Modeling and sampling optimization of soil microbial biomass for soil health assessment in agricultural systems

by

Vydehi Gadiparthi

B.S., Acharya N.G. Ranga Agricultural University, 2023

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Agronomy
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2026

Approved by:

Major Professor
Charles W. Rice

Copyright

Abstract

Soil health assessment increasingly emphasizes biological indicators, yet their integration into routine monitoring remains limited by high analytical costs, spatial variability, and interpretive challenges. Soil microbial biomass, a key indicator of soil biological functioning, plays a central role in organic matter dynamics and nutrient cycling but is typically measured using resource-intensive methods such as phospholipid fatty acid (PLFA) analysis. This thesis addresses these limitations by integrating predictive modeling and sampling optimization to develop scalable approaches for biological soil health assessment.

Chapter 1 synthesizes current knowledge on soil health frameworks, emphasizing the importance of biological indicators, the influence of spatial variability, and the limitations of existing sampling and modeling approaches. It identifies a critical gap in linking biological indicators with efficient sampling strategies and predictive modeling.

Chapter 2 evaluates the extent to which soil microbial biomass can be predicted from routinely measured soil physicochemical properties using machine learning. Results show that soil properties, particularly organic matter, carbon, nitrogen, and texture, explain a substantial proportion of variability in microbial biomass, with XGBoost achieving strong predictive performance ($R^2 \approx 0.80$). The inclusion of enzyme activity provided only modest improvements, indicating that microbial biomass can be effectively estimated using commonly measured soil variables.

Chapter 3 investigates whether reduced-intensity sampling strategies can reliably capture field-scale variability in soil health indicators. Results demonstrate that reducing sampling density from one sample per acre to one per five acres preserved field-scale means and spatial structure for most indicators. Additionally, a six-point management zone-based sampling design

estimated microbial biomass with less than 2% bias, demonstrating that targeted sampling can substantially reduce sampling effort while maintaining accuracy.

Together, these findings establish a tiered framework for soil health assessment that integrates dense physicochemical sampling, predictive modeling, and targeted biological validation. This approach enables the incorporation of biological indicators into scalable soil monitoring systems, supporting precision agriculture and sustainable land management in spatially heterogeneous agricultural landscapes.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my sincere gratitude to my advisor, Dr. Charles W. Rice, and my committee members, Dr. Vara Prasad and Dr. Gaurav Jha, for their guidance and support throughout my graduate studies.

I am thankful to the Soil Microbial Agroecology Lab (SMAL), including my lab mates, colleagues, lab manager, and visiting scholars, for their collaboration and encouragement. I also extend my appreciation to the people who have been part of my life during my time at Kansas State University.

Finally, I acknowledge Kansas State University for providing the resources and environment that made this work possible.

Chapter 1 - Integrating Biological Indicators, Spatial Variability, and Predictive Modeling for Soil Health Assessment

1.1 Soil health and the role of biological indicators

Soil health is widely defined as the continued capacity of soil to function as a living system that sustains biological productivity, maintains environmental quality, and supports plant, animal, and human health (Doran and Zeiss, 2000; Rinot et al., 2019; Mann et al., 2019, USDA, 2026). The definition reflects a shift from viewing soil as a static medium for plant growth to recognizing it as a dynamic and biologically active system, where physical, chemical, and biological components interact to regulate ecosystem processes (Lehmann et al., 2020; Fierer et al., 2021).

Conventional soil health assessments have historically focused on physical and chemical indicators such as soil texture, bulk density, nutrient availability, and pH. These indicators are well-established and provide valuable information on soil structure and fertility; however, they often respond slowly to management changes and may not adequately capture short-term shifts in soil function (Cardoso et al., 2013; Bünemann et al., 2018). In contrast, biological indicators, including microbial biomass, enzyme activity, and soil respiration, are more sensitive to environmental disturbances and management practices, making them useful for detecting early changes in soil condition (Bünemann et al., 2018; Lehmann et al., 2020; Fierer et al., 2021). The increasing emphasis on biological indicators is largely driven by the recognition that soil microorganisms play a central role in regulating key ecosystem processes. Microbial communities drive nutrient cycling, organic matter decomposition, and energy flow, thereby

influencing plant productivity and ecosystem stability (Stockdale and Brookes, 2006; Van der Heijden, 2008; Decaëns, 2010). Through these processes, microorganisms act as mediators between soil organic matter and plant-available nutrients, linking biological activity with soil fertility and ecosystem functioning (Fierer et al., 2009; Lehmann et al., 2020).

Despite their importance, defining soil health using a universal set of indicators remains challenging. Soil properties and functions vary across ecosystems, climatic conditions, and management systems, and therefore, no single "optimal" soil condition can be defined (Bünemann et al., 2018; Lehmann et al., 2020; Fierer et al., 2021). As a result, soil health is better understood as a context-dependent and functional concept, where the relevance of specific indicators depends on the environmental setting and assessment objectives (Cardoso et al., 2013; Lehmann et al., 2020).

Biological indicators remain underrepresented in many soil health assessment frameworks and are often limited to a small number of measurements, such as microbial biomass or respiration, which are sometimes treated as "black-box" indicators without clear mechanistic interpretation (Horrigue et al., 2016; Bünemann et al., 2018; Fierer et al., 2021). This limitation is largely due to challenges associated with measurement complexity, interpretation, and cost, which restrict their integration into routine monitoring systems (Dalal, 1998; Semenov et al., 2025). Consequently, there is growing interest in developing approaches that can better integrate biological indicators into soil health evaluation, particularly in agricultural systems. Among the various biological indicators, soil microbial biomass has received particular attention due to its central role in regulating soil processes and its strong association with soil organic matter dynamics.

1.2 Soil microbial biomass as a soil health indicator

Soil microbial biomass (SMB) represents the living component of soil organic matter and is widely regarded as a key indicator of soil fertility and ecosystem functioning (Gregorich et al., 1994; Wardle, 1998; Singh and Gupta, 2018). Although it typically constitutes only a small fraction of total soil organic carbon and nitrogen, SMB plays a disproportionately important role in regulating nutrient cycling and energy flow within soil systems (Gregorich et al., 1994; Jenkinson et al., 1990).

Soil microbial biomass functions both as a source and a sink of nutrients, mediating the transformation of organic matter and controlling the availability of carbon and nitrogen to plants (Templer et al., 2003; Jiang-shan et al., 2005). Through processes such as decomposition, mineralization, and immobilization, soil microorganisms drive the cycling of essential nutrients, including nitrogen, phosphorus, and sulfur, linking soil organic matter dynamics with plant productivity (Stockdale and Brookes, 2006; Van der Heijden, 2008; Decaëns, 2010). As a result, microbial biomass is often considered a key component of the soil decomposer subsystem, regulating ecosystem processes and influencing long-term soil fertility (Wardle and Ghani, 1995; Wardle, 1998).

A strong relationship between SMB and soil organic carbon (SOC) has been consistently reported across a wide range of ecosystems, reflecting the role of SOC as a primary energy source for microbial growth (Fierer et al., 2009; Lehmann et al., 2020; Wan et al., 2021). This relationship highlights the importance of organic matter inputs and soil management practices in shaping microbial communities. Changes in land use, residue inputs, and disturbance regimes can significantly influence SMB by altering resource availability and habitat conditions (Sünnemann et al., 2021; Dongdong et al., 2023).

Because of its sensitivity to environmental change, SMB is often used as an early indicator of shifts in soil condition. It has been shown to respond rapidly to management practices such as tillage, crop rotation, and organic amendments, making it useful for detecting short-term changes that may not yet be reflected in physical or chemical soil properties (Bosatta and Ågren, 1994; Dalal, 1998; Singh and Gupta, 2018). This responsiveness makes microbial biomass particularly valuable in the context of sustainable and regenerative agriculture, where monitoring early changes in soil function is critical.

Despite its advantages, the interpretation of SMB as a soil health indicator is not always straightforward. Variability in microbial communities, differences in environmental conditions, and the influence of multiple interacting factors can complicate its use as a standalone metric (Wardle and Ghani, 1995; Fierer et al., 2009). Furthermore, SMB does not always show a direct or consistent relationship with crop productivity, highlighting the need to interpret it within a broader soil health framework (Dalal, 1998; Bünemann et al., 2018).

Overall, SMB provides an integrative measure of biological activity, while also highlighting the complexity of biological systems that must be considered in its interpretation.

1.3 Challenges in measuring biological indicators

Regardless of their importance in soil health assessment, biological indicators remain difficult to measure, interpret, and implement at operational scales. These challenges limit their integration into routine soil monitoring frameworks, particularly in agricultural systems where cost, time, and scalability are critical considerations (Bünemann et al., 2018; Wilhelm et al., 2022).

One of the primary limitations of biological indicators is the complexity and cost associated with their measurement. Methods such as phospholipid fatty acid (PLFA) analysis, enzyme assays, and molecular techniques require specialized protocols and significant laboratory expertise (Horrigue et al., 2016; Semenov et al., 2025). While these approaches provide detailed insights into microbial community structure and function, they are often time-consuming and expensive, making them impractical for large-scale or routine applications (Dalal, 1998; Wilhelm et al., 2022). As a result, biological measurements are typically conducted less frequently and at fewer sampling locations compared to conventional soil tests.

In addition to cost constraints, biological indicators exhibit high spatial and temporal variability, which complicates their interpretation (Peigné et al., 2009; Gyawali et al., 2023). Microbial biomass and activity are influenced by a range of dynamic factors, including soil moisture, temperature, substrate availability, and recent management practices (Wardle and Ghani, 1995; Nunan et al., 2002). Enzyme activities, in particular, can fluctuate rapidly in response to short-term environmental changes and may persist independently of microbial biomass due to stabilization mechanisms in soil (Tabatabai, 1994; Saiya-Cork et al., 2002). These limitations highlight the need for standardized methodologies and careful interpretation of molecular data in soil health applications.

Beyond methodological challenges, scaling biological measurements from point observations to the field or regional scale remains a significant constraint. Soil systems are inherently heterogeneous, and biological properties often exhibit fine-scale spatial variability that is difficult to capture with limited sampling (Rossel, 2007; Welsch et al., 2019). Collecting sufficiently dense datasets to characterize this variability is often cost-prohibitive, leading to uncertainty in estimating field-scale biological properties. This challenge is further compounded

by the lack of established benchmarks and reference values for many biological indicators, which limits their interpretability in practical decision-making contexts (Dalal, 1998; Bünemann et al., 2018).

Together, these limitations highlight a fundamental trade-off between the ecological relevance of biological indicators and their practical applicability in soil health monitoring. While biological measurements provide valuable insights into soil processes, their cost, variability, and scaling challenges restrict their widespread use. In addition to these methodological and logistical constraints, the inherently high spatial variability of soil properties further complicates the measurement and interpretation of biological indicators, particularly when sampling intensity is limited. These challenges are further compounded by the inherent spatial variability of soil systems.

