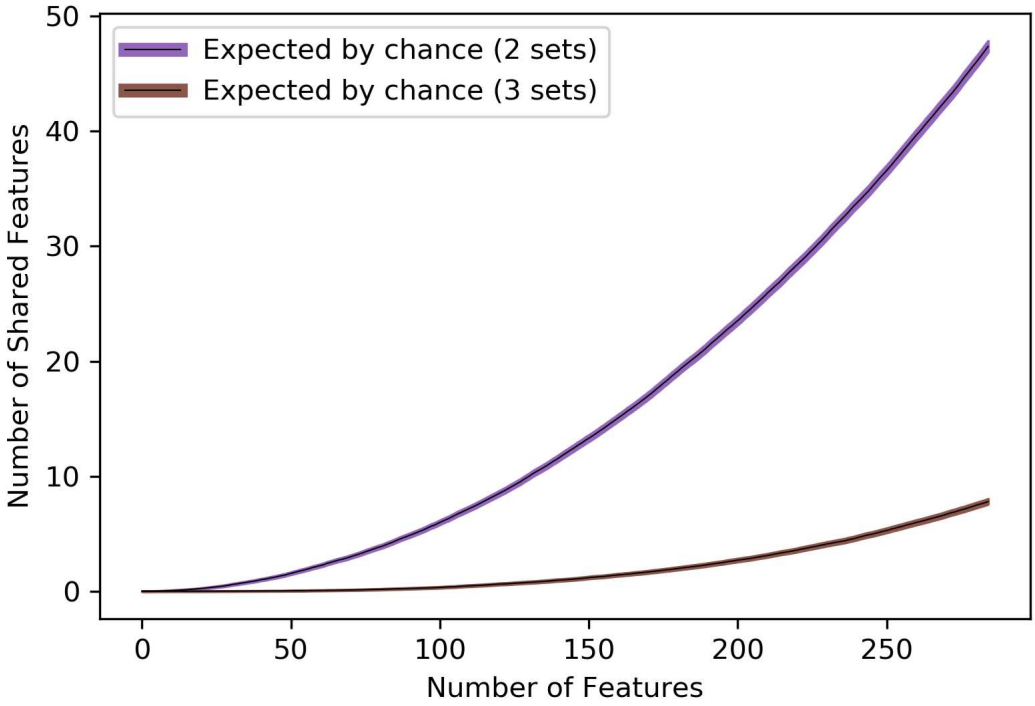

Figure 3.2: Monte Carlo simulation of the expected number of shared features after sampling from randomly organized sets of 1709 features. Monte Carlo approach samples features from three randomly organized sets of 1709 features to count the number of features commonly selected in a pair of sets (purple) or within the intersection of all three sets (brown). Plotted curves show the mean and 99confidence interval from 1,000 simulations as a function of the number of sampled features.

a histogram of bacterial abundance of the consensus set for selected features compared to the histogram of bacterial abundance for the entire data set. This shows that feature selection techniques are not biased towards selection of taxa with low or high abundance, but rather the consensus set of taxa selected by random forest, neural network, and indicator species analysis had abundance levels mostly in the moderate range. Abundance values in the figure were determined for each taxon by taking the average number of reads over the entire set of samples. Figure 3.5B shows the distribution of prevalence of bacterial species of the consensus set of selected features, where prevalence was calculated as the frequency in which taxa were present in each sample. The distribution of prevalence of selected taxa shows that prevalence was not a crucial factor in selecting features for prediction of DOC.

To test generality of the above results, we determined the Pearson's correlation coefficient for testing data under 50 randomly generated permutations of training and testing data with roughly

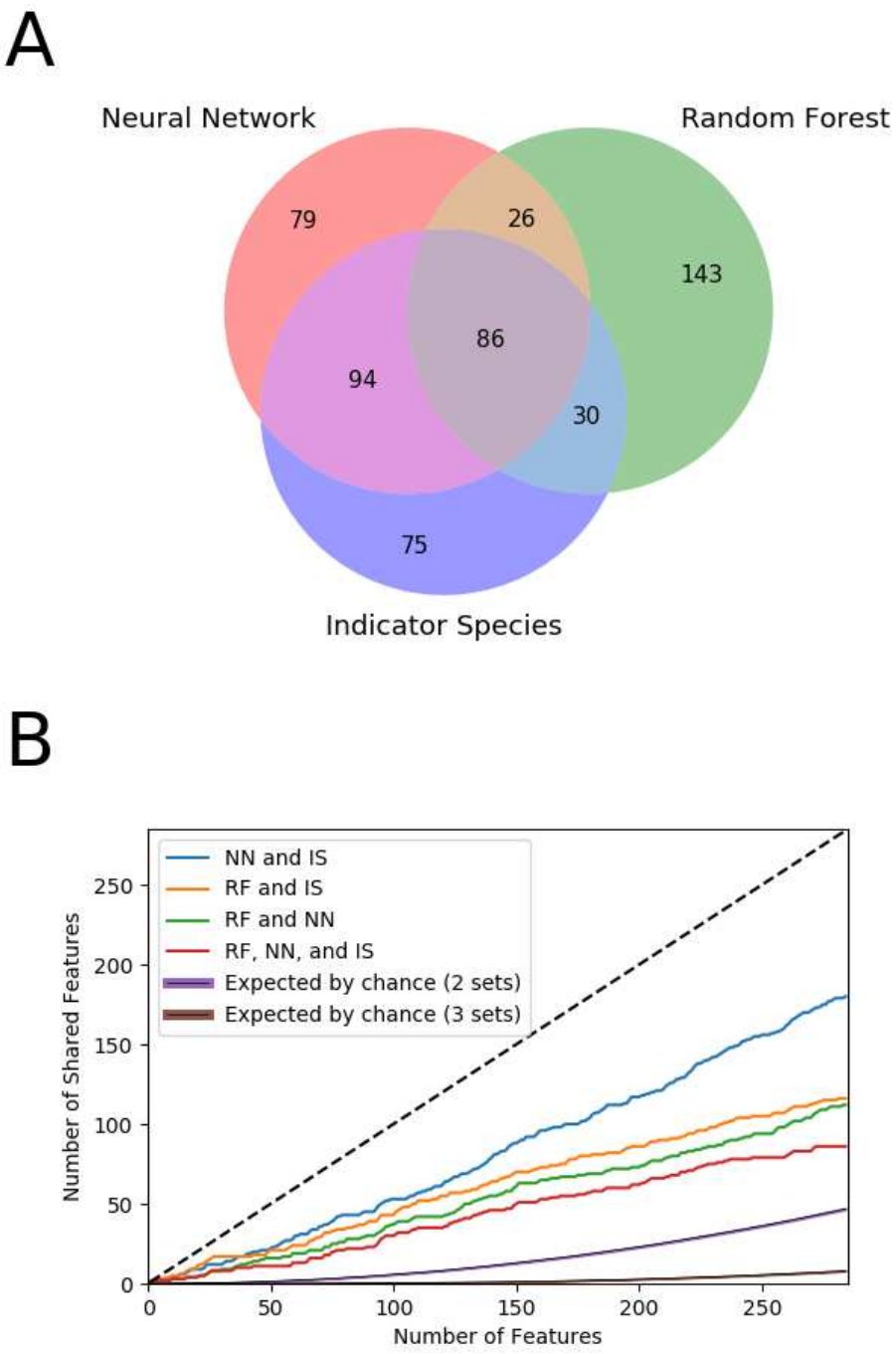

Figure 3.3: Feature ranking determined by neural network, random forest, and indicator species analysis. (A) Venn diagram demonstrates agreement of 86 bacterial taxa out of the top 285 ranked taxa from machine learning methods. (B) Plots of the number of shared features between NN and IS (blue), RF and IS (orange), RF and NN (green), and all methods (red) as a function feature rank over 285 features. Monte Carlo simulation of the number of shared features expected by randomly sampling from 3 sets of 1709 features is plotted with a 99% confidence interval (black line, purple confidence interval). The black dotted line indicates perfect agreement between the three sets of ranked features.

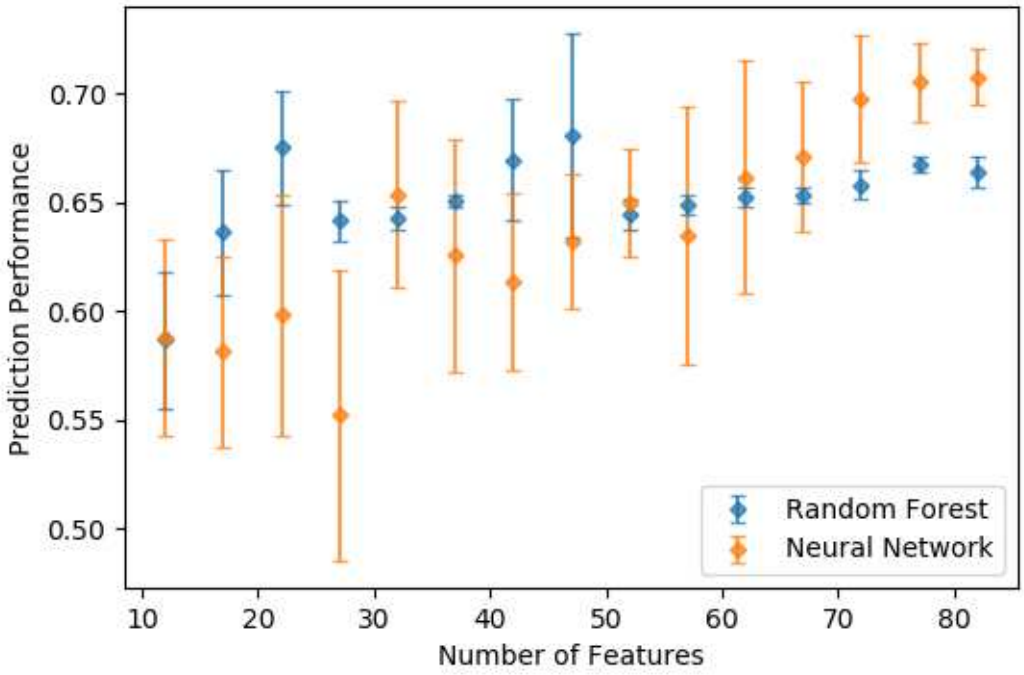

Figure 3.4: Correlation between random forest and neural network prediction accuracy on test data. Plot of prediction performance on test data as measured by Pearson's correlation coefficient versus number of features included in machine learning models. The data are binned such that each point represents the average prediction over 5 trials, where each subsequent trial includes an additional feature.

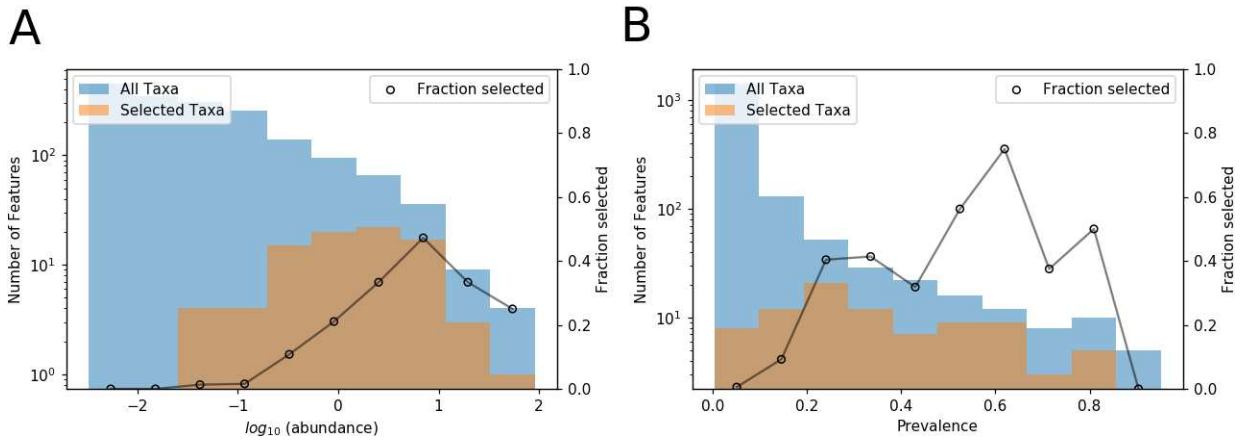

Figure 3.5: Distributions of bacterial abundance and prevalence of all taxa and the consensus set of taxa selected by all methods. (A) Histogram of abundance of taxa in the consensus set plotted over a histogram of abundance of all taxa in the data set. Abundance was calculated as the average number of taxa over the entire sample set. (B) Histogram of prevalence of taxa in the consensus set plotted over a histogram of prevalence of all taxa in the data set. Prevalence was calculated based on how frequently taxa were present in each sample.

260 training samples and 50 test samples (exact sample sizes varied between 254 and 262 samples for training data and between 46 and 54 samples for test data due to variations in the number of replicates per experimental condition). Figure 3.6 shows histograms of test performance of the neural network model and the random forest model using the full feature set (Figure 3.6A,C) and the reduced feature set (Figure 3.6B,D). While the neural network model performed better using the reduced set of 86 features (two tailed t-test, $P = .047$), the distribution of prediction errors using the random forest model with the reduced feature set was not significantly different (two tailed t-test, $P = .98$). The neural network model produced greater prediction accuracy using the reduced feature set on 70% of test samples, and the random forest model yielded greater prediction accuracy on 48% of test samples. The random forest model significantly outperformed the neural network model with the full feature set (two tailed t-test, $P < 0.001$) but only marginally so with the reduced feature set (two tailed t-test, $P = 0.11$).

To investigate how sample size affects model performance, prediction performance of the neural network and random forest regression models was measured with an increasing number of samples included in the training set (Figure 3.7). The random forest model consistently outperformed the neural network over the range of training data sample sizes, with more accurate predictions and less variability in prediction performance. Model performance of either method reaches near optimal levels after inclusion of only half of the training set or 150 training samples. Although variability in prediction performance continued to decrease as the fraction of training data increased, these results suggest that future experiments could be conducted with lower sample sizes without sacrificing model performance.