1.4 Spatial variability of soil properties

Soil systems are inherently heterogeneous, with properties varying across space due to a combination of soil-forming factors, environmental conditions, and management practices. This spatial variability is a fundamental characteristic of soils and plays a critical role in determining soil function, nutrient availability, and crop productivity (Wilding and Drees, 1983; Cambardella et al., 1994; Webster and Oliver, 2007). As a result, understanding and quantifying spatial variability is essential for accurate soil characterization and effective land management. In practice, the complexity of soil variability, particularly in managed agricultural systems, presents significant challenges for both measurement and interpretation.

Spatial variability in soils arises from both systematic and random components. Systematic variability reflects structured changes driven by soil-forming factors such as

topography, parent material, climate, and land use, while random variability represents unexplained or small-scale fluctuations that cannot be readily attributed to specific processes (Wilding and Drees, 1983; Trangmar et al., 1986). In practice, these components are often intertwined and occur across multiple spatial scales, resulting in a nested structure of variability. This scale-dependent behavior complicates efforts to characterize soil properties, as the dominant sources of variation may shift depending on the resolution of observation (Beckett and Webster, 1971; Burrough, 1993).

Geostatistical approaches, particularly semivariogram-based methods, have become the dominant framework for quantifying soil spatial variability due to their ability to explicitly model spatial dependence and support interpolation techniques such as kriging (Journel and Huijbregts, 1978; Webster and Oliver, 2007). Parameters such as nugget, sill, and range provide valuable insight into the scale and strength of spatial dependence and are widely used to guide sampling design and spatial prediction. However, the application of these methods relies on assumptions such as stationarity and isotropy, which are often difficult to satisfy in heterogeneous agricultural landscapes (Trangmar et al., 1986; Cambardella et al., 1994). As a result, semivariogram-based models may oversimplify complex spatial patterns, particularly when multiple processes operate simultaneously across different scales.

The degree and structure of spatial variability also differ substantially among soil properties. Chemical properties such as pH and nutrient concentrations often exhibit relatively smoother spatial patterns and moderate to strong spatial dependence at field scales. In contrast, biological properties, including SMB and enzyme activity, tend to exhibit greater variability and more complex spatial behavior (Cambardella et al., 1994; Nunan et al., 2002). The increased variability is driven by the sensitivity of biological processes to localized environmental

conditions, resource availability, and management disturbances. Consequently, traditional geostatistical approaches that perform well for chemical properties may be less effective in capturing the spatial dynamics of biological indicators.

These differences have important implications for soil sampling and data interpretation. Because soil properties are spatially continuous, observations collected at nearby locations are not independent, leading to spatial autocorrelation (Rossel and McBratney, 1998; Webster and Oliver, 2007). Incorporating this spatial structure into the sampling design can improve estimation efficiency and reduce redundancy. At the same time, accurately capturing spatial variability, particularly for highly variable biological properties, often requires dense sampling, which may not be feasible in practical applications (Cambardella et al., 1994; Weindorf and Zhu, 2010). This creates a fundamental challenge in balancing sampling intensity with cost and operational constraints.

In recent years, there has been growing recognition that traditional geostatistical approaches alone may be insufficient to fully capture the complexity of soil variability, particularly for dynamic biological indicators. This has led to increased interest in integrating geostatistical methods with data-driven approaches, such as machine learning and hybrid modeling frameworks, which can better accommodate nonlinear relationships and multiscale variability. These developments suggest a shift toward more flexible and integrative approaches for characterizing soil spatial variability, particularly in the context of precision agriculture and soil health assessment.

As a result, the design of soil sampling strategies must account not only for the presence of spatial variability but also for its scale, structure, and dependence on the properties being measured. This has led to the development of a range of sampling approaches aimed at balancing

accuracy, cost, and practicality, particularly in heterogeneous, biologically dynamic soil systems. These constraints directly influence how soil sampling strategies are designed, particularly in systems where capturing biological variability is a priority.

1.5 Soil sampling strategies

Soil sampling is a fundamental component of soil health assessment and nutrient management, as it provides the primary data used to characterize soil properties and inform management decisions (Lawrence et al., 2020). However, due to the inherent spatial variability of soils, designing an effective sampling strategy is challenging and requires balancing accuracy, cost, and practicality. The goal of sampling may vary depending on the application, ranging from estimating field-scale averages to delineating sub-field variability for site-specific management (Brus and Heuvelink, 2007; Lawrence et al., 2020).

One of the most widely used approaches is grid-based sampling, where samples are collected at regular intervals across a field. This method provides uniform spatial coverage and is well-suited for capturing spatial patterns and generating interpolation surfaces (Burgess and Webster, 1980; McBratney et al., 1981). Grid sampling is commonly used in precision agriculture to support variable-rate management and spatial mapping of soil properties. However, achieving high-resolution characterization requires dense sampling, which can be time-consuming and expensive, particularly for large fields or when analyzing costly soil properties (Weindorf and Zhu, 2010; Lawrence et al., 2020). While grid-based sampling provides a robust framework for capturing spatial patterns, its reliance on high sampling density limits its practicality, particularly for biological indicators that require costly laboratory analysis.

To address these limitations, reduced sampling approaches have been explored, including grid thinning and composite sampling. These methods aim to reduce sampling intensity while maintaining acceptable accuracy. Nevertheless, reducing the number of samples increases uncertainty and may fail to capture fine-scale variability, especially for properties with high spatial heterogeneity (Cambardella et al., 1994; Brus and Heuvelink, 2007). The effectiveness of reduced sampling designs, therefore, depends on the spatial structure of the soil property being measured and the objectives of the study.

An alternative approach is management zone-based sampling, in which fields are divided into relatively homogeneous areas based on soil properties, yield history, topography, or sensor data. Samples are then collected within each zone to represent average conditions (Mallarino and Wittry, 2004; Taylor et al., 2007). This approach can improve sampling efficiency by targeting variability that is relevant for management decisions and has been widely adopted in precision agriculture. However, the performance of management zones is highly dependent on the quality and selection of input data used to define them, and results have been inconsistent across studies (Flowers et al., 2005; Mzuku et al., 2005).

Most existing sampling strategies have been developed primarily for chemical soil properties such as nutrients and pH, which generally exhibit smoother spatial patterns and can be reliably estimated at lower sampling densities. In contrast, biological indicators are typically more spatially variable, limiting the direct transferability of these approaches and necessitating greater sampling intensity to achieve comparable accuracy (Cambardella et al., 1994; Nunan et al., 2002).

Overall, soil sampling strategies involve a trade-off between sampling intensity, cost, and accuracy. Dense grid sampling provides detailed spatial information but is often impractical,

while reduced or zone-based approaches improve efficiency but may introduce uncertainty. These challenges emphasize the importance of sampling frameworks tailored to highly variable soil properties, particularly biological indicators. Given these limitations, alternative approaches such as predictive modeling have been increasingly explored to enhance the scalability and efficiency of soil property estimation.

1.6 Predictive modeling in soil science

Predictive modeling has emerged as a powerful approach to addressing the challenges of measuring and scaling soil properties, particularly those that are costly or difficult to quantify directly. In soil science, modeling frameworks are increasingly used to estimate complex or hard-to-measure variables from more readily available data, an approach conceptually like pedotransfer functions (PTFs) (Bouma, 1989; McBratney et al., 2002). These methods enable the prediction of soil properties across space while reducing the need for intensive sampling and laboratory analysis.

Recent advances in machine learning (ML) have significantly expanded the potential of predictive modeling in soil science. Algorithms such as random forests, gradient boosting, and other ensemble methods can capture complex, nonlinear relationships between soil properties and environmental covariates (Hengl et al., 2018; Wadoux et al., 2021). Unlike traditional statistical approaches, which often rely on assumptions of linearity and independence, ML models can integrate diverse datasets, including soil measurements, topographic variables, and remote sensing data, to improve predictive accuracy (Wadoux et al., 2021; Wilhelm et al., 2022).

Predictive modeling has been widely applied to estimate soil properties such as soil organic carbon, texture, and nutrient cycling across different spatial scales. More recently, there

has been growing interest in applying these approaches to biological indicators, including SMB and enzyme activity. Studies have shown that microbial properties can be predicted from soil physicochemical variables such as organic carbon, pH and texture, highlighting the potential for indirect estimation of biological processes (Fierer et al., 2009; Wilhelm et al., 2022). This capability is particularly valuable given the cost and complexity associated with direct measurement of biological indicators.

Yet, the application of ML to biological indicators remains less mature compared to its use for physicochemical soil properties. While studies have demonstrated that variables such as SOC, pH, and texture can explain a substantial proportion of variability in microbial biomass (Fierer et al., 2009; Wilhelm et al., 2022), and model performance tends to be more variable and context-dependent than for more stable soil properties such as SOC or texture. This reflects the inherently dynamic and nonlinear nature of biological systems, where microbial responses are influenced by short-term environmental fluctuations and management history. Consequently, although ensemble methods such as random forests have shown relatively robust performance across diverse datasets, their predictive reliability for biological indicators is often constrained by data quality, temporal variability, and site-specific conditions.

Despite its advantages, predictive modeling introduces new challenges, particularly regarding model interpretability and generalizability. Many ML models are often considered "black-box" approaches, making it difficult to understand the underlying relationships between predictors and response variables (Lundberg and Lee, 2017; Wadoux et al., 2021). This limitation can reduce confidence in model predictions, especially in applications where understanding the drivers of variability is as important as prediction accuracy. To address this issue, recent studies have incorporated model interpretation techniques such as feature

importance metrics and Shapley Additive Explanations (SHAP) values to improve transparency and interpretability (Lundberg and Lee, 2017; Wilhelm et al., 2022). Furthermore, while advances in model interpretation have improved transparency, there remains limited consensus on how best to evaluate and compare model performance for biological indicators across studies. Differences in sampling design, data resolution, and environmental context make it difficult to generalize findings, suggesting that model performance should be interpreted within the specific conditions under which the data were collected.

Another important consideration is the dependence of predictive models on the quality and representativeness of input data. Model performance is strongly influenced by sampling design, data resolution, and the spatial distribution of observations. If training data do not adequately capture the variability of the system, model predictions may be biased or unreliable when applied to new locations (Hengl et al., 2018; Wadoux et al., 2021). This highlights a critical link between predictive modeling and sampling strategies, in which model effectiveness depends on how well spatial variability is represented in the input data.

Overall, predictive modeling offers a promising approach to improving the scalability of soil health assessment, particularly for biological indicators that are difficult to measure directly. By leveraging relationships between easily measured and complex soil properties, modeling approaches can reduce sampling costs and enhance spatial coverage. However, their successful application depends not only on model selection and interpretation but also on the quality and spatial representativeness of the input data, highlighting the need to integrate modeling approaches with effective sampling strategies.

1.7 Knowledge gap and research needs

Although biological indicators such as microbial biomass provide critical insight into soil processes, their application in routine soil monitoring remains limited due to challenges related to cost, variability, and interpretability (Bünemann et al., 2018; Fierer et al., 2021). As a result, most soil health assessments continue to rely heavily on physical and chemical indicators, with biological components often underrepresented or simplified.

At the same time, existing soil sampling strategies have primarily been developed for chemical properties such as soil nutrients and pH. These properties tend to exhibit smoother spatial patterns and can often be estimated with relatively lower sampling intensity (Cambardella et al., 1994; Lawrence et al., 2020). In contrast, biological indicators exhibit greater spatial and temporal variability, which increases the uncertainty associated with their measurements and limits the effectiveness of conventional sampling approaches (Nunan et al., 2002; Bünemann et al., 2018). Consequently, sampling frameworks that perform well for chemical properties may not be directly applicable to biological indicators.