3.5 Discussion

While random forest outperformed the neural network for prediction tasks in this study, both methods can be used to predict DOC entirely from microbial community profiles and to provide measures of feature importance. The random forest method is relatively easy to implement, and performs well with little adjustment to model hyper-parameters. Sensitivity analyses with the

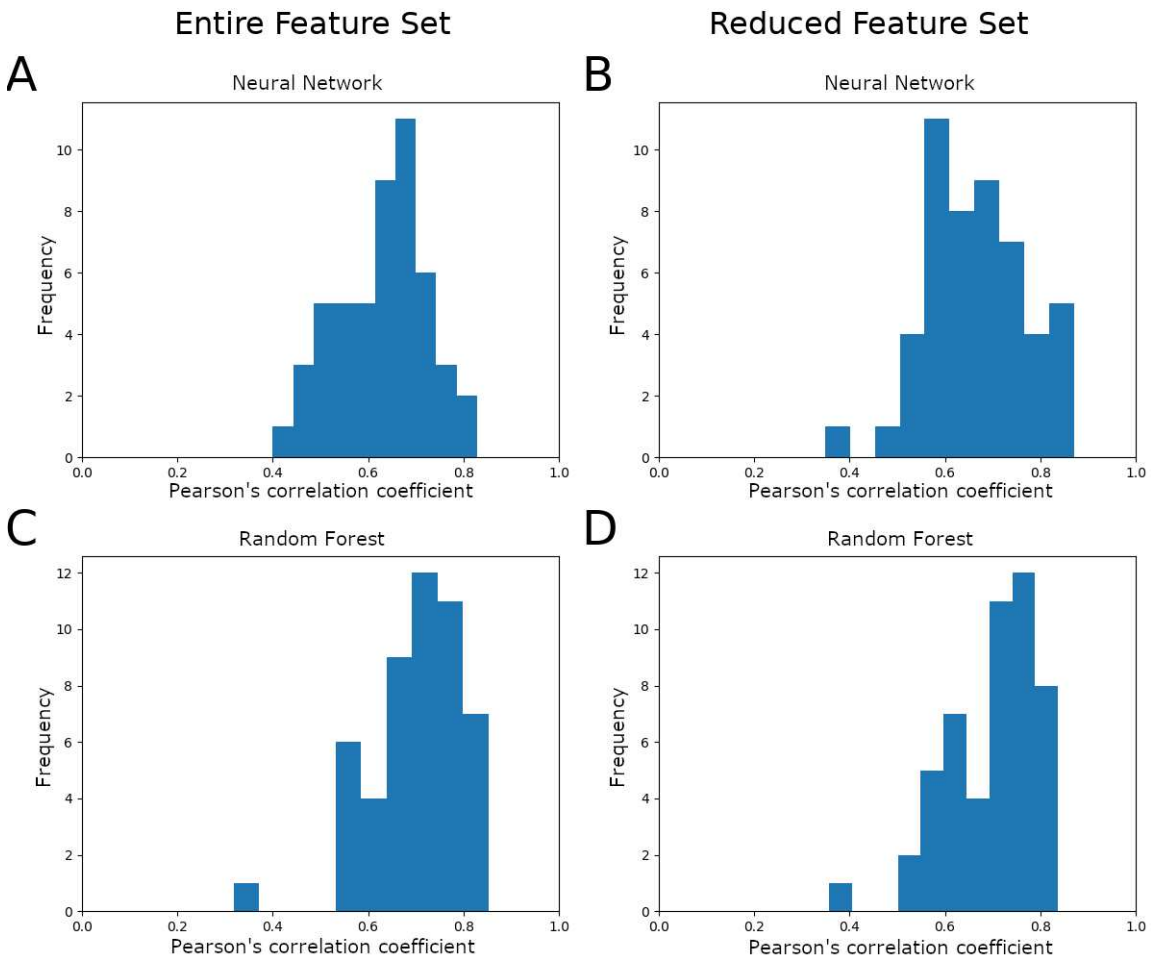

Figure 3.6: Distribution of prediction errors for 50 different permutations of training and testing data. (A) Distribution of Pearson's correlation coefficients on test data performance using the neural network model without feature reduction. Mean R value = .627, standard deviation = .097. (B) Distribution of Pearson's correlation coefficients on test data performance using the neural network model with the reduced feature set. Mean R value = .668, standard deviation = .103. (C) Distribution of Pearson's correlation coefficients on test data performance using the random forest model without feature reduction. Mean R value = .699, standard deviation = .100. (D) Distribution of Pearson's correlation coefficients on test data performance using the random forest model with the reduced feature set. Mean R value = .700, standard deviation = .095. For these permutations, feature reduction improved neural network prediction performance (two tailed t-test, $P = 0.047$), and random forest outperformed neural network with the full feature set (two tailed t-test, $P < 0.001$) and with the reduced feature set (two tailed t-test, $P = 0.11$).

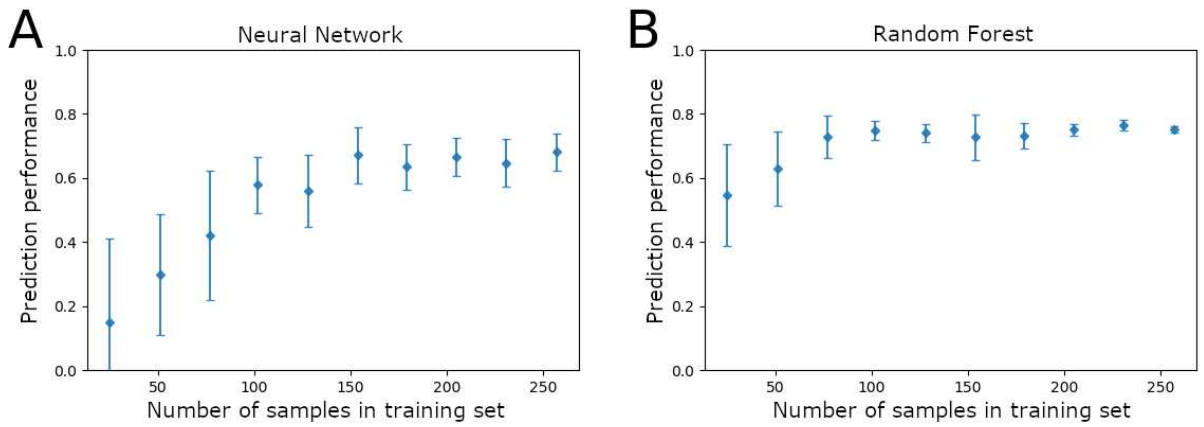

Figure 3.7: Sensitivity analysis of model prediction performance as the fraction of the total training data set (n=257) increases. Performance was measured using the average Pearson's correlation coefficient after training over 10 random samplings of a fraction of the data set, with error bars representing 1 standard deviation from the mean. (A) Prediction performance on fixed testing data by the neural network model. (B) Prediction performance on fixed testing data by the random forest model.

data set in this study (Figure 3.7) shows that the random forest model is less sensitive to sample size of the training data set, which makes random forest an attractive machine learning model for analysis of microbiome data. A benefit of the neural network model is that it provides more easily interpreted results for feature selection, which include the direction in which taxa affect environmental variables. The site correlations determined by the neural network and indicator taxa analysis show perfect agreement in sign among the entire set of taxa. Furthermore, because ground truth for which taxa drive changes in environmental variables is not known, the joint set of selected features from random forest, neural network, and indicator taxa approaches provides greater confidence than the set from one method alone (feature selection results are included in the supporting information S5 Dataset).

Machine learning approaches for analyzing microbiome data have proven successful in applications such as forensics, medicine, and agroecology [109–111]. Recently, machine learning algorithms such as random forest and K-means clustering have successfully determined the postmortem interval (PMI) using postmortem skin microbiome [109]. In medicine, machine learning models such as random forest have been used for identification of gut microbiomes associated with irritable bowel syndrome in pediatric patients [110]. In another study focusing on soil microbiomes, a

random forest model was applied to predict crop yields from soil microbiome composition [111]. With increasing access to machine learning software and high-dimensional microbiome data, machine learning is emerging as a powerful tool for understanding how microbial communities affect their environment.

Although there are several examples of platforms that facilitate use of machine learning techniques with microbial community data, our platform provides several unique options that make it more accessible and useful for microbial ecologists. QIIME [112] includes the "sample classifier" plugin [98], which provides access to a host of SCIKIT-LEARN [59] implemented machine learning classification and regression models for use with microbiome data. Although the sample classifier QIIME plugin includes hyper-parameter optimization and feature selection of important bacterial taxa, it does not provide insight into directional relationships between bacterial taxa and target variables. Moreover, the sample classifier plugin is not set up to provide combined feature selection results determined from different machine learning methods, and feature selection is not determined using different permutations of the training data. METAML [7] is another available software for implementing machine learning methods with microbiome data, but the methods are implemented exclusively for classification problems. For implementation of a neural network regression model with microbial abundance data, NEUROET [99] provides a simple GUI that can be used to train and test a single-layer, feed-forward neural network. NEUROET includes a procedure to optimize neural network architecture and identify important features for predicting model output, though optimization of hyper-parameters such as learning rate and the regularization coefficient is not available. While these platforms achieve a similar goal of applying machine learning techniques to microbiome data, no existing software packages include both neural network and random forest models and most do not provide insight into correlations between features and target variables. To provide the most confident set of important taxa, our tool produces the joint set of selected features from indicator species analysis, random forest, and neural network approaches. To aid in experimental design, our tool also provides a built-in tool for analyzing model sensitivity to experiment sample sizes.

Machine learning models offer the ability to determine hypothetical microbial communities that could promote increased levels of DOC. Recent studies have shown that microbial communities play an important role in carbon cycling and can potentially be manipulated to increase the abundance of DOC for transport and sequestration in deeper soil layers [4, 113–116]. Enhancing carbon sequestration in soil is a strategy to combat climate change, as sequestration has the potential to offset fossil-fuel emissions by 0.4 to 1.3 gigatons (5 to 15 percent) of atmospheric carbon per year [4]. Under the assumption that a trained machine learning model has learned a general relationship between microbial abundance and DOC, we can use the model to determine a hypothetical microbial community that could potentially maximize DOC. In consideration of this task, the random forest and neural network models are markedly different. Although the random forest model has been at least as good as the neural network model to predict DOC levels that lie within the range of the previous training data, the random forest model is restricted by its formulation to a finite set of values corresponding to leaf nodes of decision trees. As a result, the random forest model is incapable of predicting values outside of the range presented in the training data. Conversely, the neural network model could in principle extrapolate to make predictions outside of the range present in the training data, which would enable specification of hypothetical microbial communities predicted to increase DOC beyond empirically observed levels. Furthermore, because the feature importance vector, $\mathbf{s}$, produced by the neural network model is calculated as the gradient of the model output with respect to weights applied to features, $\mathbf{s}$ provides a potential direction in which features could be adjusted to increase levels of soluble carbon.

Figure 3.8A shows how the trained neural network model predicts responses to changes in microbial communities. In this simulation, communities (a) and (b) were initialized as the specific communities $\mathbf{x}_a$ and $\mathbf{x}_b$ that had the highest and lowest DOC and then adjusted in the direction defined by the feature importance vector according to $\mathbf{x}_{\text{new}} = \mathbf{x} + \alpha\mathbf{s}$, where α denotes the magnitude of the perturbation made to the community. The dashed trajectories represent DOC predictions made from simulated communities also initialized at the highest and lowest DOC, but with perturbations in random directions generated from a zero mean multivariate Gaussian distribution

scaled by magnitude α. As the microbial community profiles were adjusted in the direction of the gradient determined by the neural network, the level of predicted DOC increased (see communities (a) and (b) in Figure 3.8A). When the same initial communities were adjusted randomly, predicted DOC never exceeded DOC predictions determined from communities $\mathbf{x}_a$ and $\mathbf{x}_b$ (see dashed blue and orange lines stemming from the same initial values as in communities $\mathbf{x}_a$ and $\mathbf{x}_b$). For the neural network model, community (a) results in predicted DOC levels that exceed the greatest DOC prediction from the training set, thus generating testable hypotheses to supplement communities to increase dissolved organic carbon. When the same simulated microbial communities were analyzed on a trained random forest model (Figure 3.8B), the model predicted a similar trend towards increasing DOC for community (b). Due to the nature of the algorithm, however, the level of DOC predicted by the random forest model could never exceed that of community (a). Simulation results using either model suggest that simulated communities informed by the trained neural network model are not random and produce theoretical microbiomes that could promote greater levels of carbon in soil, though future experiments are needed to test these designs and verify these predictions.

Machine learning methods presented in this paper are intended to be easily applied to any data set that relates microbial communities to a scalar variable. To make this readily accessible, we have implemented all methods as a user-friendly platform available in the THOMPSON_ETAL_PLOS_ONE_2019 repository at GITHUB.COM/MUNSKYGROUP. For users without substantial knowledge of machine learning techniques, our tool enables application of machine learning regression models with optimized model parameters in a few lines of code. Tutorials for installing dependencies and using our machine learning tool can also be found on the GITHUB repository. In this study, we applied machine learning approaches to elucidate the relationship between bacterial communities and carbon flow from plant litter decomposition by developing regression models to predict dissolved organic carbon (DOC) concentrations. For the dataset we analyzed from [101], a strong relationship exists between bacterial community composition and

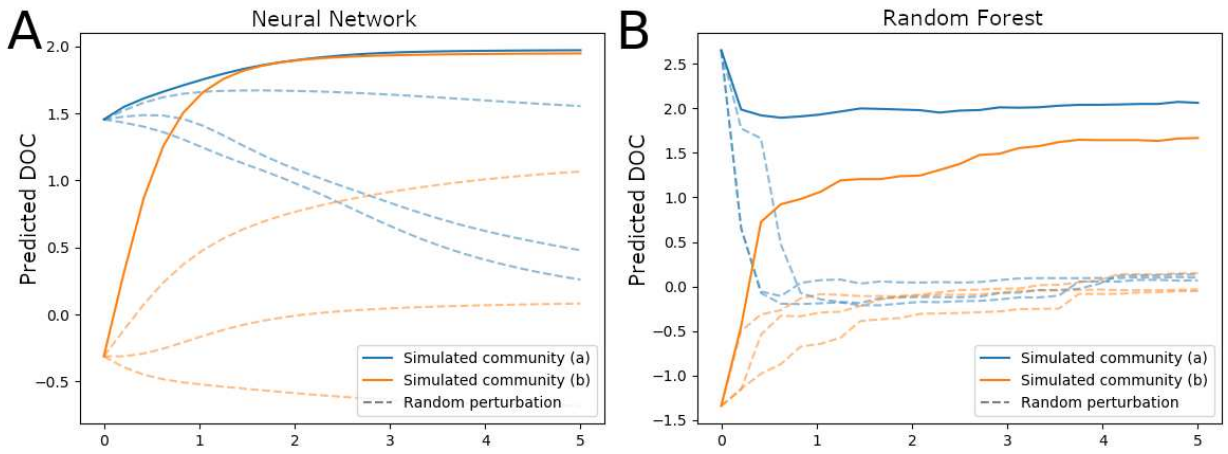

Figure 3.8: DOC predictions of trained machine learning models with synthesized microbial communities. Simulated communities (a) and (b) were specified by the training data communities with the highest and lowest DOC values, respectively. Each was then adjusted in the direction of the average gradient of maximum DOC increase determined by the neural network model, and each perturbation was scaled by magnitude α. Dashed lines stemming from the initial values of communities (a) and (b) represent DOC predictions from communities adjusted by a random vector with similar magnitude. (A) DOC prediction from hypothetical bacterial communities made by the neural network. (B) DOC prediction made by the random forest model with identical communities used in panel A.

DOC abundance. Moreover, we found a consistent set of bacterial taxa identified by multiple methods – in this case neural network, random forest, and indicator species approaches.

With our platform, a table of feature selection results from random forest, neural network, and indicator species analysis is easily produced with a built-in feature selection function. Model sensitivity to sample sizes is also easily visualized using a built-in sensitivity analysis that plots prediction performance on testing data as the size of the training data set increases. The combination of machine learning tools and indicator species analysis reduced the feature set of 1709 taxa to 86 taxa, which is a critical step towards elucidating mechanistic relationships between microbial communities and environmental factors. Sensitivity analysis performed with the neural network and random forest models suggests that future studies could be performed with smaller sets of samples. Feature importance determined by the neural network could direct future studies by proposing microbial communities that enhance a functional outcome of interest, such as increased carbon flow into soil. In this context, the proposed machine learning tools provide a framework

for designing experiments to further investigate how microbial communities function together to affect their environment.

3.6 Funding statement

JT, RJ, JD and BM were supported by grant F255LANL2018 from the U.S. Department of Energy Office of Biological and Environmental Research, Genomic Science program. BM and JT were supported under award number R35GM124747 from the National Institute of General Medical Sciences of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 4

Bayesian networks accurately estimate levels of dissolved organic carbon across independent litter decomposition studies[3]

4.1 Overview

Overwhelming complexity of microbiomes makes it difficult to decipher functional relationships between specific microbes and ecosystem properties. While machine learning analyses have demonstrated an impressive ability to correlate microbial community composition with macroscopic functions, mechanisms that dictate model predictions are often unknown and predictions often lack an assigned metric of uncertainty. In this study, we applied Bayesian networks that build on prior feature selection analyses to construct easily interpreted probabilistic models that accurately predict levels of dissolved organic carbon (DOC) from the relative abundance of soil bacteria (16S rRNA gene profiles). In addition to standard cross-validation, we show that a Bayesian network model trained using samples from a pine litter decomposition study accurately predicts DOC of samples from an independent oak litter decomposition study, suggesting that mechanisms driving variation in soil carbon storage may be conserved across different types of decomposing plant litter. Furthermore, the structure of the resulting Bayesian network model defines a minimal

---

[3] Jaron Thompson[a,*], Nicholas Lubbers[b], Marie Kroeger [c], Renee Johansen [c], John Dunbar[c], Brian Munsky[a, d]

a Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO, United States of America
b Computer, Computational and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM, United States of America
c Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, United States of America
d Keck Scholar, School of Biomedical Engineering, Colorado State University, Fort Collins, CO, United States of America

***** I was responsible for the formal analysis, investigation, methodology, software, validation, and writing of this work.

set of highly informative taxa whose abundance directly constrain the probability of high or low DOC conditions. Significant accuracy of the Bayesian network model with independent data sets supports the validity of the identified relationships between taxa abundance and DOC.

4.2 Introduction

Microbiome functions are extremely diverse and can potentially be optimized to efficiently perform industrially relevant tasks. However, engineering microbiomes requires validated models that accurately predict functions from microbial community profiles. Analytical techniques to infer microbial taxa that control macroscopic functions must overcome several obstacles. First, microbiomes are often composed of thousands of bacterial and fungal species [117]. Second, measuring microbiome composition by DNA sequencing has large measurement uncertainty, and studies are often limited to a small collection of experimental observations. Third, microbiomes are typically spatially heterogeneous, which increases measurement noise. Fourth, the composition of microbiomes varies by ecosystem [118], creating additional compositional variation that impedes efforts to identify “universal" taxa drive community functions.

Machine learning approaches are well-suited to deal with these complexities, with proven success in reducing the large microbiome feature space to a small subset of microbial taxa that are most informative for prediction tasks [119]. Machine learning techniques, such as random forest algorithms, have demonstrated impressive prediction accuracy with microbiome data [6, 101]. However, existing machine learning analyses of microbiomes are often limited by three major issues: First, machine learning models that are typically applied to metagenomic data lack the ability to quantify prediction uncertainty, which prevents users from knowing whether a prediction can be trusted in a new experimental circumstance. Second, machine learning analyses of microbiome data are often considered to be a “black box", in which the mechanisms that underpin model predictions are unknown or difficult to conceptualize, which makes models less useful from an engineering perspective. Finally, machine learning models are rarely validated with data from

independent studies, relying only on cross-validation of held-out testing data to confirm model generalizability [7].

In these regards, probabilistic machine learning models offer a clear advantage over deterministic models. Probabilistic models provide a quantified prediction confidence that reflects uncertainty resulting from noise in the data and limited sample sizes [93]. A probabilistic modeling framework that has gained popularity in modeling biological processes is the *probabilistic graphical model* [120–122]. PGMs not only provide estimates of prediction uncertainty, but also address the issue of model interpretability by providing a visual interpretation of how variables influence model predictions. The graphical structure of PGMs illustrates connections between model variables, which makes model predictions a transparent function of model inputs. Probabilistic models such as PGMs can also be more insightful than deterministic models in cross-study validation analyses, in that prediction uncertainty should be amplified when the model is extrapolated to new and untested circumstances. Uncertainty of model predictions, in addition to accuracy, can help gauge model relevance to the validation data.

A *Bayesian network* is a specific type of probabilistic graphical model that connects model variables with directed edges to describe the dependence properties of model variables [8]. The directed graph structure of Bayesian networks allows the user to identify a *minimum* set of features, known as a Markov blanket (see methods) that directly constrains the probability of the target variable or variables that the user wishes to predict. The combination of probabilistic predictions, model interpretability, and feature reduction make Bayesian networks an attractive approach for modeling microbial community behavior. However, beyond a few pilot studies [123, 124], the use of Bayesian networks for modeling microbial community behavior remains rare [125] due to the computational cost of learning Bayesian networks over many variables.

In this article, we applied Bayesian networks to predict levels of dissolved organic carbon (DOC) from the abundances of microbial taxa in decomposing plant litter. This case study is motivated by the need to understand the role of the soil microbiome in carbon cycling. The accumulation of greenhouse gases, especially carbon dioxide (CO_2), in the atmosphere is a primary

cause of global warming [126]. One potential mitigation strategy to offset greenhouse gas emissions is to increase soil carbon sequestration, which has the potential to reduce emissions by 0.4 to 1.2 gigatons of carbon per year [4]. In terrestrial ecosystems, the fate of carbon (i.e., whether released as CO_2 or stored as soil organic matter) is largely determined by microorganisms [5]. As microorganisms metabolize plant litter, CO_2 is released from microbial respiration and a variety of dissolved carbon compounds (e.g. sugars, peptides, lipids) that may be stabilized as soil organic matter are released either from deconstructed plant material or from microbial cells [118, 127]. While the composition of the soil microbiome is known to significantly influence the fate of carbon in soil, methods to elucidate specific microbial taxa that drive carbon fate are limited [3]. Overcoming this methodological challenge could enable the design of microbial communities that promote increased soil carbon storage. More powerful data analysis techniques may be a solution.

To address the high dimensionality of microbiome data, we first applied the RFINN [119] framework for feature reduction to determine the most informative features for predicting DOC. We then trained a Bayesian network structure over this reduced set of features to determine a set of bacterial genera whose abundances directly influence the probability of high or low DOC. In addition to cross-validation with held out testing data, cross-study validation was used to investigate model generalizability and transferability. We showed that a Bayesian network model trained on samples from a pine litter decomposition experiment [101] can accurately and confidently distinguish between high and low DOC samples from an independent oak litter decomposition experiment. When tasked with predicting DOC of samples from an independent grass litter decomposition experiment, the model was less accurate, but appropriately assigned less confidence in its predictions. These results suggest that the mechanisms that drive variation in soil carbon storage may be conserved at least partially across litter types. Furthermore, we showed that trained Bayesian networks correctly assign greater prediction uncertainty when presented with litter types that are less similar to the training data set. Ultimately, high-confidence predictions of such models could be used to suggest novel communities that are more likely to promote accumulation of soil organic carbon.

4.3 Materials and methods

4.3.1 Plant litter decomposition experiments

Plant litter decomposition experiments were performed in laboratory microcosms consisting of 125ml serum bottles containing 7g of sand and 0.12g of plant leaf litter (either Ponderosa pine needles, scrub oak leaves, or a 50:50 mix of two grasses (Stipa and Hilaria) common in the U.S. southwest). The microcosms were inoculated with soil microbial communities, sealed with crimp caps, and incubated at 25°C for about 6 wks. Cumulative CO_2 was determined from repeated measurements over the incubation period and the final abundance of DOC was measured at the end of the incubation. For the pine experiment, “low" and “high" DOC communities were delineated as the left and right thirds (approximately) of the observed DOC distribution, and DNA representing these microcosm community cohorts was extracted for community analysis. The communities were taxonomically profiled by PCR amplification and Illumina sequencing of fungal and bacteria ribosomal RNA gene fragments. The pine litter experiment was described in detail [101]. The oak and grass litter experiments were performed similar to the pine litter experiment, with the following modifications: only 100 (not 206) source soil communities were used to inoculate microcosms, and each source soil community was inoculated into 2 replicate (not 3) microcosms

4.3.2 Data pre-processing

The pine data were rarefied such that OTU abundance in each sample totaled to 1026 counts, and the oak and grass genera were rarefied such that OTU abundance in each sample summed to 1023 counts. The oak and grass OTU tables were normalized to match the pine data composition so that OTU abundance of each sample summed to 1026 counts. For all analyses, taxa were evaluated at the genus taxonomic rank. To calculate relative genera abundance, OTUs classified with greater than 70% confidence for a particular genus were summed together. For validation on held-out data, the pine data set of 308 samples was divided according to a K-fold partitioning scheme with 5-fold permutations of training and testing data. K-fold partitioning of the data set ensured that each set of testing samples was unique so that all samples were subject to held-out

testing. In each permutation, replicate samples were kept within either training or testing data sets to ensure independence between sets. Genera abundance was binned into categorical variables representing low, medium, and high abundance levels based on training data statistics (see next section for detail). Similarly, DOC was binned into high or low categorical variables based on the median DOC in the training data. Genera abundance in the oak and grass datasets were binned into categorical variables based on the abundance of taxa in the pine litter (training) dataset. Measured oak and grass DOC was not binned, but left as a continuous variable that was compared to the predicted probability of high DOC.

4.3.3 Probabilistic graphical models

Probabilistic graphical models combine probability and graph theory to construct a diagrammatic representation of a joint probability distribution over a set of random variables. The graph, $\mathcal{G}$, is composed of a set of nodes, $\mathcal{N}$, and edges, $\mathcal{E}$, where each node represents a random variable and edges represent a probabilistic dependence between the nodes. Bayesian networks are a type of probabilistic graphical model with directed edges that represent conditional dependencies between nodes [8, 93]. A graph is said to be fully connected if every pair of nodes is connected by an edge. While any distribution can be represented by a fully connected graph, the lack of edges in a graph constrains the distribution that the graph represents by enforcing conditional independence properties between nodes. The *Markov blanket* for a particular node in the graph is the set of nodes that, when observed, render the node conditionally independent of all other nodes in the graph. For a Bayesian network, the Markov blanket is defined as the parents, children, and other parents of children of a particular node [8]. Bayesian networks can be used for prediction tasks by determining the conditional probability of any given node in the graph given an observed set of other nodes in the graph.

A Bayesian network model was used to capture the joint probability distribution of microbial species abundances and the level of DOC. The vector $\boldsymbol{x}_i = [x_{i1}, x_{i2}, ..., x_{im}]$ represents discretized abundances of m microbial genera in the i^{th} training sample where each

$x_{ij} \in \{\text{Low}, \text{Moderate}, \text{High}\}$, and y_i denotes the abundance of DOC for the i^{th} sample, where $y_i \in \{\text{Low}, \text{High}\}$. Bin sizes for features x_{ij} were always based solely on the training data. The joint probability of microbial species abundance and DOC, $P(\boldsymbol{x}, y)$, can be represented as a graph, $\mathcal{G}$, that describes a factorization of the joint distribution as a set of conditional probability statements.

4.3.4 Bayesian network structure search and feature selection

The structure learning task aims to find a graphical structure that optimizes the log probability of the data given the model while simultaneously penalizing model complexity. Unfortunately, the computational effort for structure learning of Bayesian networks scales super-exponentially in the number of variables [8], making this impractical for microbiome data that contains hundreds of taxonomic features. For this reason, RFINN [119] was applied to the training data to reduce the feature set. An exhaustive structure search over the reduced set of variables could still be highly computationally expensive [128], but for the purpose of this study, we wish only to determine the minimum set of variables needed by the Bayesian network to accurately predict DOC for the system and provide an estimate of uncertainty. The Markov blanket of the DOC node defines a local Bayesian network with a reduced feature set that is equivalent to the global Bayesian network in its ability to predict DOC [8]. To determine the Markov blanket for DOC, the parent and child nodes of DOC were found by performing an exhaustive structure search over subsets of candidate variables in the total variable set. Directly connected nodes to DOC were kept, while indirectly connected nodes were discarded until the entire feature set was searched. In a second step, all discarded variables were checked as potential parents of child nodes to DOC. This procedure for determining the Markov blanket of a target node is similar to the *max-min* algorithm [129], which also performs an initial search for child and parent nodes of the target variable, and then determines other parents of the target's child nodes. We used the POMEGRANATE [128] package in PYTHON to perform Bayesian network structure search and training algorithms.

4.3.5 Evaluating prediction performance

The Kolmogorov-Smirnov (K-S) two sample test was performed to provide an easily interpreted metric of prediction performance. The null hypothesis of the two sample K-S test is that the cumulative distributions of observations in each sample are equivalent. This statistic was applied to determine if high and low DOC predictions separated measured DOC into distinct groups; if the p value is statistically significant, then the Bayesian network makes predictions of DOC levels that are collectively statistically significant. The test statistic is defined as

$$D = \max_x \left| \hat{F}_1(x) - \hat{F}_2(x) \right|, \tag{4.1}$$

where $\hat{F}_i$ is the empirical cumulative distribution function of the i^{th} sample. The test statistic, D, ranges from 0 to 1 and represents the maximum distance between two cumulative distributions. The p value of the K-S test represents the probability of observing the test statistic, D, if the two samples were collected from identical distributions [61]. The two sample K-S test statistic and p value were calculated using the default settings of SCIPY's KS_2SAMP function [130].

4.4 Results

For each of five training and testing permutations, we reduced the feature set using RFINN and then performed a Markov blanket search of the DOC node to construct a local Bayesian network model that predicts high or low DOC. Model training and validation with held-out data is summarized in figure 4.1. Figure 4.2 shows the receiving operating characteristic (ROC) curve representing prediction performance on held-out testing data for each permutation. In this context, the area under the ROC curve (AUC) represents the probability of ranking a randomly selected high DOC sample higher than a randomly selected low DOC sample, where the rank is the predicted probability, p(DOC=High) [131]. The AUC ranged from 0.78 to 0.92, indicating the predictions were accurate for all testing replicas. We chose the permutation with the median AUC score of .86 (permutation 1) as a representative for subsequent analyses.