Predictive modeling approaches have demonstrated strong potential for estimating soil properties across spatial scales, particularly through machine learning techniques. However, many modeling studies focus on improving predictive accuracy without explicitly considering how sampling design influences model performance (Hengl et al., 2018; Wadoux et al., 2021). The quality, density, and spatial distribution of input data play a critical role in determining model reliability, yet the interaction between sampling strategies and predictive modeling remains insufficiently explored in the context of soil health assessment.

Furthermore, studies that integrate biological indicators with spatial analysis and predictive modeling are still limited. While individual components such as microbial

measurements, geostatistical analysis, and machine learning have been widely studied, their combined application for scalable soil health assessment remains underdeveloped. This lack of integration limits the ability to efficiently characterize biologically driven soil processes at field scales, particularly in agricultural systems where both economic and environmental considerations are important.

To address these gaps, there is a need for approaches that explicitly link biological indicators, sampling design, and predictive modeling within a unified framework. Such approaches should aim to improve the feasibility of measuring biological soil properties while maintaining their ecological relevance and interpretability. In particular, understanding how sampling intensity and design influence the accuracy of both direct measurements and model predictions is critical for developing cost-effective soil health monitoring strategies. Addressing these challenges requires a shift from treating biological indicators, sampling strategies, and predictive modeling as separate components toward a more integrated framework. The limitations identified throughout this review suggest that neither dense sampling nor predictive modeling alone is sufficient to reliably characterize biologically driven soil processes at field scales. Instead, there is a need for approaches that explicitly account for the interaction between spatial variability, sampling design, and model performance.

Within this context, integrating predictive modeling with optimized sampling strategies is not only advantageous but necessary to improve the feasibility and reliability of soil health assessment. By leveraging relationships between readily measured soil properties and biological indicators, while simultaneously evaluating how sampling intensity influences both direct measurements and model predictions, it becomes possible to develop more scalable and cost-effective monitoring frameworks.

This thesis addresses these needs by combining predictive modeling and sampling optimization to evaluate soil microbial biomass as a key biological indicator of soil health. Chapter 2 focuses on developing predictive models to estimate microbial biomass from soil physicochemical properties, while Chapter 3 evaluates the performance of reduced-intensity sampling strategies for capturing spatial variability in biological indicators. Together, these components provide a structured framework for advancing the practical integration of biological indicators into soil health assessment at the field scale.

References

Beckett, P.H.T., and R. Webster. 1971. Soil variability: A review. *Soils Fertil.* 34(1): 1–15.

Bosatta, E., and G.I. Ågren. 1994. Theoretical analysis of microbial biomass dynamics in soils. *Soil Biol. Biochem.* 26(1): 143–148. doi: https://doi.org/10.1016/0038-0717(94)90206-2.

Bouma, J. 1989. Using soil survey data for quantitative land evaluation. In: Stewart, B.A., editor, *Advances in Soil Science*, Vol. 9. Springer, New York. p. 177–213.

Brus, D.J., and G.B.M. Heuvelink. 2007. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma* 138(1): 86–95. doi: https://doi.org/10.1016/j.geoderma.2006.10.016.

Bünemann, E.K., G. Bongiorno, Z. Bai, R.E. Creamer, G. De Deyn, et al. 2018. Soil quality – A critical review. *Soil Biol. Biochem.* 120: 105–125. doi: https://doi.org/10.1016/j.soilbio.2018.01.030.

Burgess, T.M., and R. Webster. 1980. Optimal interpolation and isarithmic mapping of soil properties: I. The semivariogram and punctual kriging. *J. Soil Sci.* 31(2): 315–331. doi: https://doi.org/10.1111/j.1365-2389.1980.tb02084.x.

Burrough, P.A. 1993. Soil variability: A late 20th century view. *Soils Fertil.* 56: 529–562.

Cambardella, C.A., T.B. Moorman, J.M. Novak, T.B. Parkin, D.L. Karlen, et al. 1994. Field-scale variability of soil properties in central Iowa soils. *Soil Sci. Soc. Am. J.* 58(5): 1501–1511. doi: https://doi.org/10.2136/sssaj1994.03615995005800050033x.

Cardoso, E.J.B.N., R.L.F. Vasconcellos, D. Bini, M.Y.H. Miyauchi, C.A. dos Santos, et al. 2013. Soil health: Looking for suitable indicators. *Sci. Agric.* 70: 274–289. doi: https://doi.org/10.1590/S0103-90162013000400009.

Dalal, R.C. 1998. Soil microbial biomass—What do the numbers really mean? *Aust. J. Exp. Agric.* 38(7): 649–665. doi: https://doi.org/10.1071/EA97142.

Decaëns, T. 2010. Macroecological patterns in soil communities. *Glob. Ecol. Biogeogr.* 19(3): 287–302. doi: https://doi.org/10.1111/j.1466-8238.2009.00517.x

Dongdong, C., L. Qi, H. Lili, X. Qian, C. Xin, et al. 2023. Soil nutrients directly drive soil microbial biomass and carbon metabolism in the Sanjiangyuan alpine grassland. *J. Soil Sci. Plant Nutr.* 23(3): 3548–3560. doi: https://doi.org/10.1007/s42729-023-01270-y.

Doran, J.W., and M.R. Zeiss. 2000. Soil health and sustainability: Managing the biotic component of soil quality. *Appl. Soil Ecol.* 15(1): 3–11. doi: https://doi.org/10.1016/S0929-1393(00)00067-6.

Fierer, N., A.S. Grandy, J. Six, and E.A. Paul. 2009. Searching for unifying principles in soil ecology. *Soil Biol. Biochem.* 41(11): 2249–2256. doi: https://doi.org/10.1016/j.soilbio.2009.06.009.

Fierer, N., S.A. Wood, and C.P. Bueno de Mesquita. 2021. How microbes can, and cannot, be used to assess soil health. *Soil Biol. Biochem.* 153: 108111. doi: https://doi.org/10.1016/j.soilbio.2020.108111

Flowers, M., R. Weisz, and J.G. White. 2005. Yield-based management zones and grid sampling strategies. *Agron. J.* 97(3): 968–982. doi: https://doi.org/10.2134/agronj2004.0224

Gregorich, E.G., M.R. Carter, D.A. Angers, C.M. Monreal, and B.H. Ellert. 1994. Towards a minimum data set to assess soil organic matter quality. *Can. J. Soil Sci.* 74(4): 367–385. doi: https://doi.org/10.4141/cjss94-051.

Gyawali, A.J., H. Neely, J. Foster, C. Neely, K. Lewis, et al. 2023. Sampling for biological indicators of soil health: How does sampling methodology affect results? *Geoderma* 435: 116513. doi: https://doi.org/10.1016/j.geoderma.2023.116513

Hengl, T., M. Nussbaum, M.N. Wright, G.B.M. Heuvelink, and B. Gräler. 2018. Random forest as a framework for spatial prediction. *PeerJ* 6: e5518. doi: https://doi.org/10.7717/peerj.5518.

Horrigue, W., S. Dequiedt, N. Chemidlin Prévost-Bouré, C. Jolivet, N.P.A. Saby, et al. 2016. Predictive model of soil molecular microbial biomass. *Ecol. Indic.* 64: 203–211. doi: https://doi.org/10.1016/j.ecolind.2015.12.004.

Jenkinson, D.S., S.P.S. Andrew, J.M. Lynch, M.J. Goss, and P.B. Tinker. 1990. The turnover of organic carbon and nitrogen in soil. *Philosophical Transactions of the Royal Society B: Biological Sciences* 329(1255): 361–368. doi: https://doi.org/10.1098/rstb.1990.0177

Jiang-shan, Z., G. Jian-fen, C. Guang-shui, and Q. Wei. 2005. Soil microbial biomass and its controls. *J. For. Res.* 16(4): 327–330. doi: https://doi.org/10.1007/BF02858201.

Lawrence, P.G., W. Roper, T.F. Morris, and K. Guillard. 2020. Guiding soil sampling strategies using spatial statistics: A review. *Agron. J.* 112(1): 493–510. doi: https://doi.org/10.1002/agj2.20048.

Lehmann, J., D.A. Bossio, I. Kögel-Knabner, and M.C. Rillig. 2020. The concept and future prospects of soil health. *Nat. Rev. Earth Environ.* 1(10): 544–553. doi: https://doi.org/10.1038/s43017-020-0080-8.

Lundberg, S.M., and S.-I. Lee. 2017. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.*

Mallarino, A.P., and D.J. Wittry. 2004. Efficacy of grid and zone soil sampling. *Precis. Agric.* 5(2): 131–144. doi: https://doi.org/10.1023/B:PRAG.0000022358.24102.1b

Mann, C., D. Lynch, S. Fillmore, and A. Mills. 2019. Relationships between field management and soil health. *Appl. Soil Ecol.* 144: 12–21. doi: https://doi.org/10.1016/j.apsoil.2019.06.012.

McBratney, A.B., B. Minasny, S.R. Cattle, and R.W. Vervoort. 2002. From pedotransfer functions to soil inference systems. *Geoderma* 109(1): 41–73. doi: https://doi.org/10.1016/S0016-7061(02)00139-8.

McBratney, A.B., R. Webster, and T.M. Burgess. 1981. Optimal sampling schemes. *Comput. Geosci.* 7(4): 331–334. doi: https://doi.org/10.1016/0098-3004(81)90077-7

Mzuku, M., R. Khosla, R. Reich, D. Inman, F. Smith, et al. 2005. Spatial variability across management zones. *Soil Sci. Soc. Am. J.* 69(5): 1572–1579. doi: https://doi.org/10.2136/sssaj2005.0062.

Nunan, N., K. Wu, I.M. Young, J.W. Crawford, and K. Ritz. 2002. Spatial patterns of soil bacteria. *Microb. Ecol.* 44(4): 296–305. doi: https://doi.org/10.1007/s00248-002-2021-0.

Peigné, J., J.-F. Vian, M. Cannavacciuolo, B. Bottollier, and R. Chaussod. 2009. Soil sampling based on the spatial variability of microbial indicators. *Eur. J. Soil Biol.* 45(5): 488–495. doi: https://doi.org/10.1016/j.ejsobi.2009.09.002

Rinot, O., G.J. Levy, Y. Steinberger, T. Svoray, and G. Eshel. 2019. Soil health assessment: A critical review. *Sci. Total Environ.* 648: 1484–1491. doi: https://doi.org/10.1016/j.scitotenv.2018.08.259

Rossel, R.A.V. 2007. Modelling soil reflectance spectra. *J. Infrared Spectrosc.* 15(1): 39–47. doi: https://doi.org/10.1255/jnirs.694

Rossel, R.A.V., and A.B. McBratney. 1998. Soil analytical accuracy and costs. *Aust. J. Exp. Agric.* 38(7): 765–775. doi: https://doi.org/10.1071/EA97158.

Saiya-Cork, K.R., R.L. Sinsabaugh, and D.R. Zak. 2002. Nitrogen deposition effects on enzymes. *Soil Biol. Biochem.* 34(9): 1309–1315. doi: https://doi.org/10.1016/S0038-0717(02)00074-3.

Semenov, M.V., A.D. Zhelezova, N.A. Ksenofontova, et al. 2025. Microbiological indicators for soil health. *Agronomy* 15(2): 335. doi: https://doi.org/10.3390/agronomy15020335.

Singh, J.S., and V.K. Gupta. 2018. Soil microbial biomass: A key driver. *Sci. Total Environ.* 634: 497–500. doi: https://doi.org/10.1016/j.scitotenv.2018.03.373.

Stockdale, E.A., and P.C. Brookes. 2006. Soil microbial biomass measurement. *J. Agric. Sci.* 144(4): 285–302. doi: 10.1017/S0021859606006228

Sünnemann, M., C. Alt, J.E. Kostin, et al. 2021. Land-use effects on microbial biomass. *J. Appl. Ecol.* 58(11): 2614–2625. doi: https://doi.org/10.1111/1365-2664.14004.

Tabatabai, M.A. 1994. Soil enzymes. In: *Methods of Soil Analysis*. p. 775–833.

Taylor, J.A., A.B. McBratney, and B.M. Whelan. 2007. Management classes for agriculture. *Agron. J.* 99(5): 1366–1376. doi: https://doi.org/10.2134/agronj2007.0070

Templer, P., S. Findlay, and G. Lovett. 2003. Soil microbial biomass and nitrogen transformations. *Soil Biol. Biochem.* 35(4): 607–613. doi: https://doi.org/10.1016/S0038-0717(03)00006-3

Trangmar, B.B., R.S. Yost, and G. Uehara. 1986. Application of geostatistics to spatial studies of soil properties. *Adv. Agron.* 38: 45–94.

USDA. 2026. Soil health. United States Department of Agriculture. Available at: https://www.usda.gov/ (accessed 16 Jan. 2026).

Van der Heijden, M.G.A. 2008. Soil microbes and plant productivity. *Ecol. Lett.* 11(3): 296–310. doi: https://doi.org/10.1111/j.1461-0248.2007.01139.x

Wadoux, A.M.J.-C., M. Román-Dobarco, and A.B. McBratney. 2021. Perspectives on data-driven soil research. Eur. J. Soil Sci. 72(4): 1675–1689. doi: 10.1111/ejss.13071.

Wan, X., X. Chen, Z. Huang, and H.Y.H. Chen. 2021. Global soil microbial biomass decreases with aridity and land-use intensification. Glob. Ecol. Biogeogr. 30(5): 1056–1069. doi: 10.1111/geb.13282.

Wardle, D.A. 1998. Controls of temporal variability of the soil microbial biomass: A global-scale synthesis. Soil Biol. Biochem. 30(13): 1627–1637. doi: 10.1016/S0038-0717(97)00201-0.

Wardle, D.A., and A. Ghani. 1995. Why is the strength of relationships between pairs of methods for estimating soil microbial biomass often so variable? Soil Biol. Biochem. 27(6): 821–828. doi: 10.1016/0038-0717(94)00229-T.

Webster, R., and M.A. Oliver. 2007. *Geostatistics for Environmental Scientists*. Wiley.

Weindorf, D.C., and Y. Zhu. 2010. Spatial variability and sampling strategy. *Pedosphere* 20(2): 185–197. doi: https://doi.org/10.1016/S1002-0160(10)60006-9.

Welsch, J., C. Songling, H.L. Buckley, et al. 2019. Soil variability and sampling confidence. *Ecol. Indic.* 102: 401–409. doi: https://doi.org/10.1016/j.ecolind.2019.02.065.

Wilding, L.P., and L.R. Drees. 1983. Spatial variability and pedology. *Developments in Soil Science.* p. 83–116.

Wilhelm, R.C., H.M. van Es, and D.H. Buckley. 2022. Predicting soil health using ML. *Soil Biol. Biochem.* 164: 108472. doi: https://doi.org/10.1016/j.soilbio.2021.108472.

Chapter 2 - Soil microbial biomass as an emergent property of soil health: A predictive modeling approach

Abstract

Soil microbial biomass is a key indicator of soil health, integrating organic matter dynamics and nutrient cycling processes. However, direct measurement methods such as phospholipid fatty acid (PLFA) analysis are costly and time-intensive, limiting their use in routine soil health monitoring. This study aimed to (i) quantify how well routinely measured soil properties predict PLFA-derived microbial biomass, (ii) evaluate whether enzyme activity data improve predictions beyond soil properties alone, and (iii) identify the most influential predictors using interpretable machine learning.

Soils were sampled in 2024 from two Kansas sites with contrasting land uses: the Kansas State University North Farm (81 locations, grid-based sampling) and a Dickinson County site with pasture and no-till cropland (75 locations, zone-based sampling). Composite samples were collected at 0-5 cm and 5-15 cm depths (312 observations). Predictors included commonly measured soil properties: organic matter (OM), total carbon (C) and nitrogen (N), texture, pH, nutrients and an enzyme-augmented feature set including β-glucosidase (βG), N-acetyl-β-glucosaminidase (NAG), acid phosphatase (AP), and arylsulfatase (AS). Five machine learning algorithms were compared, including ensemble methods (Random Forest, XGBoost) and linear models (ElasticNet). Models were trained on transformed data to address response skewness and were evaluated using leakage-safe GroupKFold cross-validation.

Soil properties alone accurately predicted microbial biomass. XGBoost achieved the best performance (R^2 = 0.80; RMSE = 41.5; MAE = 23.9), with Random Forest performing comparably (R^2 = 0.79). Adding enzyme activity resulted in modest changes in prediction

accuracy, with slight improvements in error metrics but no consistent gains in explained variance (XGBoost R^2 = 0.86; Random Forest = 0.87). Interpretability analyses (permutation importance and SHAP) identified soil carbon (C%), organic matter (OM%), and nitrogen (N%) as the dominant predictors, with additional contributions from sand content and phosphorus. These results demonstrate that routinely measured soil properties contain substantial biological information for estimating microbial biomass.

Keywords: soil microbial biomass, PLFA, machine learning, Random Forest, SHAP, enzyme activity, soil health indicators

2.1 Introduction

2.1.1 Importance of soil microbial biomass

Soil health has been defined as the continued capacity of soil to function as a living system that sustains biological productivity, maintains environmental quality, and supports plant, animal, and human health (USDA, 2026). Within this framework, soil microbial communities represent a central component of soil functioning, regulating organic matter decomposition, nutrient cycling, and energy flow through terrestrial ecosystems (Wardle and Ghani, 1995; Wardle, 1998; Decaëns, 2010; van der Heijden et al., 2008; Sünnemann et al., 2021).

Soil microbial biomass (SMB) is widely perceived as a key indicator of soil fertility, ecosystem productivity, and soil quality, reflecting the size of the living microbial pool that mediates organic matter decomposition and nutrient cycling (Singh and Gupta, 2018). Acting both as a sink and source of carbon and nutrients, microbial biomass regulates the transformation and stabilization of soil organic matter (SOM) and directly affects the retention and release of organic carbon and nitrogen in soils (Templer et al., 2003; Jiang-shan et al., 2005). Because

microbial biomass integrates substrate availability, environmental constraints, and biological demand, it is particularly sensitive to management practices, land use change, and disturbance, often responding more rapidly than bulk soil pools (Harden et al., 1993; García and Hernández, 1997; Dalal, 1998; Horwath, 1994; Horrigue et al., 2016).

Biological indicators, including microbial biomass, enzymes, respiration, and biodiversity, often respond more rapidly to anthropogenic and natural drivers than physical or chemical indicators such as texture or total nutrient pools (Cardoso et al., 2013). Consequently, microbial biomass has been recommended as a rapid indicator of soil state and change (Bosatta and Ågren, 1994). However, interpreting microbial biomass remains challenging because higher biomass does not necessarily indicate higher activity or improved soil functioning, and responses can vary with soil properties, climate, and measurement methods (Fierer et al., 2021).

Phospholipid fatty acid (PLFA) analysis is a well-established technique for quantifying total microbial biomass and broad community structure by targeting phospholipids derived from the membranes of living microbial cells. Because phospholipids degrade rapidly following cell death, PLFA provides an estimate of the active microbial biomass rather than historical or relic material (Zelles, 1999). Despite this ecological relevance, PLFA analysis is technically demanding, time-consuming, and costly, requiring specialized extraction protocols and analytical expertise (Dalal, 1998; Semenov et al., 2025). Hence, PLFA-based measurements are typically limited to research-focused studies or small-scale sampling campaigns and are difficult to implement at the spatial and temporal scales required for routine soil health monitoring or management decision-making.

2.1.2 Need for predictive modeling

Monitoring soil health at meaningful spatial scales requires comprehensive diagnostics that integrate physical, chemical, and biological indicators; however, such assessments are frequently constrained by analytical costs, labor, and technical expertise (Wilhelm et al., 2022). Although modern soil monitoring programs routinely collect large volumes of physicochemical data, including organic carbon, nutrients, texture, and bulk density, biological indicators remain underrepresented due to the expense and complexity of laboratory analyses (Bünemann et al., 2018).

This imbalance raises a fundamental question for soil health assessment: to what extent does routinely measured soil data contain sufficient biological signal to infer microbial biomass? Strong evidence suggests that microbial abundance and activity are intrinsically linked to soil organic carbon pools, nutrient availability, texture, and pH, with microbial biomass often representing a relatively stable fraction (approximately 1–3%) of total SOC across diverse ecosystems (Anderson and Domsch, 1989; Wardle, 1992; Fierer et al., 2009). Across broad soil and ecosystem gradients, microbial properties and enzyme activities are remarkably predictable from a small number of abiotic variables, particularly pH and SOC (Fierer and Jackson, 2006; Sinsabaugh et al., 2008; Fierer et al., 2009).

Simultaneously, intensive laboratory measurements of microbial biomass and function are limited. Methods based on fumigation-extraction, PLFA analysis, or DNA-based approaches require careful sample handling, standardized protocols, and specialized equipment, and are subject to biases arising from extraction efficiency, gene copy number variability, and seasonal dynamics (Gregorich et al., 1994; Wardle and Ghani, 1995; Semenov et al., 2025). As a result,

microbial indicators are often measured infrequently or at limited locations, restricting their utility for spatial upscaling and long-term monitoring (Welsch et al., 2019).

Predictive modeling offers a potential pathway to overcome these constraints by leveraging routinely collected soil data to indirectly estimate microbial biomass. Data-driven approaches have gained momentum in soil science for extracting empirical relationships from complex datasets without imposing restrictive assumptions on functional form (Wadoux et al., 2021). When applied meticulously, such models can support scalable assessments of soil health while reducing dependence on intensive laboratory measurements (Wilhelm et al., 2022).

2.1.3 Gaps in existing studies

Despite growing interest in predictive approaches, several limitations remain in existing studies of soil microbial biomass modeling. Many predictive models of microbial biomass are site-specific, limiting cross-study comparisons and operational deployment. In addition, soil datasets frequently include measurements from multiple depths, yet statistical dependence across depth layers is often ignored, leading to optimistic performance estimates and reduced ecological realism (Wardle and Ghani, 1995).