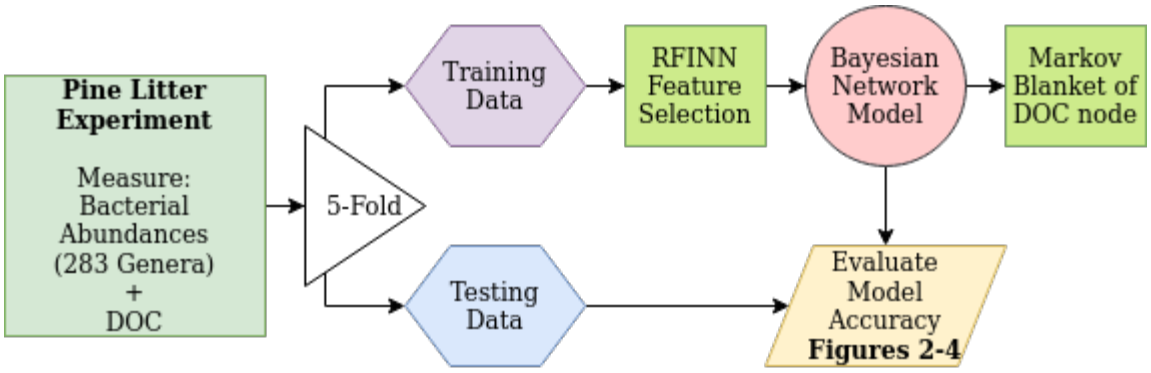

Figure 4.1: Process flow chart of feature selection, training, and validation with held-out data. The pine litter data set is partitioned into 5 permutations of training and testing data. RFINN is applied to each permutation of training data to reduce the feature set. Using the reduced feature set, a Bayesian network structure search algorithm is performed to identify and train a local Bayesian network composed of the Markov blanket of the DOC node. Prediction accuracy of the Bayesian network model is validated with held-out testing data.

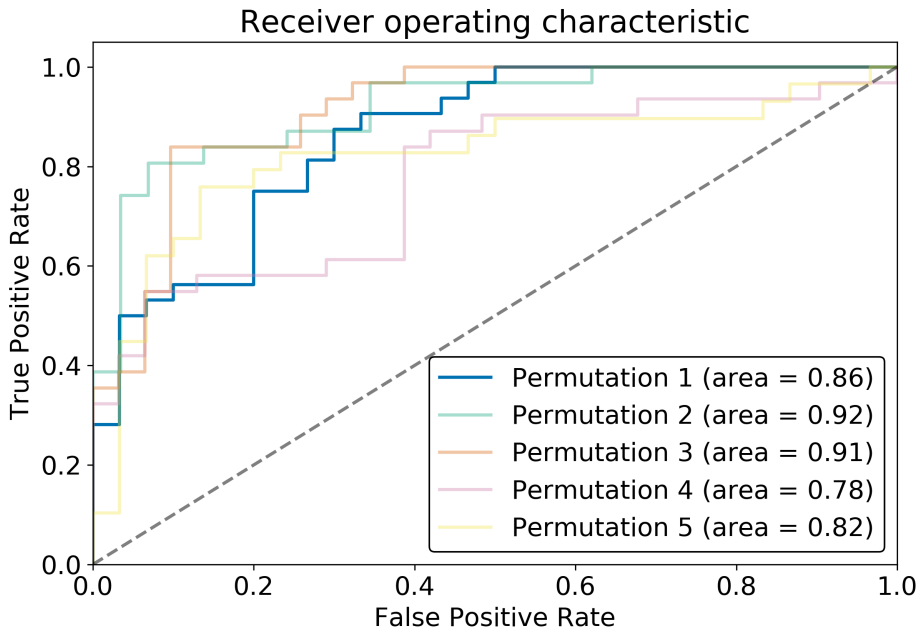

Figure 4.2: Receiver operating characteristic curves of test data sets. The full data set was partitioned into 5 unique sets of training and testing data such that the samples in each set of held-out testing data were a unique set. The area under the ROC curve for each permutation ranged from .78 to .92 with a median of .86 corresponding to permutation 1.

We show a confusion matrix in Figure 4.3A to visualize the predictive power of the Bayesian network model on held-out testing data, and the corresponding ROC curve for this permutation

is isolated in Figure 4.3B. The upper left section of the confusion matrix shows the number of accurately identified low DOC samples, or *true negatives*, and the central block of the confusion matrix shows the number of accurately identified high DOC samples, or *true positives*. The lower right block shows the accuracy, where accuracy is defined as the number of true positives and true negatives divided by the total number of evaluated samples. Additional metrics such as the positive predictive value, which is the probability of a true positive, are shown in the remaining blocks.

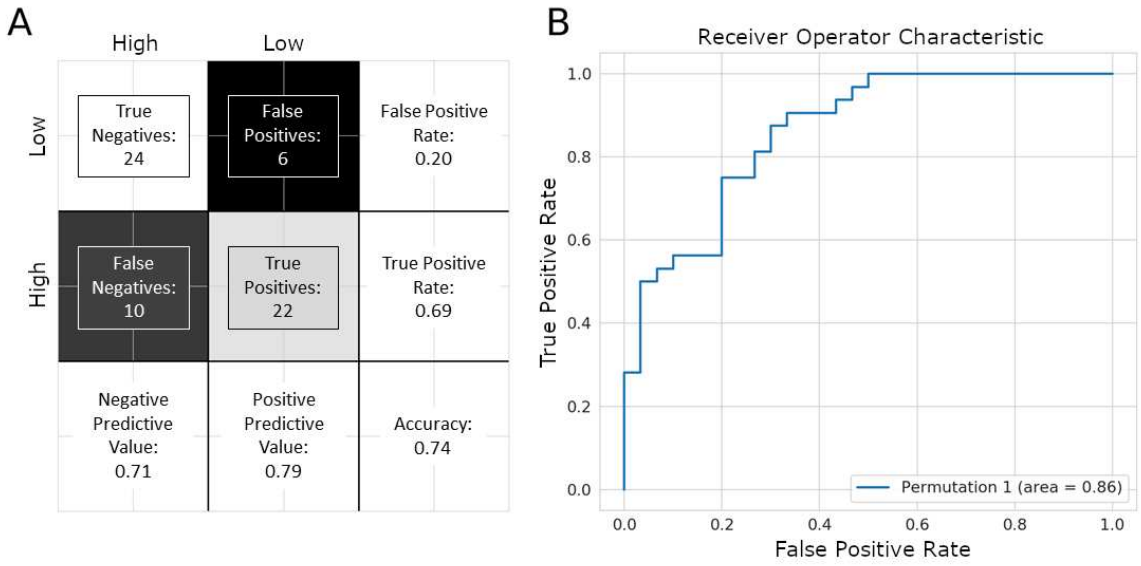

Figure 4.3: Confusion matrix and receiver operator characteristic on held-out pine samples. (A) Accuracy on held-out testing data was .74, where accuracy is the sum of true positives and true negatives divided by the total number of testing samples. (B) The area under the receiver operating characteristic curve was .86.

Although the model was trained using categorical labels for DOC, figure 4.4A shows a scatter plot of continuous DOC versus the predicted probability of high DOC for held-out testing samples. We defined 0.8 and 0.2 cutoffs to distinguish between high confidence and low confidence predictions, where predicted probabilities greater than 0.8 indicate highly confident predictions that DOC is high, and predicted probabilities less than .2 indicate highly confident predictions that DOC is low. Samples predicted with high confidence to have high and low DOC are labeled orange and cyan, respectively. Predictions for which confidence is less certain are shown in magenta. Of the 62 total communities in the pine testing data, 14 predictions were discarded as uncertain, but of the remaining 48 communities that were classified as certain, 89.6% were correctly predicted. Figure 4.4B shows the cumulative probability distribution of measured DOC between samples predicted to have high DOC with high confidence, low DOC with high confidence, and DOC from low confidence samples. The Kolmogorov-Smirnov (K-S) distance between the cumulative probability distribution of measured DOC corresponding to samples predicted to have high DOC and samples predicted to have low DOC was .814, P value $<<$ .001.

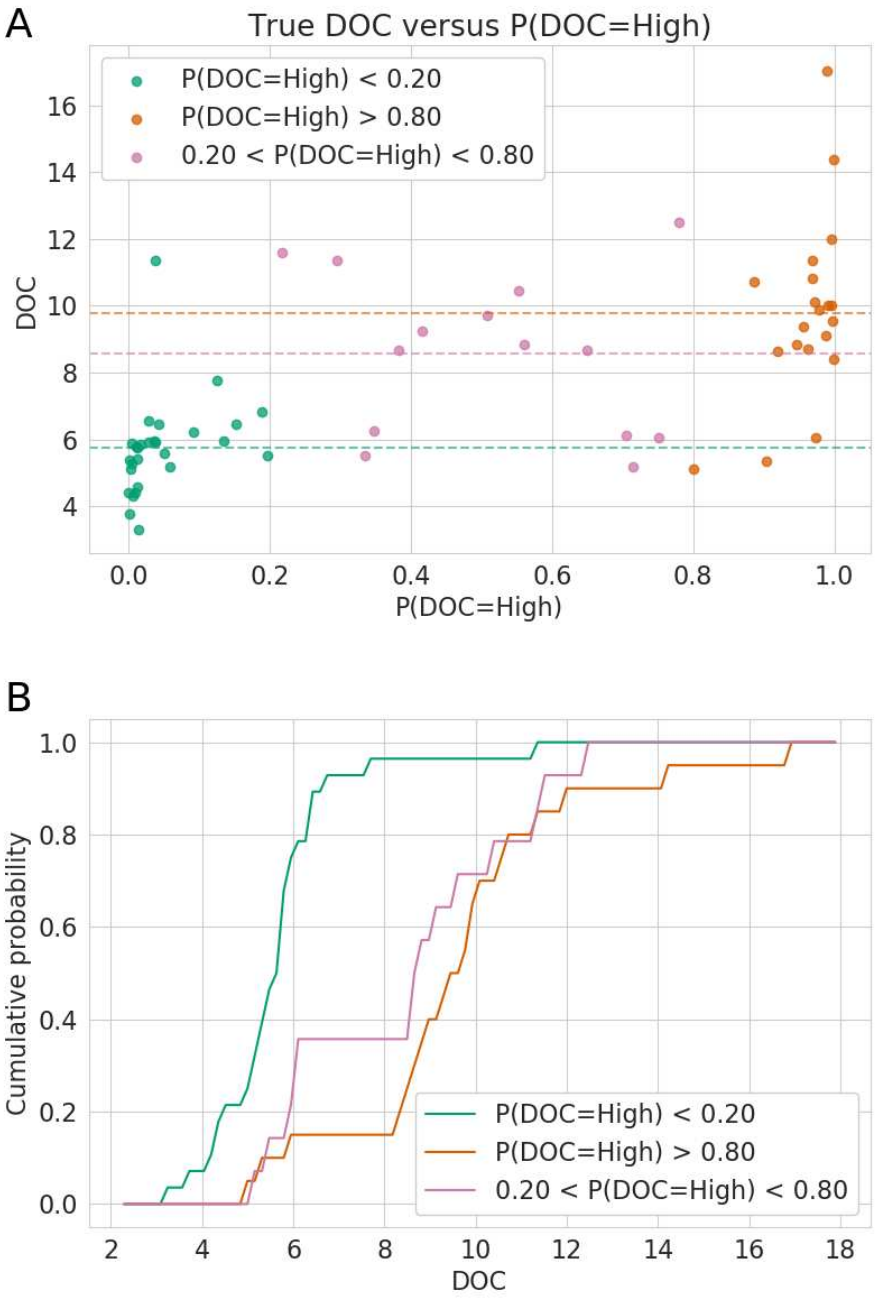

Figure 4.4: Measured DOC versus predicted $P(DOC = High)$. (A) Samples predicted with high confidence to have low DOC (P(DOC=High) $<$ .2) and high DOC (P(DOC=High) $>$ .8) are shown in cyan and orange, respectively. Predictions for which confidence is less certain are shown in magenta. Dotted horizontal lines represent the average measured DOC for each group. (B) Cumulative distribution plot of the set of measured DOC values in the held-out testing set binned into low, uncertain, and high DOC categories according to the predicted probability of high DOC. The Kolmogorov-Smirnov (K-S) distance between true DOC values of samples with a high predicted DOC and DOC values of samples with low predicted DOC was .814, P value $<<$ .001.

To investigate the transferability of the Bayesian network model, we next sought to apply a Bayesian network model trained on the pine data set to predict DOC outcomes of independent oak and grass litter data sets. A two-dimensional embedding of pine, oak, and grass samples using dimension reduction techniques such as principal component analysis (PCA) or t-distributed stochastic neighbor embedding (tSNE) can provide a visualization of the similarity of samples from each litter type. We applied tSNE using SCIKIT-LEARN [59] to vizualize a qualitative degree of separation between pine, oak, and grass litter samples (Figure 4.5). Using an analogous procedure to the cross-validation analyses, RFINN was first applied to the entire pine data set to determine a reduced feature set of 33 genera. Using these 33 features, a Bayesian network structure search algorithm identified the Markov blanket of the DOC node resulting in a final set of 10 genera. The 10 feature Bayesian network model trained using the pine data set was applied to predict DOC of oak and grass samples. Model training with pine data and validation with oak and grass data is summarized in figure 4.6.

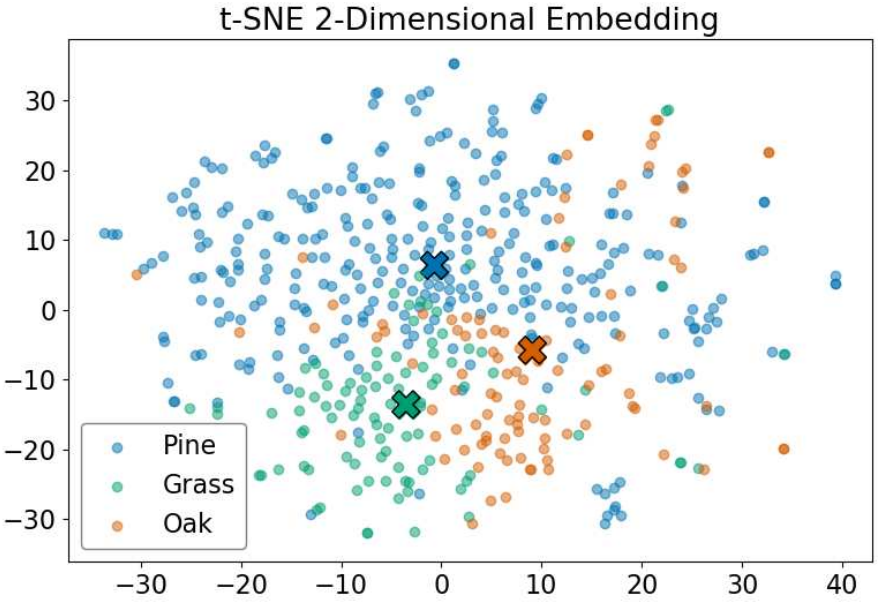

Figure 4.5: tSNE two-dimensional embedding of pine, oak, and grass samples. t-distributed Stochastic Neighbor Embedding (tSNE) was used to find a two-dimensional embedding of the high dimensional feature space of pine, oak, and grass samples. The lower dimensional embedding shows a clear qualitative separation between litter types, with X marking the average representation of each litter type.