A further challenge lies in the interpretability of predictive models. Machine learning methods are increasingly used to model soil biological properties because they can accommodate nonlinear relationships and interactions among predictors (Wadoux et al., 2021; Pellegrini et al., 2021). However, complex models can function as 'black boxes', where predictive accuracy is achieved without clear insight into which soil properties exert dominant control or how management influences microbial abundance (Wadoux et al., 2021). This limitation is significant in the context of soil health, where indicators must be interpretable and actionable to advise management decisions (Doran and Zeiss, 2000; Fierer et al., 2021).

Additionally, increasing the number of indicators can exacerbate collinearity, inflate costs, and complicate interpretation, reinforcing the need to identify minimal yet informative indicator sets (O'Sullivan et al., 2017; Bünemann et al., 2018). Addressing these gaps requires modeling frameworks that balance predictive performance with comprehensibility, account for sampling constraints, such as depth dependence, and explicitly link microbial biomass to soil health indicators meaningful beyond individual study sites.

2.1.4 Objectives

The objectives of this chapter were to:

1. Quantify the extent to which soil physicochemical properties explain variability in total microbial biomass, thereby evaluating microbial biomass as an emergent property of soil health rather than an isolated biological measurement.
2. Assess whether microbial enzyme activity provides additional explanatory power for predicting microbial biomass and distinguish the relative contributions of stable soil properties compared to dynamic biological processes.
3. Identify key soil health indicators that exert dominant control over microbial biomass, using interpretable machine learning approaches to link soil chemical, physical, and biological attributes with microbial abundance.
4. Develop a robust and reproducible modeling framework that reflects realistic soil sampling constraints, ensuring leakage-safe evaluation across soil depths and heterogeneous field conditions.

2.2 Materials and Methods

2.2.1 Study sites and data integration

Site description

This study was conducted at two locations in northeastern and central Kansas, representing contrasting land uses and management histories. The first site was the Kansas State University Regenerative Agriculture Innovation Farm (North Farm; Riley County, KS; 39.2162° N, 96.6007° W), an approximately 40-ha research facility subdivided into four fields. Three fields have been managed under continuous no-till for around 22 years (sorghum-wheat-corn rotation), and one field is conventionally tilled for instructional and experimental use. Dominant soil series include Smolan silt loam, Wymore silty clay loam, and Reading silt loam. Soil samples were collected early spring 2024 using an approximate one-core-per-acre design. Each sampling location represents one core and consists of a composite soil sample formed by combining multiple subsamples collected within a small radius around a georeferenced point. Composite samples were obtained at two depth intervals (0-5 cm and 5-15 cm). The site has a mean annual temperature of 12.8°C and a mean annual precipitation of 908 mm based on 30-year climate normals (NOAA, 2023).

The second site was in Dickinson County, KS (38.6200°N, 97.1977°W) and included both pasture and cropland systems. Approximately 400 acres of pasture are managed under continuous stocking (~180 grazing days yr^{-1}), and approximately 500 acres of cropland are managed under continuous no-till for around 20 years (sorghum-soybean-wheat rotations and mixed fallow). Dominant soil series include Clime and Irwin silty clay loam in pasture areas and Irwin silty loam and Hobbs and Muir silt loam in cropland areas. At this site, soil samples were collected in late spring 2024 using a zone-based design, with management zones delineated by

spatial patterns in soil organic carbon and soil pH. Samples were collected at 0-5 and 5-15 cm to match the North Farm protocol. The site has a mean annual temperature of 12.2°C and mean annual precipitation of 870 mm based on 30-year climate normals (NOAA, 2023). Soil series information was obtained from the USDA NRCS Web Soil Survey (USDA NRCS, 2024). Across both sites, the final dataset comprised 156 sampling locations (North Farm: 81; Dickinson County: 75), with paired depth samples at each location, resulting in 312 depth-specific observations. (Table 2; Fig 2.1)

Depth restriction and rationale

Analyses were restricted to the 0-5 and 5-15 cm depth intervals because microbial biomass and associated biological processes are concentrated near the soil surface and are most responsive to land management and disturbance (Fierer et al., 2003; Sun et al., 2021). Restricting depth to the surface profile ensured consistency across sites and focused model development on the biologically active soil zone most relevant for soil health assessment.

Data integration strategy

Data from both sites were pooled to develop predictive models of soil microbial biomass applicable across contrasting agroecosystems. Pooling was appropriate because both sites used identical depth intervals (0-5 and 5-15 cm) and measured the same soil chemical, physical, and biological variables using consistent laboratory protocols. Pooling increased the effective sample size, improving model stability when predictor dimensionality is moderate-to-high relative to the sample size and reducing the risk of overfitting (Babyak, 2004).

Because samples were collected using different spatial designs (grid-based at North Farm and zone-based at Dickinson County), pooled modeling was evaluated using leakage-safe

validation to ensure generalization across sampling intensities and landscape contexts (Fletcher et al., 2019).

To account for depth-paired observations arising from the same soil core, all model evaluations used grouped cross-validation with NEW_id as the grouping variable. This ensured that samples from the same location were never split between the training and validation folds, preventing information leakage and optimistic performance estimates.

2.2.2 Soil sampling, response variable and depth pairing

Soil sampling design and depth pairing

At North Farm, sampling locations were generated as georeferenced points using a one-acre grid in ArcGIS Pro. At each location, soil cores were collected using a manual soil punch probe (1.9 cm in diameter) at 0-5 cm and 5-15 cm depths. Multiple cores were collected within a small radius at each location and composited by depth to form a representative sample.

At the Dickinson County site, a zone-based strategy was used to capture within-field variability at reduced sampling intensity. Management zones were delineated based on spatial patterns in soil organic carbon and soil pH, and georeferenced sampling locations were selected within those zones. Zone-based designs are commonly used to reduce sampling density while maintaining representation of spatial variability when zones reflect stable spatial patterns (Franzen et al., 2000; Mallarino and Wittry, 2004).

After collection, samples were homogenized, and visible plant residues and stones were removed. Samples were stored at 4°C before laboratory analysis, and a subsample was freeze-dried for PLFA analysis. Each sampling location was assigned a unique core identifier (*NEW_id*). As two depth increments were collected from the same physical core, observations

were depth-paired and not statistically independent. This pairing reflects shared microsite conditions and vertical continuity within individual soil profiles.

Response variable: PLFA derived soil microbial biomass

Total microbial biomass was quantified using phospholipid fatty acid (PLFA) analysis. Freeze-dried subsamples were processed using a modified Bligh and Dyer extraction (Bligh and Dyer, 1959) following established soil PLFA protocols (White and Rice, 2009). Phospholipids were separated using silicic acid chromatography, saponified with potassium hydroxide, and methylated to form fatty acid methyl esters (FAMEs).

FAMEs were analyzed using Thermo Scientific Trace GC-ISQ-GC-MS equipped with a DB-5MS capillary column (30 m × 250 μm i.d. × 0.25 μm film thickness). Peaks were identified using a bacterial acid methyl ester (BAME) standard mix (Matreya 1114) with additional authentic standards as needed. Quantification used methyl nonadecanoate (19:0 FAME) as an internal standard. Total microbial biomass was calculated as the sum of identified PLFA biomarkers and reported as nmol PLFA g^{-1} dry soil. PLFA-derived biomass was used because it estimates the living microbial community and is sensitive to management-driven changes in soil conditions (Frostegård et al., 1993; Zelles, 1999; White and Rice, 2009).

2.2.3 Predictor variables

Operational soil variables

Soil samples were air-dried, homogenized, and sieved (2 mm) following removal of visible residues. Soil chemical analyses were conducted by the Kansas State University Soil Testing Laboratory using standard procedures. Soil pH was measured using standard laboratory protocols. Plant-available phosphorus was determined using Mehlich-3 extraction (Frank et al., 1998) and quantified colorimetrically (Lachat QuickChem 8000). Exchangeable cations (Ca^{2+},

K^+, Mg^{2+}, Na^+) were extracted using 1M ammonium acetate (pH 7.0) and quantified by ICP-OES (Warncke and Brown, 1998). Soil organic carbon and total nitrogen were measured by dry combustion using the LECO TruSpec CN Analyzer, and carbonate carbon was removed by dilute phosphoric acid pretreatment for soils with pH above 7.2. Plant-available micronutrients (Fe, Zn, Cu, Mn) were extracted using DTPA and quantified by ICP-OES (Whitney, 1998). Soil texture was determined using the hydrometer method following dispersion with sodium hexametaphosphate (Gee and Or, 2002).

Soil depth was included as a contextual predictor. Management system was initially considered but excluded from final modeling to prevent dominance of categorical effects and to focus on soil-driven variability (Section 2.6).

Enzyme activity variables

The potential activities of four enzymes were quantified using colorimetric assays with p-nitrophenyl (PNP)-linked substrates, following standard procedures (Tabatabai, 1994): β-glucosidase (βG), arylsulfatase (AS), acid phosphatase (AP), and N-acetyl-β-glucosaminidase (NAG). Briefly, 0.5 g of soil was incubated with appropriate buffer and substrate at 37°C for 1 hour, with controls to account for background absorbance.

Reactions were terminated using $CaCl_2$ and alkaline stop solution, and absorbance was measured at 400 nm (Genesys 50 UV visible spectrophotometer). Activities were quantified using p-nitrophenol calibration curves and expressed as μg PNP g^{-1} soil h^{-1}. Enzyme variables were included to test whether these indicators improved prediction beyond operational soil properties.

2.2.4 Exploratory data analysis

Exploratory data analysis (EDA) was performed to summarize the distribution of samples across sites, depths, and management systems, assess missing data, and identify potential data issues before modeling. EDA was used only for understanding and quality control, not for feature selection or model optimization (Tukey, 1993).

Data structure, distribution and missingness

The combined dataset comprised 156 sampling locations (North Farm: 81; Dickinson County: 75) and paired depth observations at depths of 0-5 and 5-15 cm. Missingness was assessed across predictors before model development. Predictors with excessive missingness were excluded, and the remaining missing values were handled in the modeling pipeline using imputation procedures applied during cross-validation. Samples with missing response measurements were excluded from modeling.

Total microbial biomass exhibited right-skewness. To improve distributional symmetry and reduce the influence of extreme values during model fitting, a Yeo–Johnson power transformation was applied to the response variable before modeling (Yeo and Johnson, 2000). All models were trained and evaluated using the transformed response.

Exploratory relationships and multivariate structure

Pairwise correlations among continuous soil predictors were examined to understand data structure and anticipate multicollinearity. These analyses were exploratory, and no predictors were removed based on correlation at this stage.

Principal component analysis (PCA) was used as an exploratory visualization tool to assess the multivariate overlap among samples by management system and soil depth. PCA was not used for dimensionality reduction, feature selection, or model input.