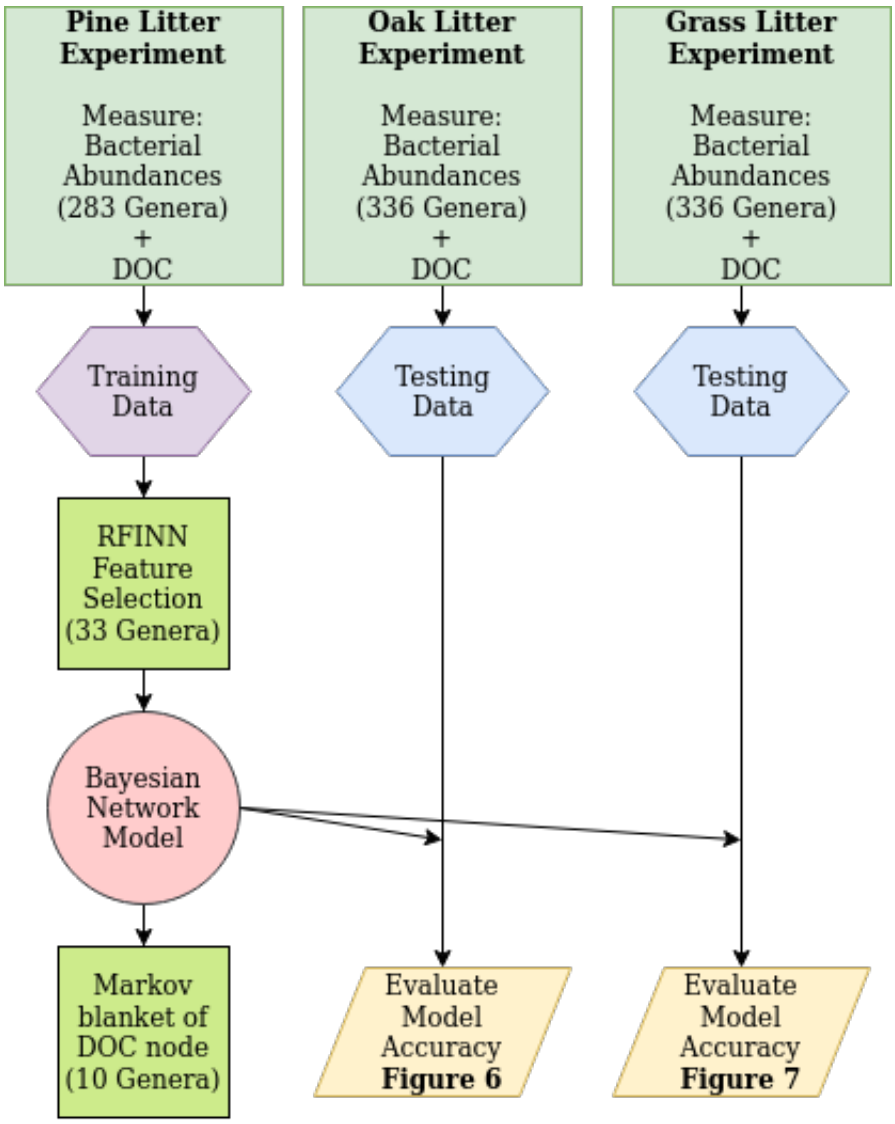

Figure 4.6: Process flow chart of feature selection, training, and validation with oak and grass data. RFINN is applied to the entire pine data set for feature reduction. Using the reduced feature set, a Bayesian network structure search algorithm is performed to identify and train a local Bayesian network composed of the Markov blanket of the DOC node. Prediction accuracy of the Bayesian network model is validated with oak and grass data.

The 10 feature Bayesian network model was found to be *Markov equivalent* to a Naive Bayes model. Markov equivalent structures satisfy equivalent independence assumptions despite having different edge directions [132]. Figure 4.7 shows the network structure of the Bayesian network model that links DOC with abundance of bacterial genera. Arrow directions illustrate the independence assumptions made by the model and do not necessarily suggest that variation in DOC

abundance drives variation in the abundance of the genera in the model. The independence assumptions made by this model (and, equivalently, a Naive Bayes model) is that abundance of each of the bacterial genera in the model are conditionally independent given the target node [8]. In other words, the structure search algorithm selected a model where the abundance of each genus is considered an independent random variable under high or low DOC conditions.

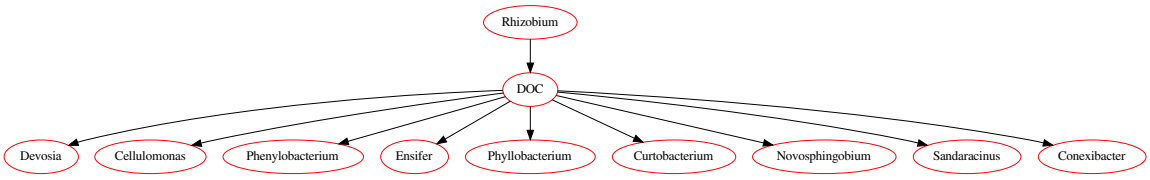

Figure 4.7: Bayesian network model of DOC using abundance of genera. The Bayesian network model composed of the Markov blanket of the DOC node is Markov equivalent to a Naive Bayes model. The independence properties of model specify that features are conditionally independent given the target (DOC) node.

Figure 4.8A shows a scatter plot of measured DOC from the 99 held-out oak litter data samples versus the predicted probability of high DOC from the 10 feature Bayesian network model trained on pine data. Using the same confidence cutoff as before, predictions are categorized into highly confident high DOC predictions (orange), highly confident low DOC (cyan), and low confidence predictions (magenta). Of the 99 total communities in the oak data, 37 predictions were discarded as uncertain, but of the remaining 62 communities that were classified as certain, 88.71% were correctly predicted to be either below or above the mean oak DOC. Figure 4.8B compares the distribution of measured DOC corresponding to high, low, and uncertain samples, with a K-S distance of .82 between high and low DOC samples, p value $<<$ .001.

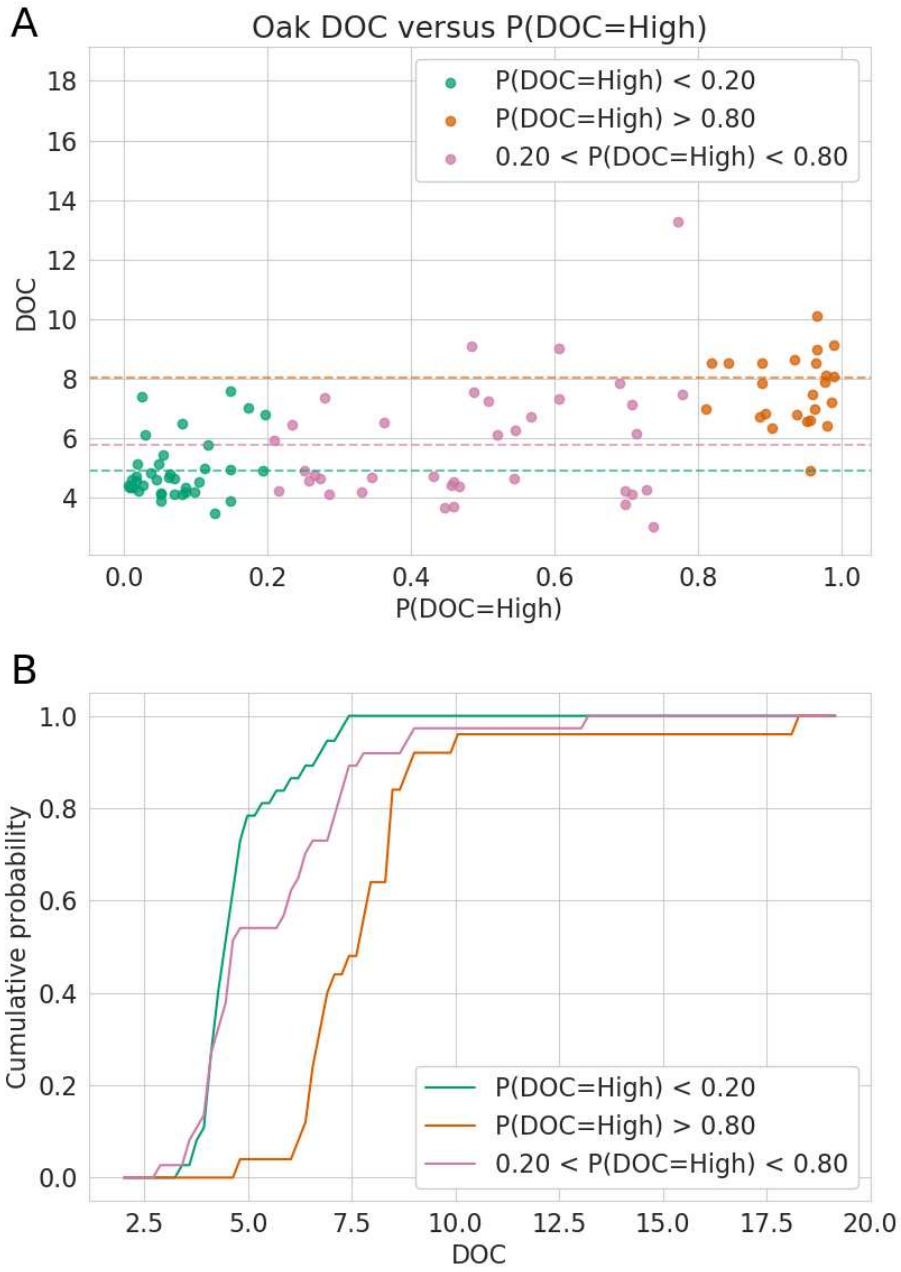

Figure 4.8: Oak DOC versus predicted $P(DOC = High)$. (A) Samples predicted with high confidence to have low DOC (P(DOC=High) $<$ .2) and high DOC (P(DOC=High) $>$ .8) are shown in cyan and orange, respectively. Predictions for which confidence is less certain are shown in magenta. Dotted horizontal lines represent the average measured DOC for each group.(B) The set of measured DOC values in the oak data set binned into low, uncertain, and high DOC categories according to the predicted probability of high DOC using a Bayesian network model trained on pine samples. The Kolmogorov-Smirnov (K-S) distance between true DOC values of samples with a high predicted DOC and DOC values of samples with low predicted DOC was .82, p value $<<$ .001.

To further investigate the transferability of the Bayesian network model, we applied the same model trained on pine data to samples from a grass litter decomposition experiment (Figure 4.9A,B). Unlike with the oak data, the model was not able to predict DOC for grass litter samples with statistically significant accuracy (K-S distance = .44, p value = .078). Poor performance on the grass data set represents an important failure case, demonstrating that the model loses relevance when applied to disparate litter types. However, we note that the percentage of uncertain predictions in the grass litter samples (54%) was higher than that in the oak litter test (45%) and in the pine litter test (23%), suggesting that the analysis is correctly able to identify its own fidelity when applied to contexts outside of its original training circumstances.

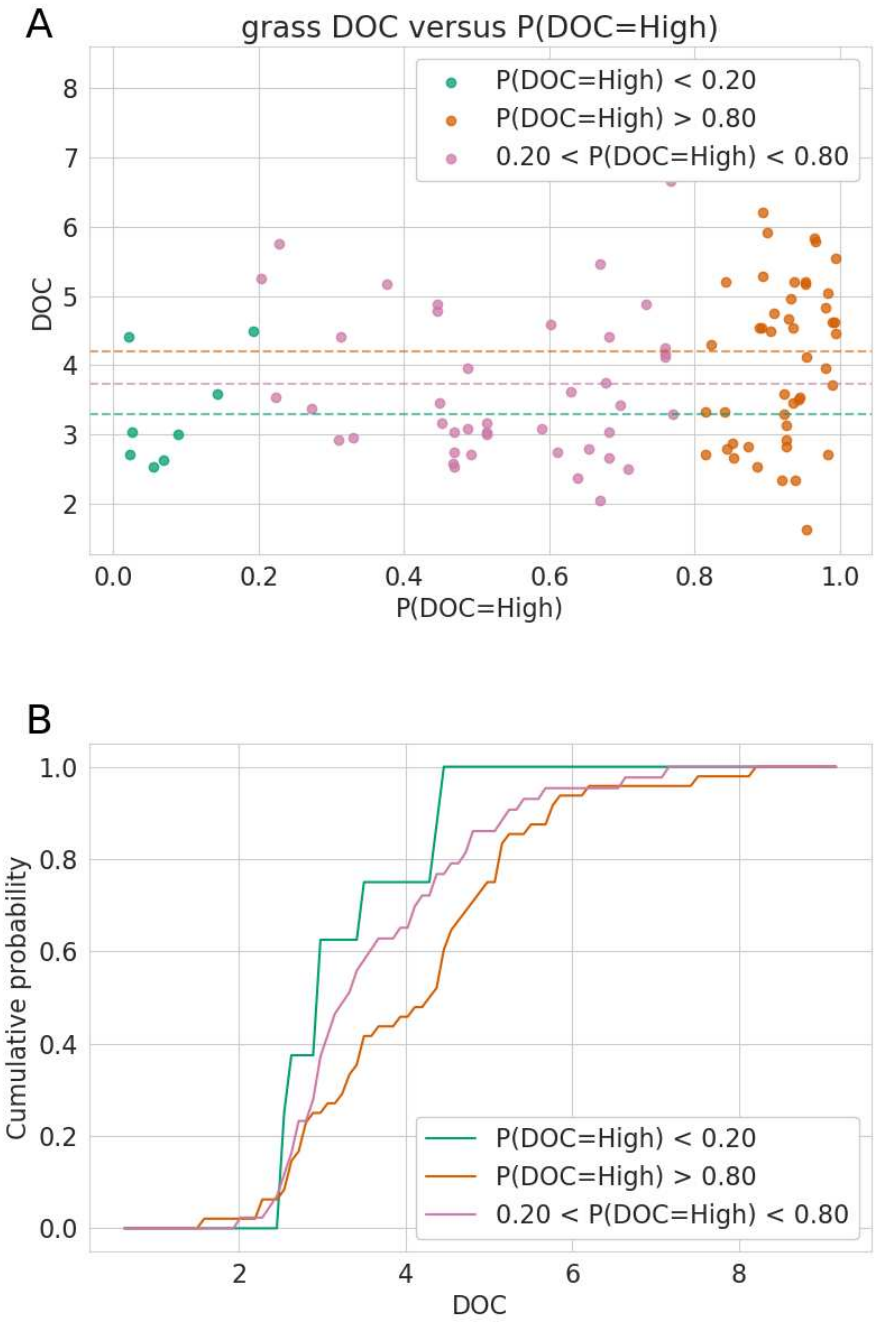

Figure 4.9: grass DOC versus predicted $P(DOC = High)$. (A) The average DOC of samples predicted to have low DOC (P(DOC=High) $<$ .2), high DOC (P(DOC=High) $>$ .8), and uncertain samples are shown as cyan, orange, and magenta dotted horizontal lines.(B) The set of measured DOC values in the grass data set binned into low, uncertain, and high DOC categories according to the predicted probability of high DOC using a Bayesian network model trained on pine samples. The Kolmogorov-Smirnov (K-S) distance between true DOC values of samples with a high predicted DOC and DOC values of samples with low predicted DOC was .44, p value $=$ .078.