2.2.5 Feature selection strategy

To evaluate the incremental predictive value of biologically intensive measurements, predictors were organized into two pre-specified feature sets: (i) a soil-only feature set consisting of operational soil properties plus depth, and (ii) a bio-augmented feature set consisting of the soil-only predictors plus enzyme activities. All modeling and selection procedures were conducted within the cross-validation framework to prevent information leakage.
Predictors with excessive missingness, duplicates, unit-equivalent variables, or directly derived variables were removed before modeling. Collinearity was addressed primarily within the modeling framework using regularization and model classes that tolerate correlated predictors, rather than aggressive correlation-based removal (Breiman, 2001; Zou and Hastie, 2005).

Model-based feature importance was used to characterize influential predictors while preserving leakage-safe evaluation. For tree-based models, permutation importance was used; for ElasticNet, standardized coefficient magnitudes were examined. Importance measures were interpreted descriptively to explain model behavior rather than as evidence of causality (Strobl et al., 2007). All machine learning models were trained using an identical predictor set to ensure comparability across algorithms.

2.2.6 Modeling framework and validation strategy

Cross-validation design

Model performance was evaluated using GroupKFold cross-validation with NEW_id as the grouping variable to prevent depth-based leakage (Kuhn and Johnson, 2013). All algorithms and feature sets were evaluated using the same folds, and performance was summarized as the mean ± standard deviation across folds.

Preprocessing pipeline

All preprocessing was implemented within a unified pipeline and performed independently within each cross-validation fold. Continuous predictors were imputed using median imputation and standardized. Categorical predictor (depth) was one-hot encoded. Preprocessing parameters were learned from training folds and applied to validation folds without recalibration.

Machine learning models and evaluation metrics

Five regression algorithms were evaluated: ElasticNet, Random Forest, Support Vector Regression, XGBoost, and k-Nearest Neighbors. Hyperparameters were tuned within the grouped cross-validation framework to optimize predictive performance while avoiding overfitting (Cortes and Vapnik, 1995; Breiman, 2001; Zou and Hastie, 2005; Chen and Guestrin, 2016). Performance was assessed using the coefficient of determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). Metrics were computed within each fold and summarized as mean ± standard deviation across folds. When sensitivity analyses involved alternative transformations, predictions were back-transformed as needed before computing error metrics.

2.2.7 Model explainability and interpretation

Global feature importance and SHAP (Shapley Additive exPlanations) were used to interpret the model's behavior. For ElasticNet, standardized coefficients were used to assess direction and relative influence. For tree-based models, permutation feature importance was computed on held-out folds (Strobl et al., 2007). SHAP summary and dependence plots were generated for selected models to characterize the direction and magnitude of predictor effects (Lundberg and Lee, 2017). Explainability outputs were interpreted as describing the model structure rather than as causal.

All analyses were conducted in Python. Model training and preprocessing were implemented using scikit-learn pipelines, with grouped cross-validation (GroupKFold) to prevent depth-based leakage across paired samples. Hyperparameters were tuned within the grouped cross-validation framework, and the final selected settings for the primary model (XGBoost) are reported in (Table 2.3)

The entire workflow is described in (Fig 2.2)

Statistical software and implementation

All statistical analyses and machine learning models were implemented in Python 3.12 (Python Software Foundation, 2023). Model training, preprocessing, and cross-validation were conducted using scikit-learn version 1.6.1 (Pedregosa et al., 2011). Gradient boosting models were implemented using XGBoost version 3.1.3 (Chen and Guestrin, 2016). Model interpretability was assessed using SHAP version 0.50.0 (Lundberg and Lee, 2017). Data manipulation and exploratory analyses were performed using pandas version 2.2.2 (McKinney, 2010) and NumPy version 2.0.2 (Harris et al., 2020). Visualizations were generated using matplotlib version 3.10.0 (Hunter, 2007) and seaborn version 0.13.2 (Waskom, 2021).

2.3 Results

2.3.1 Correlation structure of soil and biological predictors

To characterize relationships among predictors and assess potential redundancy within the bio-augmented feature set, pairwise correlations were examined using a correlation matrix (Fig 2.3). Strong positive correlations were observed among soil organic indicators, including organic matter (OM%), total carbon (C%), and total nitrogen (N%), indicating substantial shared variance among these variables. Enzyme activities (β-glucosidase, N-acetylglucosaminidase, phosphatase, and arylsulfatase) were positively correlated with organic matter-related variables,

particularly OM%, C%, and N%, reflecting their close association with substrate availability and microbial activity.

Sand % had a positive correlation with organic indicators and enzyme activities, whereas clay content was inversely related to most of these variables. Several micronutrients (e.g., Zn, Fe, Mn) demonstrated moderate correlations with both organic and biological variables, whereas pH had weaker and more variable associations across predictors. Overall, the correlation matrix revealed substantial multicollinearity among organic and biological indicators, particularly within functional groups, motivating the use of modeling approaches that handle correlated predictors rather than relying on variable pre-filtering.

2.3.2 Principal component analysis (PCA)

Principal component analysis revealed a coherent multivariate structure among predictors (Fig 2.4). PC1 (34.9% variance explained) captured a gradient of soil organic status, with strong positive loadings for OM%, C%, N%, and enzyme activities. PC2 (18.9% variance explained) primarily reflected variation in soil chemical properties, including pH, Ca, Fe, Mn, and P. Samples were visualized by management system for exploratory context; however, management was not included as a predictor in subsequent modeling analyses. While some separation among management groups was observed along PC1, substantial overlap indicates that variability is largely driven by continuous gradients in soil properties and biological activity rather than discrete categories.

Surface soils (0–5 cm) had greater dispersion than subsurface samples, consistent with higher biological heterogeneity near the soil surface. These patterns provide context for the relationships among predictors but were not used for dimensionality reduction in model development.

2.3.3 Model performance comparison

Model performance was evaluated using grouped cross-validation to compare predictive accuracy across machine learning algorithms within the soil-only feature set. The predictive performance was assessed using R^2, RMSE, and MAE as mean ± standard deviation across cross-validation folds (Table 2.4).

Model performance varied substantially across algorithms for the operational (soil-only) feature set. XGBoost achieved the highest predictive performance, with a mean R^2 of 0.80 (± 0.06), along with the lowest RMSE (41.52 ± 5.72) and MAE (23.94 ± 1.96), indicating strong predictive accuracy and consistency across validation folds.

K-Nearest Neighbors (KNN) and Random Forest demonstrated comparable performance, with mean R^2 values of 0.79 (± 0.08) and 0.79 (± 0.07), respectively. However, both models exhibited slightly higher RMSE (43.14–43.34) and MAE (24.15–24.22) relative to XGBoost. KNN had greater variability across folds, reflecting sensitivity to local data structure and feature scaling.

In contrast, Support Vector Machines (SVM) and ElasticNet regression had substantially lower predictive performance, with mean R^2 values of 0.31 and 0.30, respectively, and considerably higher RMSE (78.17–79.60) and MAE (44.51–44.91). These results indicate limited ability of linear and kernel-based methods to capture the underlying relationships in the data.

Overall, tree-based ensemble models outperformed linear, kernel-based, and instance-based approaches, highlighting the importance of nonlinear relationships and interactions among soil properties in explaining variability in soil microbial biomass. XGBoost was selected as the

primary model for subsequent analyses due to its superior predictive performance, lower error metrics, and consistent behavior across cross-validation folds.

2.3.4 Effect of enzyme activity inclusion

Model performance was compared between soil-only and bio-augmented feature sets to evaluate the contribution of enzyme activity variables (Table 2.5). Across models, the inclusion of enzyme variables resulted in modest changes in predictive performance, with patterns varying by model type.

For tree-based ensemble models, the bio-augmented feature set had no consistent improvement in R^2 compared to soil-only models. Random Forest achieved a mean R^2 of 0.87 (± 0.04), while XGBoost had a mean R^2 of 0.86 (± 0.04), both of which were marginally lower than their soil-only counterparts. However, these models exhibited improved error metrics, with lower RMSE (33.35–34.77) and MAE (20.21–20.24), suggesting improved prediction stability despite minor reductions in explained variance.

K-Nearest Neighbors demonstrated moderate performance ($R^2 = 0.81 \pm 0.06$), with higher RMSE (40.43) and MAE (22.60) relative to ensemble models. Similar to the soil-only results, Support Vector Machines and ElasticNet regression demonstrated substantially lower predictive performance, with mean R^2 values of 0.26 and 0.39, respectively, and considerably higher error metrics across all folds.

While enzyme activities provide functional insight into nutrient cycling processes, their contribution to large-scale microbial biomass prediction appears limited when compared to foundational soil properties such as organic matter, carbon, and texture. These findings suggest that enzyme measurements may be more valuable for process-level interpretation than for broad predictive modeling applications.

2.3.5 Key predictors of microbial biomass

Interpretability analyses were conducted using the XGBoost model, selected as the best-performing model for the operational (soil-only) feature set. Feature importance was evaluated using permutation feature importance (PFI; Table 2.7) and SHAP-based interpretations (Fig 2.5 and 2.6; Table 2.8), with strong agreement in identifying the dominant predictors of soil microbial biomass.

Across both interpretability approaches, soil organic properties emerged as the most influential predictors. Soil carbon (C%) was consistently ranked as the most important variable, followed by organic matter (OM%) and total nitrogen (N%). These variables exhibited the highest PFI scores and the largest mean absolute SHAP values, indicating their dominant role in explaining variability in microbial biomass.

In addition to organic indicators, selected soil physical and chemical properties contributed to model predictions. Sand content and phosphorus (P) ranked among the next most important predictors, followed by exchangeable cations such as Mg and K. Soil pH was moderately important, while micronutrients (Zn, Mn, Fe) and other variables contributed more modestly to predictive performance.

SHAP dependence patterns (Fig 2.6) further revealed the direction and magnitude of predictor effects. Higher values of C%, OM%, and N% were consistently associated with positive contributions to model predictions, indicating increased microbial biomass with greater organic resource availability. In contrast, sand content exhibited a mixed but generally negative influence, reflecting reduced microbial biomass in coarser-textured soils. Nutrient variables such as P, Mg, and K have nonlinear and heterogeneous effects, suggesting complex interactions with soil biological processes.

Soil texture components such as clay and silt fractions, along with micronutrients, exhibited relatively less important, indicating a secondary role in explaining microbial variability in the observed dataset. Soil depth was included in the model but had minimal importance, suggesting that depth-related differences in microbial biomass were largely captured indirectly through associated variations in soil organic matter and nutrient availability.

The strong agreement between permutation importance and SHAP analyses highlights the robustness of the identified predictors. These results emphasize that soil organic properties, particularly carbon and nitrogen pools, are the primary drivers of microbial biomass variability, while other physicochemical properties contribute additional but comparatively smaller explanatory power.

2.3.6 Model performance using top predictors

To evaluate whether a reduced set of predictors could maintain predictive performance, models were re-trained using the top-ranked operational variables identified from feature importance analysis (C%, OM%, N%, sand content, and P, along with soil depth; Table 2.6). Among the evaluated models, Random Forest achieved the highest performance for the reduced feature set, with a mean R^2 of 0.79 (± 0.07), RMSE of 43.01 (± 7.01), and MAE of 24.70 (± 2.02). XGBoost and KNN demonstrated comparable but slightly lower performance ($R^2 \approx 0.77$), while ElasticNet and SVM continued to exhibit substantially lower predictive accuracy.