4.5 Discussion

Rigorous feature selection and probabilistic predictions make Bayesian networks a powerful tool for modeling microbial communities. To our knowledge, this study presents the first example of a cross-study validated probabilistic graphical model approach that predicts microbiome function. We applied Bayesian networks to infer a model that captures the statistical interactions between relative microbial abundance and DOC in a pine litter decomposition experiment. By determining the Markov blanket of the DOC node in the resulting Bayesian network, we found a small set of specific bacterial genera whose abundance directly influence the probability of high or low DOC levels. In addition to standard cross-validation analysis with held-out data from the training set, we applied the Bayesian network model trained on pine data to predict DOC outcomes of two independent litter decomposition experiments.

We found that the Bayesian network model trained on samples from a pine litter decomposition experiment demonstrated significant prediction accuracy when tested on samples from an independent oak litter decomposition experiment. These strong predictions suggest that relationships between the abundances of key bacterial genera and dissolved organic carbon are conserved across pine and oak litter types. The observed overlap in correlation between community composition and collective functional state (i.e., high or low DOC) between pine and oak litter supports the hypothesis that common microbial drivers of DOC abundance can occur in diverse ecosystems despite the strong role of litter chemistry in shaping distinctive decomposer microbiomes [118].

Despite strong prediction accuracy when applied to samples from an oak litter decomposition experiment, the model failed to significantly distinguish high and low DOC samples from a grass litter decomposition experiment. The degree of this failure was predicted by the model in that uncertainty analysis showed that 54% of samples from the grass litter data set were classified as uncertain while 45% and 23% of samples from the oak and pine litter test sets were categorized as uncertain. This result suggests that the Bayesian network model correctly assigns greater prediction uncertainty in response to unfamiliar combinations of microbial abundances. Consequently, cross-validation with independent data sets not only demonstrated that the Bayesian network model

can successfully make accurate predictions when applied beyond its training circumstances, but that the model also assigns greater collective uncertainty to less accurate predictions. Models validated to make accurate predictions and accurate uncertainty estimates could be especially useful for designing microbial communities that optimize the probability of increased dissolved carbon in soil.

Engineering microbial communities to promote increased carbon fixation in soil will require accurate prediction models and a mechanistic understanding of how microbiomes influence soil carbon outcomes. While we have shown that the Bayesian network model accurately predicts DOC, other methods designed to reconstruct interaction networks could be explored in future work. PGM based approaches have recently been developed that offer a promising approach for comprehensive reconstruction of interaction networks, but they have not been bench-marked for their ability to make accurate predictions of target variables, as was the focus of this study. Network inference software such as SPIEC-EASI [122] or MERLIN [121] could be used to determine a more comprehensive network of interactions between microbial species. Designed to address the sparse and compositional nature of microbial abundance data, SPIEC-EASI has shown promise for inferring microbial interaction networks compared to frequently used [133] correlation-based techniques. MERLIN has proven more effective than Bayesian networks for determining interaction networks on synthetic gene expression data but has not been applied to model microbial communities. The combination of accurate prediction models with reconstructed interaction networks could guide strategic manipulation of microbiomes to promote increased levels of dissolved organic carbon.

We applied Bayesian network structure learning to identify microbial genera whose abundance directly influence DOC with an emphasis on prediction and feature selection. By searching for the Markov blanket of the DOC node in the graph, we identified a set of bacterial genera whose relative abundance directly influences the probability of high or low DOC. Furthermore, successful cross-validation with test data and with samples from independent studies provides strong evidence against false-discoveries in the model (over-fitting) [134]. While many studies using

metagenomics data have applied machine learning approaches to model microbial communities, validation of these models with data sets across studies is rare [7]. The ability of the Bayesian network model to apply concepts learned using pine litter decomposition samples and accurately predict DOC abundance of independent oak litter samples is compelling evidence to suggest that the identified taxa play a strong role in mediating soil carbon flow. In this targeted study of the relationships between microbial communities and DOC, the Bayesian network was trained only using bacterial abundances and DOC levels and was blind to any knowledge of litter type or litter chemistry. We envision future applications in which extended network models could consider a mixture of chemical and taxonomic features to improve further upon predictions and uncertainty quantification when extrapolated to disparate environments.

4.6 Funding statement

JT, RJ, JD and BM were supported by grant F255LANL2018 from the U.S. Department of Energy Office of Biological and Environmental Research, Genomic Science program. NL was supported by the U.S. Department of Energy through the Laboratory Directed Research and Development (LDRD) program at Los Alamos National Laboratory. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 5

Summary and Conclusions

5.1 Summary

Soil is a massive carbon pool that plays a major role in the natural carbon cycle. By directing carbon flow to favor storage in soil, levels of atmospheric carbon dioxide could reduce dramatically. With increasing evidence to suggest that microbial communities influence soil carbon sequestration, elucidating the mechanisms of this process could inform methods of improving microbially driven carbon storage. Furthermore, increasing ability to accurately profile microbial communities has led to an explosion of high dimensional sequencing data that motivates the need for improved data driven methods to link microbial community features with functions of interest. Through a progression of machine learning methods, we showed how significant feature reduction coupled with probabilistic modeling could be used to identify transparent and accurate models that link microbial community abundance with levels of dissolved organic carbon (DOC) in soil.

To study the role of the soil microbiome in plant litter decomposition, soil samples collected throughout the Western United States served as source material for inoculation of microbial communities into leaf litter microcosms. To evaluate carbon flow in the microcosms, DOC was quantified after a 44 day period. Ribosomal RNA gene profiles of the original and final microbial communities were taxonomically profiled, providing the relative abundance of microbial species in initial and final microcosm microbiomes. Each experiment yields on the order of hundreds of samples, where each sample includes the relative abundance of thousands of identified OTUs (operational taxonomic units). The disproportionate number of identified OTUs relative to the number of samples in each experiment poses a difficult problem from a data science perspective, and motivates the use of advanced computational tools to discover generalizable patterns in the data.

Using a simple logistic regression model, we showed that microbial community traits such as biomass and richness forecast DOC outcomes after a period of 44 days. The ability to forecast levels of DOC from original microbial traits suggests a causal relationship in which the soil microbiome directly influences soil carbon cycling. Also using a logistic regression model, the same microbial community traits measured at day 44 accurately predicted DOC. While the model using original microbial community features suggests that the soil microbiome influences the progression of carbon cycling in soil, accurate prediction accuracy of day 44 DOC outcomes using day 44 microbiome profiles supports a direct link between variation in instantaneous microbial community profiles with soil carbon cycling. The most significant features identified by the logistic regression model included biomass, fungal richness (when measured at day 0), and bacterial richness (when measured at day 44).

To further explore the direct link between the soil microbiome and carbon sequestration, we applied a host of machine learning approaches to predict DOC levels using the relative abundance of bacterial OTUs (operational taxonomic units). We showed that random forest and neural network regression accurately predict DOC on held-out testing data, with random forest performing better over bootstrapped permutations of the data. Feature ranking using machine learning approaches was compared to indicator species analysis to rank OTUs based on their importance in prediction tasks. The combination of several feature selection approaches provided a relatively small consensus set of taxa that resulted in equally accurate predictions compared to models using the entire feature set. We refer to this feature selection procedure as RFINN (Random Forest Indicator Species Neural Network).

By applying RFINN to reduce the set of bacterial genera, we were able to explore sophisticated probabilistic models that would otherwise be computationally intractable with the entire feature set. Probabilistic machine learning models offer a clear advantage over deterministic models such as random forest regression because they provide a metric of prediction uncertainty and therefore are capable of qualifying predictions with prediction confidence. Bayesian networks are a particular type of probabilistic machine learning model that offer a visual interpretation of how

the model links microbial taxa with DOC. We showed that confident predictions of DOC using a Bayesian network model trained on pine litter decomposition samples accurately separated high and low DOC samples in a held-out testing set. To investigate model transferability, we applied a Bayesian network trained using pine litter samples to predict DOC levels of samples derived from independent oak and grass litter decomposition experiments. Although the model could not distinguish high and low DOC outcomes of grass litter samples with statistical significance, the model achieved exceptional accuracy in distinguishing high and low DOC of oak samples. The ability of the model to predict levels of DOC across separate litter types suggests that correlations between carbon flow and microbial species are at least partially conserved across litter types.

5.2 Future Directions

In future studies, trained machine learning models could be used to identify hypothetical microbial communities that are predicted to promote increased levels of DOC based on trends in the training data. Uncertainty analysis using probabilistic models could guide the design of microbiomes optimized to promote high levels of DOC while avoiding improbable combinations of microbial taxa. Engineering the soil microbiome to match hypothetical communities proposed by machine learning models could be accomplished by either inoculating soil samples isolated microbial taxa, or by iteratively selecting and mixing soil samples whose individual microbiomes form optimized communities when combined. The ability to explore such approaches will require further research in understanding how to construct communities that exhibit stability as well as carbon sequestration capability.

The structure of Bayesian networks offers limited insight into the potential interactions that exist between microbial taxa. Other types of probabilistic graphical models have been developed to construct interaction networks using larger feature sets of microbial taxa. In future work, the exploration of such techniques could provide greater insight into microbial community interaction networks. Mechanistic insight coupled with accurate models that predict DOC outcomes could be a powerful combination of tools to inform the strategic manipulation of microbiomes to foster

increased carbon sequestration. Currently, the validation of models designed to discover interaction networks is limited due to the lack of ground truth. In addition to exploring other techniques to reconstruct interaction networks, future research could be applied to examine smaller microbial communities in which ground truth interaction networks have been proposed.

Another possible direction forward would be to study the individual taxa that were identified as significantly correlated with increased levels of DOC by machine learning methods. Current approaches such as genome-scale metabolic reconstruction provide a means for modeling the metabolic activity of microorganisms based on their genome. Such models could be used to identify metabolic pathways linked with the conversion of plant litter into forms of carbon that remain stable in the soil carbon pool. These models can then be used to identify rate limiting steps in metabolic pathways, thus providing strategies to increase rates of reactions that promote the degradation of soil organic matter into soluble forms of carbon.

5.3 Conclusions

The work in this thesis demonstrates how machine learning methods can be used to discover patterns in microbial communities. An important application of this work is the discovery of models that link the soil microbiome with carbon sequestration. With the potential to significantly reduce atmospheric carbon levels, microbially mediated carbon sequestration could be a viable path towards combating climate change. By training and validating machine learning models that predict levels of dissolved organic carbon using microbial community profiles, we show that there exist consistent patterns linking the soil microbiome with carbon flow. Insights from such models identify key taxa that are significantly linked with abundance of dissolved organic carbon produced during early litter decomposition. Developing accurate models of microbially driven carbon sequestration is an important step towards optimizing the soil microbiome to increase carbon storage in soil.

Bibliography

[1] Dieter Luthi, Martine Floch, Bernhard Bereiter, Thomas Blunier, Jean-Marc Barnola, Urs Siegenthaler, Dominique Raynaud, Jean Jouzel, Hubertus Fischer, Kenji Kawamura, and Thomas Stocker. High-resolution carbon dioxide concentration record 650,000-800,000 years before present. *Nature*, 453:379–82, 06 2008.

[2] Harry Cikanek, Ned Cyr, Ming Ji, Gary C. Matlock, and Steve Thur. *NOAA Chief Scientist's Annual Report : 2018*, chapter 2018 NOAA Science Report. NOAA Technical Memorandum NOAA Research Council ; 001. Silver Spring, MD, 2019. Technical Memorandum.

[3] Christos Gougoulias, Joanna Clark, and Liz Shaw. The role of soil microbes in the global carbon cycle: Tracking the below-ground microbial processing of plant-derived carbon for manipulating carbon dynamics in agricultural systems. *Journal of the science of food and agriculture*, 94, 09 2014.

[4] R. Lal. Soil carbon sequestration impacts on global climate change and food security. *Science*, 304(5677):1623–1627, 2004.

[5] Joshua Schimel and Sean Schaeffer. Microbial control over carbon cycling in soil. *Frontiers in Microbiology*, 3:348, 2012.

[6] Alexander Statnikov, Mikael Henaff, Varun Narendra, Kranti Konganti, Zhiguo Li, Liying Yang, Zhiheng Pei, Martin Blaser, Constantin Aliferis, and Alexander Alekseyenko. A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome*, 1, 04 2013.

[7] Edoardo Pasolli, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata. Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. *PLOS Computational Biology*, 12(7):1–26, 07 2016.

[8] Daphne Koller and Nir Friedman. Probabilistic graphical models: Principles and techniques (adaptive computation and machine learning series). *Foundations. The MIT Press*, 2009.

[9] Cindy Prescott. Litter decomposition: What controls it and how can we alter it to sequester more carbon in forest soils? *Biogeochemistry*, 101:133–149, 12 2010.

[10] William Wieder, Stuart Grandy, Cynthia Kallenbach, P. Taylor, and G. Bonan. Representing life in the earth system with soil microbial functional traits in the mimics model. *Geoscientific Model Development Discussions*, 8:2011–2052, 06 2015.

[11] William Wieder, Melannie Hartman, Benjamin Sulman, Yingping Wang, Charles Koven, and Gordon Bonan. Carbon cycle confidence and uncertainty: Exploring variation among soil biogeochemical models. *Global Change Biology*, 24, 11 2017.

[12] Pierre Friedlingstein, Malte Meinshausen, Vivek Arora, Chris Jones, Alessandro Anav, Spencer Liddicoat, and Reto Knutti. Uncertainties in cmip5 climate projections due to carbon cycle feedbacks. *Journal of Climate*, 27, 01 2014.

[13] Kefeng Wang, Changhui Peng, Qiuan Zhu, Xiaolu Zhou, Meng Wang, Kerou Zhang, and Gangsheng Wang. Modeling global soil carbon and soil microbial carbon by integrating microbial processes into the ecosystem process model triplex-ghg. *Journal of Advances in Modeling Earth Systems*, 9, 09 2017.

[14] William Wieder, Gordon Bonan, and Steven Allison. Global soil carbon predictions are improved by modeling microbial processes. *Nature Climate Change*, 3:909–912, 07 2013.

[15] Katherine Todd-Brown, J. Randerson, W. Post, Forrest Hoffman, C. Tarnocai, Edward Schuur, and S. Allison. Causes of variation in soil carbon simulations from cmip5 earth system models and comparison with observations. *Biogeosciences*, 10, 03 2013.

[16] Joshua Schimel and Sean Schaeffer. Microbial control over carbon cycling in soil. *Frontiers in microbiology*, 3:348, 09 2012.

[17] Dominic Woolf and Johannes Lehmann. Microbial models with minimal mineral protection can explain long-term soil organic carbon persistence. *Scientific Reports*, 9, 12 2019.

[18] Ashish Malik, Jennifer Martiny, Eoin Brodie, Adam Martiny, Kathleen Treseder, and Steven Allison. Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. *The ISME Journal*, 14:1–9, 09 2019.

[19] Raven Bier, Emily Bernhardt, Claudia Boot, Emily Graham, Edward Hall, Jay Lennon, Diana Nemergut, Brooke Osborne, Clara Ruiz-González, Joshua Schimel, Mark Waldrop, and Matthew Wallenstein. Linking microbial community structure and microbial processes: An empirical and conceptual overview. *FEMS Microbiology Ecology*, 91, 09 2015.

[20] Jacques Ravel, Pawel Gajer, Zaid Abdo, G. M. Schneider, Sara Koenig, Stacey McCulle, Shara Karlebach, Reshma Gorle, Jennifer Russell, Carol Tacket, Rebecca Brotman, Catherine Davis, Kevin Ault, Ligia Peralta, and Larry Forney. Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences of the United States of America*, 108 Suppl 1:4680–7, 03 2011.

[21] Gwen Falony, Marie Joossens, Sara Vieira-Silva, Jun Wang, Youssef Darzi, Karoline Faust, Alexander Kurilshikov, Marc Jan Bonder, Mireia Valles-Colomer, Doris Vandeputte, Raul Tito, Samuel Chaffron, Leen Rymenans, Chloë Verspecht, Lise Sutter, Gipsi Lima-Mendez, Kevin Dhoe, Karl Jonckheere, Daniel Homola, and Jeroen Raes. Population-level analysis of gut microbiome variation. *Science*, 352:560–564, 04 2016.

[22] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel Mende, Gabriel Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, Marcelo Bertalan, Natalia Borruel, Francesc Casellas, Leyden Fernández, Laurent Gautier, Torben Hansen, Masahira Hattori, Tetsuya Hayashi, Michiel Kleerebezem, and Peer Bork. Enterotypes of the human gut microbiome. *Nature*, 473:174–80, 05 2011.

[23] Raul Ochoa-Hueso. Global change and the soil microbiome: A human-health perspective. *Frontiers in Ecology and Evolution*, 5, 07 2017.

[24] Bonnie Waring, Colin Averill, and Christine Hawkes. Differences in fungal and bacterial physiology alter soil carbon and nitrogen cycling: Insights from meta-analysis and theoretical models. *Ecology letters*, 16, 05 2013.

[25] Benjamin Louis, Pierre-Alain Maron, Valérie Viaud, Philippe Leterme, and Safya Menasseri-Aubry. Soil c and n models that integrate microbial diversity. *Environmental Chemistry Letters*, 14, 07 2016.

[26] Uffe Nielsen, Edward Ayres, Diana Wall, and R.D. Bardgett. Soil biodiversity and carbon cycling: A review and synthesis of studies examining diversity-function relationships. *European Journal of Soil Science*, 62, 02 2011.

[27] M.-A Graaff, Jaron Adkins, Paul Kardol, and Heather Throop. A meta-analysis of soil biodiversity impacts on the carbon cycle. *SOIL*, 1:257–271, 03 2015.

[28] Brad Degens. Decreases in microbial function diversity do not result in corresponding changes in decomposition under different moisture conditions. *Soil Biology and Biochemistry*, 30, 12 1998.

[29] B. Griffiths, K Ritz, Ron Wheatley, H.L Kuan, Brian Boag, Søren Christensen, F Ekelund, Søren Sørensen, S Muller, and J Bloem. An examination of the biodiversity-ecosystem function relationship in arable soil microbial communities. *Soil Biology and Biochemistry*, 33:1713–1722, 10 2001.

[30] Sophie Wertz, Valérie Degrange, James Prosser, Franck Poly, Claire Commeaux, Thomas Freitag, Nadine Guillaumaud, and X. Roux. Maintenance of soil functioning following erosion of microbial diversity. *Environmental microbiology*, 8:2162–9, 12 2006.

[31] B. Griffiths, K. Ritz, R.D. Bardgett, R. Cook, Søren Christensen, F. Ekelund, Søren Sørensen, E. Bccth, J. Bloem, Ruiter, Jan Dolfing, and Bernard Nicolardot. Ecosystem

response of pasture soil communities to fumigation-induced microbial diversity reductions: An examination of the biodiversity-ecosystem function relationship. *Oikos 90 (2000), 2: 279-294*, 90, 08 2000.

[32] Sabrina Juarez, Naoise Nunan, Anne-Claire Duday, Valérie Pouteau, and Claire Chenu. Soil carbon mineralisation responses to alterations of microbial diversity and soil structure. *Biology and Fertility of Soils*, 49, 03 2013.

[33] Pierre-Alain Maron, Amadou Sarr, Aurore Kaisermann, Jean Lévêque, Olivier Mathieu, Julien Guigue, Battle Karimi, Laetitia Bernard, Samuel Dequiedt, Sebastien Terrat, Abad Chabbi, and Lionel Ranjard. High microbial diversity promotes soil ecosystem functioning. *Applied and environmental microbiology*, 84, 02 2018.

[34] Cameron Wagg, Klaus Schlaeppi, Samiran Banerjee, Eiko Kuramae, and Marcel Van der Heijden. Fungal-bacterial diversity and microbiome complexity predict ecosystem functioning. *Nature Communications*, 10:4841, 10 2019.

[35] William Schlesinger and Jeffrey Andrews. Soil respiration and the global carbon cycle. *Biogeochemistry*, 48:7–20, 01 2000.

[36] M. Francesca Cotrufo, Jennifer Soong, Andrew Horton, Eleanor Campbell, Michelle Haddix, Diana Wall, and William Parton. Formation of soil organic matter via biochemical and physical pathways of litter mass loss. *Nature Geoscience*, 8, 09 2015.

[37] Karolin Müller, Sven Marhan, Ellen Kandeler, and Christian Poll. Carbon flow from litter through soil microorganisms: From incorporation rates to mean residence times in bacteria and fungi. *Soil Biology and Biochemistry*, 115:187–196, 12 2017.

[38] Mauro Rubino, Jennifer Dungait, Richard Evershed, Teresa Bertolini, Paolo Angelis, Antonio D'Onofrio, Alessandra Lagomarsino, Carmine Lubritto, A. Merola, Filippo Terrasi, and M. Francesca Cotrufo. Carbon input belowground is the major c flux contributing to

leaf litter mass loss: Evidences from a 13c labelled-leaf litter experiment. *Soil Biology and Biochemistry*, 42:1009–1016, 07 2010.

[39] James Galloway, Frank Dentener, Elizabeth Boyer, R. Howarth, Sybil Seitzinger, Gregory Asner, Cory Cleveland, Pamela Green, Elisabeth Holland, D. Karl, Anthony Michaels, J. Porter, Alan Townsend, and C. Vöosmarty. Nitrogen cycles: Past, present, and future. *Biogeochemistry*, 70:153–226, 01 2004.

[40] Scott Bates, Donna Berg-Lyons, J Caporaso, William Walters, Rob Knight, and Noah Fierer. Examining the global distribution of dominant archaeal populations in soil. *The ISME journal*, 5:908–17, 11 2010.

[41] Rebecca Mueller, Jayne Belnap, and Cheryl Kuske. Soil bacterial and fungal community responses to nitrogen addition are constrained by microhabitat in an arid shrubland. *Frontiers in microbiology*, 6:891, 09 2015.

[42] Jennifer Talbot, Tom Bruns, John Taylor, Dylan Smith, Sara Branco, Sydney Glassman, Sonya Erlandson, Rytas Vilgalys, Hui-Ling Liao, Matthew Smith, and Kabir Peay. Endemism and functional convergence across the north american soil mycobiome. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 04 2014.

[43] Andrea Porras-Alfaro, Kuan-Liang Liu, Cheryl Kuske, and Gary Xie. From genus to phylum: Large-subunit and internal transcribed spacer rrna operon regions show similar classification accuracies influenced by database composition. *Applied and environmental microbiology*, 80, 11 2013.

[44] Gregory Gloor, Ruben Hummelen, Jean Macklaim, Russell Dickson, Andrew Fernandes, Roderick MacPhee, and Gregor Reid. Microbiome profiling by illumina sequencing of combinatorial sequence-tagged pcr products. *PloS one*, 5:e15406, 10 2010.

[45] Jiajie Zhang, Kassian Kobert, Tomá Flouri, and Alexandros Stamatakis. Pear: a fast and accurate illumina paired-end read merger. *Bioinformatics (Oxford, England)*, 30, 10 2013.

[46] J Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic Bushman, Elizabeth Costello, Noah Fierer, Antonio Peña, Julia Goodrich, Jeffrey Gordon, Gavin Huttley, Scott Kelley, Dan Knights, Jeremy Koenig, Ruth Ley, Catherine Lozupone, Daniel Mcdonald, Brian Muegge, Meg Pirrung, and Rob Knight. Caparose jgkj, stombaugh j, bittinger k, bushman fd.. qiime allows analysis of high-throughput community sequencing data. nat met 7: 335-336. *Nature methods*, 7:335–6, 04 2010.

[47] Robert Edgar. Uparse: Highly accurate otu sequences from microbial amplicon reads. *Nature methods*, 10, 08 2013.

[48] Edgar R. Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*, 27:2194–2200, 01 2011.

[49] Qiong Wang, George Garrity, James Tiedje, and J.R. Cole. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and Environment Microbiology*, 73:5264–5267, 01 2007.

[50] James Borneman and R. Hartin. Borneman j , hartin r. pcr primers that amplify fungal rrna genes from environmental samples. applied and environmental microbiology. *Applied and environmental microbiology*, 66:4356–60, 11 2000.

[51] D.J. Lane. 16s/23s rrna sequencing. nucleic acid techniques in bacterial systematics. *J.Wiley and Sons*, 01 1991.

[52] Gerard Muyzer, Ellen de Waal, and A.G. Uitterlinden. Profiling complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16s rrna. *Applied and environmental microbiology*, 59:695–700, 04 1993.

[53] Hector Castro, Aimée Classen, Emily Austin, Richard Norby, and Christopher Schadt. Soil microbial community responses to multiple experimental climate change drivers. *Applied and environmental microbiology*, 76:999–1007, 12 2009.

[54] Markus Kleber, Karin Eusterhues, Marco Keiluweit, Christian Mikutta, Robert Mikutta, and Peter Nico. Mineralorganic associations: Formation, properties, and relevance in soil environments. *Advances in Agronomy*, 130:1–140, 12 2015.

[55] Jari Oksanen, F. Guillaume Blanchet, Roeland Kindt, P Legendre, Peter Minchin, RB OHara, Gavin Simpson, Peter Solymos, MHH Stevens, and Helene Wagner. Vegan: community ecology package. r package version 2.0–7. h ttp. *cran. rproject. org/web/packages/vegan/index. html*, 01 2013.

[56] M. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26:32 – 46, 02 2001.

[57] Sarah Goslee and Dean Urban. The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, 22:1–19, 09 2007.

[58] Core Team. *R: A Language and Environment for Statistical Computing*. 01 2017.

[59] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Edouard Duchesnay, and Gilles Louppe. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 01 2012.

[60] Larry Wasserman. *A Concise Course in Statistical Inference*. 01 2005.

[61] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference (Springer Texts in Statistics)*. 12 2003.

[62] Petr Baldrian, Mark Bradford, Robert Warren, Thomas Crowther, Daniel Maynard, Emily Oldfield, William Wieder, Stephen Wood, and Joshua King. Climate fails to predict wood decomposition at regional scales. *Nature Climate Change*, 4:625–630, 06 2014.

[63] Mark Bradford, Ciska Veen, Anne Bonis, Ella Bradford, Aimee Classen, Johannes Cornelissen, Thomas Crowther, Jonathan De Long, Grégoire Freschet, Paul Kardol, Marta Manrubia, Daniel Maynard, Gregory Newman, Richard Logtestijn, Maria Viketoft, David Wardle, William Wieder, Stephen Wood, and Wim Putten. A test of the hierarchical model of litter decomposition. *Nature Ecology and Evolution*, 1, 12 2017.

[64] Michael Strickland, Christian Lauber, Noah Fierer, and Mark Bradford. Strickland ms, lauber c, fierer n, bradford ma.. testing the functional significance of microbial community composition. ecology 90: 441-451. *Ecology*, 90:441–51, 03 2009.

[65] Kristin Matulich and Jennifer Martiny. Microbial composition alters the response of litter decomposition to environmental change. *Ecology*, 96:154–163, 08 2015.

[66] Petr Baldrian. Forest microbiome: Diversity, complexity and dynamics. *FEMS microbiology reviews*, 41, 11 2016.

[67] Sydney Glassman, Claudia Weihe, Junhui Li, Michaeline Nelson Albright, Caitlin Looby, Adam Martiny, Kathleen Treseder, Steven Allison, and Jennifer Martiny. Decomposition responses to climate depend on microbial community composition. *Proceedings of the National Academy of Sciences*, 115:201811269, 11 2018.

[68] Michael Ryan and Beverly Law. Interpreting, measuring, and modeling soil respiration. *Biogeochemistry*, 73:3–27, 03 2005.

[69] Yun Gao, Xiang Gao, and Xiaohua Zhang. The 2 řc global temperature target and the evolution of the long-term goal of addressing climate changefrom the united nations framework convention on climate change to the paris agreement. *Engineering*, 3:272–278, 04 2017.

[70] Karsten Kalbitz and Klaus Kaiser. Contribution of dissolved organic matter to carbon storage in forest mineral soils. *Journal of Plant Nutrition and Soil Science*, 171:52–60, 02 2008.

[71] Klaus Kaiser and Karsten Kalbitz. Cycling downwardsdissolved organic matter in soils. *Soil Biology and Biochemistry*, 52:29–32, 09 2012.

[72] C. Newcomb, Nikolla Qafoku, J. Grate, Vanessa Bailey, and J. Yoreo. Developing a molecular picture of soil organic mattermineral interactions by quantifying organomineral binding. *Nature Communications*, 8, 12 2017.

[73] Cornelia Rumpel and Ingrid Kögel-Knabner. Deep soil organic matter-a key but poorly understood component of terrestrial c cycle. *Plant and Soil*, 338:143–158, 05 2011.

[74] Ingo Schöning and Ingrid Kögel-Knabner. Chemical composition of young and old carbon pools throughout cambisol and luvisol profiles under forests. *Soil Biology and Biochemistry*, 38:2411–2424, 08 2006.

[75] C Trigo and Andrew Ball. Is the solubilized product from the degradation of lignocellulose by actinomycetes a precursor of humic substances? microbiology. *Microbiology (Reading, England)*, 140 (Pt 11):3145–52, 12 1994.

[76] Christopher Fernandez and RT Koide. The role of chitin in the decomposition of ectomycorrhizal fungal litter. *Ecology*, 93:24–8, 01 2012.

[77] Ingrid Kögel-Knabner. The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter. *Soil Biology and Biochemistry*, 34:139–162, 02 2002.

[78] Cheta Siletti, Carolyn Zeiner, and Jennifer Bhatnagar. Distributions of fungal melanin across species and soils. *Soil Biology and Biochemistry*, 113:285–293, 10 2017.

[79] Jason Neff and Gregory Asner. Dissolved organic carbon in terrestrial ecosystems: Synthesis and a model. *Ecosystems*, 4:29–48, 01 2001.

[80] Vanessa Bailey, A. Peyton Smith, Malak Tfaily, S Fansler, and Ben Bond-Lamberty. Differences in soluble organic carbon chemistry in pore waters sampled from different pore size domains. *Soil Biology and Biochemistry*, 107:133–143, 04 2017.

[81] Sebastian Doetterl, Antoine Stevens, J. Six, Roel Merckx, Kristof Oost, Manuel Casanova, Angélica Casanova-Katny, Cristina Muñoz, Mathieu Boudin, Erick Zagal, and Pascal Boeckx. Soil carbon storage controlled by interactions between geochemistry and climate. *Nature Geoscience*, 8, 08 2015.