Compared to the full operational model ($R^2 \approx 0.80$), the reduced model had a modest decline in performance, indicating that a small subset of key soil variables captures most of the explainable variability in microbial biomass.

These results demonstrate that simplified models with a limited number predictors can achieve performance comparable to full-feature models, highlighting the potential for more cost-effective and operationally feasible soil health assessments.

2.4 Discussion

2.4.1 Predictability of microbial biomass using soil properties

The strong predictability of soil microbial biomass from operational soil physicochemical variables underscores that soil properties serve as integrated proxies of microbial habitat quality. Microbial abundance is governed by a combination of resource availability, physicochemical constraints, and habitat structure, all of which are indirectly captured by commonly measured soil properties such as organic matter, carbon and nitrogen content, texture, and pH (Fig 2.5).

The strong influence of pH in our model reflects its role in governing microbial-driven biogeochemical cycling of C, N, and S, thereby integrating biological activity with geochemical buffering (Bolan and Hedley, 2003; Philippot et al., 2024). In parallel, mineral-associated nutrients and trace elements, including P, Zn, Mg, and other micronutrients, play essential roles in microbial metabolism, often serving as catalytic centers in enzymes or as structural components of microbial cells (Philippot et al., 2024). These variables, therefore, represent not isolated drivers but cumulative indicators of the chemical environment regulating microbial growth.

The importance of organic matter-related predictors is consistent with extensive evidence that soil organic carbon governs microbial abundance across spatial scales. Soil organic matter influences nutrient storage, water holding capacity, aggregate stability, and substrate availability, all of which shape microbial habitats (Gregorich et al., 1994). As the living component of soil organic matter, microbial biomass plays a central role in transforming organic matter and cycling

nutrients (Jenkinson, 1988; Moore et al., 2000). Numerous studies have demonstrated strong relationships between microbial abundance and SOC concentration, with land use influencing microbial biomass largely through its effects on SOC, C:N ratio, and pH (Wan et al., 2021; Williams et al., 2025).

Texture-related variables further explain the observed predictability. Fine-textured soils promote the stabilization of organic matter through organo-mineral associations and physical protection within aggregates, limiting microbial access to substrates while simultaneously providing stable habitats (Hassink, 1997; Six et al., 2002; Dungait et al., 2012). The preferential enrichment of microbial-derived compounds in silt and clay fractions has been widely reported and supports the role of texture in regulating microbial habitat size and persistence (Sessitsch et al., 2001; Six et al., 2002).

Together, these relationships align with soil ecological theory, suggesting that, despite the inherent complexity of soil systems, microbial abundance exhibits consistent, predictable patterns governed by a limited set of dominant physicochemical constraints (Fierer et al., 2009; Philippot et al., 2012). Our findings suggest that microbial biomass follows predictable patterns not because soils are uniform, but because the same physicochemical controls apply across diverse landscapes.

2.4.2 Added value of enzyme activity

Beyond standing biomass, soil enzymes capture process-level functionality. Soil enzymes are directly involved in the decomposition of organic matter and the mobilization of nutrients and are therefore sensitive to management, substrate availability, and environmental conditions (Dick, 1994; Tabatabai, 1994; Cui and Holden, 2015). Consequently, enzyme activities have been widely used as indicators of soil functional status and management effects (Saiya-Cork et

al., 2002; Waldrop et al., 2004; Tian et al., 2010). Numerous studies have indicated that enzyme activities increase under systems with higher organic matter inputs and reduced disturbance, such as pasture and conservation tillage systems (Tian et al., 2010; Avellaneda-Torres et al., 2013; Zhu et al., 2024).

Furthermore, the relatively modest gains associated with enzyme inclusion are consistent with known properties of enzyme dynamics. Enzyme activity reflects the size of the enzyme pool at a given moment and can become increasingly decoupled from carbon and nitrogen mineralization over time due to enzyme stabilization and legacy effects (Geisseler and Horwath, 2009; Philippot et al., 2012). As a result, enzyme measurements may capture short-term functional potential rather than long-term microbial abundance, particularly in heterogeneous field environments.

Additionally, enzyme activities are often strongly correlated with organic matter and microbial biomass, leading to partial redundancy with soil physicochemical predictors (Fig 2.3). Correlation analyses commonly indicate strong associations between enzymes, SOC, and nitrogen pools, while relationships with pH and texture are more variable (Williams et al., 2025). This overlap limits the incremental predictive gain from including enzymes, especially in linear or distance-based modeling frameworks.

Despite these limitations, enzyme variables provide a valuable functional context. Their inclusion is best interpreted as enhancing biological realism rather than fundamentally altering predictive capacity. This distinction is important for soil health assessment, where enzyme activity provides insight into nutrient-cycling potential and ecosystem functioning, even when it does not substantially improve the prediction of microbial biomass per se (Dick, 1994; de Almeida and Naves, 2015).

2.4.3 Implications for soil sampling and cost reduction

The ability to predict microbial biomass using frequently measured soil properties has important implications for soil health monitoring and sampling design. Biological indicators are recognized as sensitive and informative, yet their adoption in large-scale monitoring programs is often constrained by analytical costs, labor requirements, and methodological complexity (Doran and Zeiss, 2000; Bünemann et al., 2018). As a result, many soil quality frameworks rely primarily on physical and chemical indicators, with biological metrics absent from a substantial proportion of minimum data sets (Bünemann et al., 2018).

Predictive modeling provides a pragmatic pathway to bridge this gap. By leveraging soil physicochemical measurements as proxies, it becomes possible to infer biological status across fields while reserving intensive laboratory analyses for targeted validation or high-interest zones. This approach aligns with earlier concepts of pedotransfer functions, in which difficult-to-measure soil properties are estimated from more readily measured variables, while acknowledging that such estimates should be interpreted as approximations rather than mechanistic representations (Bouma, 1989; Baveye, 2025).

The use of grouped cross-validation further strengthens the relevance of this framework by ensuring that predictive performance reflects realistic deployment scenarios rather than being artificially inflated by depth or spatial dependence. Multi-model consistency among high-performing nonlinear algorithms increases confidence that the identified relationships are robust and not model-specific. Importantly, integrating biological indicators with soil physical and chemical variables supports a holistic soil health perspective without requiring exhaustive biological sampling at every location.

2.4.4 Methodological strengths and limitations

Tree-based ensemble models, particularly XGBoost, demonstrated strong performance in this study due to their ability to capture nonlinear relationships and complex interactions characteristic of soil biological systems. Microbial biomass is influenced simultaneously by organic matter availability, nutrient status, texture-mediated habitat constraints, and physicochemical conditions, many of which interact in ways that violate assumptions of linearity or additivity. Gradient boosting algorithms such as XGBoost iteratively improve model performance by minimizing residual error across sequential decision trees, allowing for the representation of subtle nonlinear patterns and hierarchical interactions within the data (Chen and Guestrin, 2016). This flexibility is particularly advantageous in soil systems, where threshold responses and coupled biogeochemical processes are common.

The comparable performance of Random Forest and XGBoost suggests that the dominant relationships governing microbial biomass are robust across ensemble modeling approaches. Both methods consistently outperformed linear, kernel-based, and instance-based models, indicating that nonlinear interactions among soil properties play a central role in explaining microbial variability. While Random Forest constructs independent decision trees using bootstrapped samples (Breiman, 2001), XGBoost refines predictions through boosting, leading to improved accuracy and reduced bias in many cases. The consistency between these approaches increases confidence that model performance is driven by underlying data structure rather than model-specific artifacts (Wadoux et al., 2020a; Zhang et al., 2025).

An additional advantage of ensemble models lies in their robustness to multicollinearity among predictors. Soil variables such as organic matter, carbon, nitrogen, and enzyme activities are inherently correlated because they share biogeochemical controls. Unlike parametric or

distance-based methods, tree-based models do not require explicit decorrelation or variable exclusion to maintain predictive stability, making them well-suited for integrated soil health datasets where redundancy is expected rather than avoidable (Hengl et al., 2018). This property is particularly relevant for biological indicators, which often covary as part of coupled ecological processes rather than acting independently.

From an interpretability perspective, the use of permutation importance and SHAP values provides transparency in identifying dominant predictors while maintaining the predictive strength of machine learning models. Although these approaches do not establish causality, they enable consistent identification of influential variables and allow for ecological interpretation of model outputs (Jenny, 1994; Wadoux et al., 2020b). In this study, agreement between permutation-based and SHAP-derived importance supports the robustness of identified predictors and highlights the dominant role of soil organic properties in shaping microbial biomass.

The methodological framework highlights both the strengths and inherent constraints of predictive modeling of soil biological properties. The use of ensemble-based machine learning methods accommodates nonlinear relationships and complex interactions that are difficult to represent in parametric models (Wadoux et al., 2020a; Liu et al., 2022; Zhang et al., 2025). Interpretability tools such as permutation importance and SHAP provide transparency regarding dominant predictors while acknowledging that importance rankings should not be equated with causal inference (Jenny, 1994; Wadoux et al., 2020b).

Residual spatial dependence remains an unavoidable limitation. Soil microorganisms are spatially heterogeneous and often spatially disconnected from the substrates they decompose, leading to patchiness that point-based predictors cannot fully capture (Ettema and Wardle, 2002;

Baveye, 2023; Manzoni and Schimel, 2024). Additionally, biological variables frequently exhibit collinearity due to shared dependence on organic matter and nutrient pools, complicating attribution among closely coupled predictors.

Temporal variability represents a further constraint. Microbial biomass and enzyme activities respond dynamically to climate, management, and substrate inputs, and single-season observations provide only a snapshot of these processes (Wilhelm et al., 2022). These limitations do not undermine the predictive utility but emphasize that modelled estimates represent conditional biological states rather than fixed properties.

From a practical standpoint, these findings support the development of tiered soil health monitoring strategies that balance analytical cost with biological insight. High-density sampling of physicochemical properties can be used to generate spatial predictions of microbial biomass, while targeted biological measurements can be reserved for validation or areas of uncertainty. Such approaches enable more efficient allocation of resources, particularly in large-scale agricultural systems where comprehensive biological sampling is often cost-prohibitive.

2.5 Conclusions

Soil microbial biomass exhibited strong, consistent relationships with commonly measured soil physicochemical properties, supporting the use of these variables as effective proxies for biological condition at the field scale. Organic matter, carbon and nitrogen content, and texture emerged as dominant predictors of microbial biomass, reflecting their role in shaping resource availability and habitat structure. High predictive performance achieved using soil-only variables further demonstrates that routinely measured soil properties capture much of the variability in microbial biomass across sites and depths. Enzyme activity provides additional

functional insight into nutrient-cycling potential but yields only marginal gains in large-scale biomass prediction.

Predictive modeling therefore, offers a viable pathway to reduce analytical costs while maintaining biological relevance, enabling adaptive sampling strategies that balance feasibility with ecological insight. For soil health monitoring programs, these findings support a tiered approach: (1) collect routine physicochemical data at high spatial density, (2) use predictive models to estimate microbial biomass across landscapes, and (3) validate predictions with targeted PLFA analysis at a subset of locations representing key environmental gradients or zones of uncertainty. This framework enables biological insights at operational scales while controlling analytical costs.