[82] Wenming Dong, Jiamin Wan, Tetsu Tokunaga, Benjamin Gilbert, and Kenneth Williams. Transport and humification of dissolved organic matter within a semi-arid floodplain. *Journal of Environmental Sciences*, 57, 12 2016.

[83] Fernando Maestre, Cristina Escolar, Mónica Ladrón de Guevara, Jose Quero, Roberto Lazaro, Manuel Delgado-Baquerizo, Victoria Ochoa, Miguel Berdugo, Beatriz Gozalo, and Antonio Gallardo. Changes in biocrust cover drive carbon cycle responses to climate change in drylands, 02 2015.

[84] Bo Tu, Xavier Domene, Minjie Yao, Chaonan Li, Shiheng ZHANG, Yongping Kou, Yansu Wang, and Xiangzhen Li. Microbial diversity in chinese temperate steppe: unveiling the most influential environmental drivers. *FEMS microbiology ecology*, 93, 03 2017.

[85] Mohammad Bahram, Falk Hildebrand, Sofia Forslund, Jennifer Anderson, Nadejda Soudzilovskaia, Peter Bodegom, Johan Bengtsson-Palme, Sten Anslan, Luis Pedro Coelho, Helery Harend, Jaime Huerta-Cepas, Marnix Medema, Mia Maltz, Sunil Mundra, Pal Olsson, Mari Pent, Sergei Põlme, Shinichi Sunagawa, Martin Ryberg, and Peer Bork. Structure and function of the global topsoil microbiome. *Nature*, 560, 08 2018.

[86] Pascale Frey-Klett, P Burlinson, Aurélie Deveau, Matthieu Barret, Mika Tarkka, and Alain Sarniguet. Bacterial-fungal interactions: Hyphens between agricultural, clinical, environ-

mental, and food microbiologists. *Microbiology and molecular biology reviews : MMBR*, 75:583–609, 12 2011.

[87] Renee (Liz) Sockett. Sockett re.. predatory lifestyle of bdellovibrio bacteriovorus. ann rev microbiol 63: 523-539. *Annual review of microbiology*, 63:523–39, 06 2009.

[88] Kurt Williamson, Jeffry Fuhrmann, K Eric Wommack, and Mark Radosevich. Viruses in soil ecosystems: An unknown quantity within an unexplored territory. *Annual Review of Virology*, 4:201–219, 09 2017.

[89] Daniel Laughlin. The intrinsic dimensionality of plant traits and its relevance to community assembly. *Journal of Ecology*, 102, 01 2013.

[90] Sheerli Shabat, Goor Sasson, Adi Faigenboim, Thomer Durman, Shamay Yaacoby, Margret Berg, Bryan White, Naama Shterzer, and Itzhak Mizrahi. Specific microbiome-dependent mechanisms underlie the energy harvest efficiency of ruminants. *The ISME journal*, 10, 05 2016.

[91] Emmanuelle Le Chatelier, Trine Nielsen, Junjie Qin, Edi Prifti, Falk Hildebrand, Gwen Falony, Mathieu Almeida, Manimozhiyan Arumugam, Jean-Michel Batto, Sean Kennedy, Pierre Leonard, Junhua Li, Kristoffer Burgdorf, Niels Grarup, Torben Jorgensen, Ivan Brandslund, Henrik Nielsen, Agnieszka Juncker, Marcelo Bertalan, and Takuji Yamada. Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500:541–6, 08 2013.

[92] Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 04 2015.

[93] Christopher M Bishop. *Pattern recognition and machine learning*. Springer Science+ Business Media, 2006.

[94] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 856–863, 2003.

[95] Peter E. Larsen, Dawn Field, and Jack A. Gilbert. Predicting bacterial community assemblages using an artificial neural network approach. *Nature Methods*, 9:621 EP –, Apr 2012. Article.

[96] Brandi L. Cantarel, Claire Fraser-Liggett, Elliott Franco Drábek, William Hsiao, and Zhenqiu Liu. Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics*, 27(23):3242–3249, 10 2011.

[97] Kevin Riehle, Cristian Coarfa, Andrew Jackson, Jun Ma, Arpit Tandon, Sameer Paithankar, Sriram Raghuraman, Toni-Ann Mistretta, Delphine Saulnier, Sabeen Raza, Maria Alejandra Diaz, Robert Shulman, Kjersti Aagaard, James Versalovic, and Aleksandar Milosavljevic. The genboree microbiome toolset and the analysis of 16s rrna microbial sequences. *BMC Bioinformatics*, 13(13):S11, Aug 2012.

[98] Nicholas Bokulich, Matthew Dillon, Evan Bolyen, Benjamin D Kaehler, Gavin A Huttley, and J Gregory Caporaso. q2-sample-classifier: machine-learning tools for microbiome classification and regression. *bioRxiv*, 2018.

[99] Peter A. Noble and Erik H. Tribou. Neuroet: An easy-to-use artificial neural network for ecological and biological modeling. *Ecological Modelling*, 203(1):87 – 98, 2007. Special Issue on Ecological Informatics: Biologically-Inspired Machine Learning.

[100] Ryan S King and Matthew Baker. *Use, Misuse, and Limitations of Threshold Indicator Taxa Analysis (TITAN) for Natural Resource Management*, pages 231–254. 02 2014.

[101] M Albright, R Johansen, J Thompson, D Lopez, L Gallegos-Gravesa, A Runde, R Mueller, A Washburne, B Munsky, T Yoshida, and J Dunbar. Fungal and bacterial richness forecast patterns of early pine litter decomposition. *The ISME Journal (Submitted)*, 2020.

[102] Marc Dufrene and Pierre Legendre. Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological monographs*, 67:345–366, 08 1997.

[103] Miquel De Caceres and Pierre Legendre. Associations between species and groups of sites: indices and statistical inference. *Ecology*, 90(12):3566–3574, 12 2009.

[104] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[105] Yifeng Li, Chih-Yu Chen, and Wyeth Wasserman. Deep feature selection: Theory and application to identify enhancers and promoters. 04 2015.

[106] N. Challita, M. Khalil, and P. Beauseroy. New feature selection method based on neural network and machine learning. In *2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET)*, pages 81–85, Nov 2016.

[107] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[108] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[109] Hunter R. Johnson, Donovan D. Trinidad, Stephania Guzman, Zenab Khan, James V. Parziale, Jennifer M. DeBruyn, and Nathan H. Lents. A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval. *PLOS ONE*, 11(12):1–23, 12 2016.

[110] Mistretta TA Saulnier DM, Riehle K. Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. *Gastroenterology*, 141, 11 2011.

[111] Hao-Xun Chang, James S. Haudenshield, Charles R. Bowen, and Glen L. Hartman. Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity. *Frontiers in Microbiology*, 8:519, 2017.

[112] Evan Bolyen, Jai Ram Rideout, Matthew R Dillon, Nicholas A Bokulich, Christian Abnet, Gabriel A Al-Ghalith, Harriet Alexander, Eric J Alm, Manimozhiyan Arumugam, Francesco Asnicar, Yang Bai, Jordan E Bisanz, Kyle Bittinger, Asker Brejnrod, Colin J Brislawn, C Titus Brown, Benjamin J Callahan, Andrés Mauricio Caraballo-Rodríguez, John Chase, Emily Cope, Ricardo Da Silva, Pieter C Dorrestein, Gavin M Douglas, Daniel M Durall, Claire Duvallet, Christian F Edwardson, Madeleine Ernst, Mehrbod Estaki, Jennifer Fouquier, Julia M Gauglitz, Deanna L Gibson, Antonio Gonzalez, Kestrel Gorlick, Jiarong Guo, Benjamin Hillmann, Susan Holmes, Hannes Holste, Curtis Huttenhower, Gavin Huttley, Stefan Janssen, Alan K Jarmusch, Lingjing Jiang, Benjamin Kaehler, Kyo Bin Kang, Christopher R Keefe, Paul Keim, Scott T Kelley, Dan Knights, Irina Koester, Tomasz Kosciolek, Jorden Kreps, Morgan GI Langille, Joslynn Lee, Ruth Ley, Yong-Xin Liu, Erikka Loftfield, Catherine Lozupone, Massoud Maher, Clarisse Marotz, Bryan D Martin, Daniel McDonald, Lauren J McIver, Alexey V Melnik, Jessica L Metcalf, Sydney C Morgan, Jamie Morton, Ahmad Turan Naimey, Jose A Navas-Molina, Louis Felix Nothias, Stephanie B Orchanian, Talima Pearson, Samuel L Peoples, Daniel Petras, Mary Lai Preuss, Elmar Pruesse, Lasse Buur Rasmussen, Adam Rivers, Michael S Robeson, II, Patrick Rosenthal, Nicola Segata, Michael Shaffer, Arron Shiffer, Rashmi Sinha, Se Jin Song, John R Spear, Austin D Swafford, Luke R Thompson, Pedro J Torres, Pauline Trinh, Anupriya Tripathi, Peter J Turnbaugh, Sabah Ul-Hasan, Justin JJ van der Hooft, Fernando Vargas, Yoshiki Vázquez-Baeza, Emily Vogtmann, Max von Hippel, William Walters, Yunhu Wan, Mingxun Wang, Jonathan Warren, Kyle C Weber, Chase HD Williamson, Amy D Willis, Zhenjiang Zech Xu, Jesse R Zaneveld, Yilong Zhang, Qiyun Zhu, Rob Knight, and J Gregory Caporaso. Qiime 2: Reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ Preprints*, 6:e27295v2, December 2018.

[113] Jacqueline M. Chaparro, Amy M. Sheflin, Daniel K. Manter, and Jorge M Vivanco. Manipulating the soil microbiome to increase soil health and plant fertility. *Biology and Fertility of Soils*, 48:489–499, 2012.

[114] Chao Liang, Joshua Schimel, and Julie Jastrow. The importance of anabolism in microbial control over soil carbon storage. *Nature Microbiology*, 2:17105, 07 2017.

[115] William Schlesinger and Jeffrey Andrews. Soil respiration and global carbon cycle. *Biogeochemistry*, 48:7–20, 01 2000.

[116] Cindy Prescott. Litter decomposition: What controls it and how can we alter it to sequester more carbon in forest soils? *Biogeochemistry*, 101:133–149, 12 2010.

[117] Luiz Roesch, Roberta Fulthorpe, Alberto Riva, George Casella, Alison Hadwin, Angela Kent, S. Daroub, Flavio Camargo, William Farmerie, and Eric Triplett. Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME journal*, 1:283–90, 09 2007.

[118] M. P. Krishna and Mahesh Mohan. Litter decomposition in forest ecosystems: a review. *Energy, Ecology and Environment*, 2(4):236–249, Aug 2017.

[119] Jaron Thompson, Renee Johansen, John Dunbar, and Brian Munsky. Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition. *PLOS ONE*, 14(7):1–16, 07 2019.

[120] Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, 2004.

[121] Sushmita Roy, Stephen Lagree, Zhonggang Hou, James A. Thomson, Ron Stewart, and Audrey P. Gasch. Integrated module and gene-specific regulatory inference implicates upstream signaling networks. *PLOS Computational Biology*, 9(10):1–20, 10 2013.

[122] Zachary D. Kurtz, Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLOS Computational Biology*, 11(5):1–25, 05 2015.

[123] Cory J. Butz, André E. dos Santos, Jhonatan S. Oliveira, and John Stavrinides. Efficient examination of soil bacteria using probabilistic graphical models. In Malek Mouhoub, Samira

Sadaoui, Otmane Ait Mohamed, and Moonis Ali, editors, *Recent Trends and Future Technology in Applied Intelligence*, pages 315–326, Cham, 2018. Springer International Publishing.

[124] Michael McGeachie, Joanne Sordillo, Travis Gibson, George Weinstock, Yang-Yu Liu, Diane Gold, Scott Weiss, and Augusto Litonjua. Longitudinal prediction of the infant gut microbiome with dynamic bayesian networks. *Scientific Reports*, 6:20359, 02 2016.

[125] Mehdi Layeghifard, David M. Hwang, and David S. Guttman. Disentangling interactions in the microbiome: A network perspective. *Trends in Microbiology*, 25(3):217 – 228, 2017.

[126] John Cook, Naomi Oreskes, Peter T Doran, William R L Anderegg, Bart Verheggen, Ed W Maibach, J Stuart Carlton, Stephan Lewandowsky, Andrew G Skuce, Sarah A Green, Dana Nuccitelli, Peter Jacobs, Mark Richardson, Bärbel Winkler, Rob Painting, and Ken Rice. Consensus on consensus: a synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters*, 11(4):048002, apr 2016.

[127] Mathias Neumann, Liisa Ukonmaanaho, James Johnson, Sue Benham, Lars Vesterdal, Radek Novotný, Arne Verstraeten, Lars Lundin, Anne Thimonier, Panagiotis Michopoulos, and Hubert Hasenauer. Quantifying carbon and nutrient input from litterfall in european forests using field observations and modeling. *Global Biogeochemical Cycles*, 32(5):784–798, 2018.

[128] Jacob Schreiber. Pomegranate: fast and flexible probabilistic modeling in python. *Journal of Machine Learning Research*, 18(164):1–6, 2018.

[129] Ioannis Tsamardinos, Constantin F. Aliferis, and Alexander Statnikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 673–678, New York, NY, USA, 2003. ACM.

[130] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy. Scipy 1.0-fundamental algorithms for scientific computing in python. *CoRR*, abs/1907.10121, 2019.

[131] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861 – 874, 2006. ROC Analysis in Pattern Recognition.

[132] Steen Andersson, David Madigan, and Michael Perlman. A characterization of markov equivalence classes for acyclic digraphs. *Annals of Statistics*, 25, 02 2000.

[133] Sophie Weiss, Will Treuren, Catherine Lozupone, Karoline Faust, Jonathan Friedman, Ye Deng, Li Xia, Zhenjiang Xu, Luke Ursell, Eric Alm, Amanda Birmingham, Jacob Cram, Jed Fuhrman, Jeroen Raes, Fengzhu Sun, Jizhong Zhou, and Rob Knight. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal*, 10, 02 2016.

[134] Christoph Bernau, Markus Riester, Anne-Laure Boulesteix, Giovanni Parmigiani, Curtis Huttenhower, Levi Waldron, and Lorenzo Trippa. Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, 30(12):i105–i112, 06 2014.

Appendix A

Supporting Information for Chapter 3

S1 Dataset OTU table. The bacteria OTU table used for all results in the paper organized with samples as rows and features as columns. https://doi.org/10.1371/journal.pone.0215502.s001

S2 Dataset Training data set. Partition OTU table used for training machine learning models to produce Figure 3.1. https://doi.org/10.1371/journal.pone.0215502.s002

S3 Dataset Testing data set. Partition OTU table used for testing machine learning models to produce Figure 3.1. https://doi.org/10.1371/journal.pone.0215502.s003

S4 Dataset Feature selection table. A table with feature importance values for the consensus set of taxa sorted by the indicator species statistic. https://doi.org/10.1371/journal.pone.0215502.s004