These findings provide a foundation for developing scalable soil health assessment frameworks and further demonstrate that reduced predictor sets can achieve comparable predictive performance to full models. This motivates continued exploration of sampling optimization strategies in subsequent chapters.

References

Anderson, T.-H., and K.H. Domsch. 1989. Ratios of microbial biomass carbon to total organic carbon in arable soils. Soil Biol. Biochem. 21(4): 471–479. doi: 10.1016/0038-0717(89)90117-X.

Avellaneda-Torres, L.M., L.M. Melgarejo, C.E. Narváez-Cuenca and J. Sánchez. 2013. Enzymatic activities in potato crop soils under conventional management and in grassland soils. J. Soil Sci. Plant Nutr. 13(2): 301–312. doi: 10.4067/S0718-95162013005000025.

Babyak, M.A. 2004. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. Psychosom. Med. 66(3): 411–421. doi: 10.1097/01.psy.0000127692.23278.a9

Baveye, P.C. 2023. Ecosystem-scale modelling of soil carbon dynamics: Time for a radical shift of perspective? Soil Biol. Biochem. 184: 109112. doi: 10.1016/j.soilbio.2023.109112.

Baveye, P.C. 2025. Why a mechanistic theory of soils is crucially important: Another line of supportive argument exists, seldom invoked in soil science. SOIL 11(2): 1131–1140. doi: 10.5194/soil-11-1131-2025.

Bligh, E.G., and W.J. Dyer. 1959. A rapid method of total lipid extraction and purification. Can. J. Biochem. Physiol. 37(8): 911–917. doi: 10.1139/y59-099

Bolan, N.S., and M.J. Hedley. 2003. Role of carbon, nitrogen, and sulfur cycles in soil acidification. In: Z. Rengel, ed. Handbook of Soil Acidity. Marcel Dekker, New York. p. 29–56. doi: https://doi.org/10.1201/9780203912317.

Bosatta, E., and G.I. Ågren. 1994. Theoretical analysis of microbial biomass dynamics in soils. Soil Biol. Biochem. 26(1): 143–148. Bosatta, E., and G.I. Ågren. 1994. Theoretical analysis of microbial biomass dynamics in soils. Soil Biol. Biochem. 26(1): 143–148. doi: 10.1016/0038-0717(94)90206-2.

Bouma, J. 1989. Using soil survey data for quantitative land evaluation. In: B.A. Stewart, ed. Advances in Soil Science. Vol. 9. Springer, New York. p. 177–213. doi: https://doi.org/10.1007/978-1-4612-3532-3_4.

Breiman, L. 2001. Random Forests. Mach. Learn. 45(1): 5–32. doi: 10.1023/A:1010933404324.

Bünemann, E.K., G. Bongiorno, Z. Bai, R.E. Creamer, G. De Deyn, et al. 2018. Soil quality – A critical review. Soil Biol. Biochem. 120: 105–125. doi: 10.1016/j.soilbio.2018.01.030.

Cardoso, E.J.B.N., R.L.F. Vasconcellos, D. Bini, M.Y.H. Miyauchi, C.A. dos Santos, et al. 2013. Soil health: Looking for suitable indicators. What should be considered to assess the effects of use and management on soil health? Sci. Agric. 70: 274–289. doi: 10.1590/S0103-90162013000400009.

Chen, T., and C. Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA. p. 785–794. doi: 10.1145/2939672.2939785

Cortes, C., and V. Vapnik. 1995. Support-Vector Networks. Mach. Learn. 20(3): 273–297. doi: 10.1023/A:1022627411411.

Cui, J., and N.M. Holden. 2015. The relationship between soil microbial activity, microbial biomass, soil structure, and grassland management. Soil Tillage Res. 146: 32–38. doi: 10.1016/j.still.2014.07.005.

Dalal, R. 1998. Soil microbial biomass—what do the numbers really mean? Aust. J. Exp. Agric. 38: 649–665. doi: 10.1071/EA97142.

de Almeida, R.F., and E.R. Naves. 2015. Soil quality: Enzymatic activity of soil β-glucosidase. Glob. J. Agric. Res. Rev. 3(3): 146–150.

Decaëns, T. 2010. Macroecological patterns in soil communities. Glob. Ecol. Biogeogr. 19(3): 287–302. doi: https://doi.org/10.1111/j.1466-8238.2009.00517.x.

Dick, R.P. 1994. Soil enzyme activities as indicators of soil quality. In: J.W. Doran, D.C. Coleman, D.F. Bezdicek, and B.A. Stewart, eds. Defining Soil Quality for a Sustainable Environment. SSSA Spec. Publ. 35. SSSA, Madison, WI. p. 107–124. doi: 10.2136/sssaspecpub35.c7.

Doran, J.W., and M.R. Zeiss. 2000. Soil health and sustainability: managing the biotic component of soil quality. Appl. Soil Ecol. 15(1): 3–11. doi: 10.1016/S0929-1393(00)00067-6.

Dungait, J.A.J., D.W. Hopkins, A.S. Gregory, and A.P. Whitmore. 2012. Soil organic matter turnover is governed by accessibility not recalcitrance. Glob. Change Biol. 18(6): 1781–1796. doi: 10.1111/j.1365-2486.2012.02665.x.

Ettema, C.H., and D.A. Wardle. 2002. Spatial soil ecology. Trends Ecol. Evol. 17(4): 177–183. doi: 10.1016/S0169-5347(02)02496-5.

Fierer, N., A.S. Grandy, J. Six, and E.A. Paul. 2009. Searching for unifying principles in soil ecology. Soil Biol. Biochem. 41(11): 2249–2256. doi: 10.1016/j.soilbio.2009.06.009.

Fierer, N., and R.B. Jackson. 2006. The diversity and biogeography of soil bacterial communities. Proc. Natl. Acad. Sci. 103(3): 626–631. doi: 10.1073/pnas.0507535103.

Fierer, N., J.P. Schimel, and P.A. Holden. 2003. Variations in microbial community composition through two soil depth profiles. Soil Biol. Biochem. 35(1): 167–176. doi: 10.1016/S0038-0717(02)00251-1.

Fierer, N., S.A. Wood, and C.P. Bueno de Mesquita. 2021. How microbes can, and cannot, be used to assess soil health. Soil Biol. Biochem. 153: 108111. doi: 10.1016/j.soilbio.2020.108111.

Fletcher, R.J., T.J. Hefley, E.P. Robertson, B. Zuckerberg, R.A. McCleery, et al. 2019. A practical guide for combining data to model species distributions. Ecology 100(6): e02710. doi: 10.1002/ecy.2710.

Frank, K., D. Beegle, and J. Denning. 1998. Phosphorus. In: J.R. Brown, ed. Recommended Chemical Soil Test Procedures for the North Central Region. NCR Publ. No. 221 (Revised). Missouri Agric. Exp. Stn., Columbia, MO. p. 21–30.

Franzen, D.W., A.D. Halvorson, and V.L. Hofman. 2000. Management zones for soil N and P levels in the Northern Great Plains. In: P.C. Robert, R.H. Rust, and W.E. Larson, eds. Proc. 5th Int. Conf. Precision Agriculture, Bloomington, MN. 16–19 July 2000. ASA, CSSA, and SSSA, Madison, WI. p. 1–10.

Frostegård, Å., A. Tunlid, and E. Bååth. 1993. Phospholipid Fatty Acid Composition, Biomass, and Activity of Microbial Communities from Two Soil Types Experimentally Exposed to Different Heavy Metals. Appl. Environ. Microbiol. 59(11): 3605–3617. doi: 10.1128/aem.59.11.3605-3617.1993.

García, C., and T. Hernández. 1997. Biological and biochemical indicators in derelict soils subject to erosion. Soil Biol. Biochem. 29(2): 171–177. doi: 10.1016/S0038-0717(96)00294-5.

Gee, G.W., and D. Or. 2002. Particle-size analysis. In: J.H. Dane and G.C. Topp, eds. Methods of Soil Analysis. Part 4. Physical Methods. SSSA Book Ser. 5. SSSA, Madison, WI. p. 255–293. doi: https://doi.org/10.2136/sssabookser5.4.c12.

Geisseler, D., and W.R. Horwath. 2009. Relationship between carbon and nitrogen availability and extracellular enzyme activities in soil. Pedobiologia 53(1): 87–98. doi: 10.1016/j.pedobi.2009.06.002.

Gregorich, E.G., M.R. Carter, D.A. Angers, C.M. Monreal, and B.H. Ellert. 1994. Towards a minimum data set to assess soil organic matter quality in agricultural soils. Can. J. Soil Sci. 74(4): 367–385. doi: 10.4141/cjss94-051.

Grimm, R., E. Behrens, M. Marker, H. Elsenbeer, and I. Dreibrodt. 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis. Geoderma 146(1-2): 102–113. doi: 10.1016/j.geoderma.2008.05.008.

Harden, T., R.G. Joergensen, B. Meyer, and V. Wolters. 1993. Mineralization of straw and formation of soil microbial biomass in a soil treated with simazine and dinoterb. Soil Biol. Biochem. 25(9): 1273–1276. doi: 10.1016/0038-0717(93)90224-Y.

Harris, C.R., K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, et al. 2020. Array programming with NumPy. Nature 585(7825): 357–362. doi: 10.1038/s41586-020-2649-2.

Hassink, J. 1997. The capacity of soils to preserve organic C and N by their association with clay and silt particles. Plant Soil 191(1): 77–87. doi: 10.1023/A:1004213929699.

Hengl, T., M. Nussbaum, M.N. Wright, G.B.M. Heuvelink, and B. Gräler. 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6: e5518. doi: 10.7717/peerj.5518.

Horrigue, W., S. Dequiedt, N. Chemidlin Prévost-Bouré, C. Jolivet, N.P.A. Saby, et al. 2016. Predictive model of soil molecular microbial biomass. Ecol. Indic. 64: 203–211. doi: 10.1016/j.ecolind.2015.12.004.

Horwath, W.R. 1994. Microbial biomass. In: R.W. Weaver, S. Angle, and P. Bottomley, eds. Methods of Soil Analysis. Part 2. Microbiological and Biochemical Properties. SSSA Book Ser. 5. SSSA, Madison, WI. p. 753–773. doi: 10.2136/sssabookser5.2.c36.

Hunter, J.D. 2007. Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 9(3): 90–95. doi: 10.1109/MCSE.2007.55

Jenkinson, D.S. 1988. Determination of microbial biomass carbon and nitrogen in soil. In: J.R. Wilson, ed. Advances in Nitrogen Cycling in Agricultural Ecosystems. CAB International, Wallingford, UK. p. 368–386.

Jenny, H. 1994. Factors of Soil Formation: A System of Quantitative Pedology. Courier Corporation.

Jiang-shan, Z., G. Jian-fen, C. Guang-shui, and Q. Wei. 2005. Soil microbial biomass and its controls. J. For. Res. 16(4): 327–330. doi: 10.1007/BF02858201.

Kuhn, M., and K. Johnson. 2013. Applied Predictive Modeling. Springer, New York. 600 p.

Liu, F., H. Wu, Y. Zhao, D. Li, J.-L. Yang, et al. 2022. Mapping high resolution National Soil Information Grids of China. Sci. Bull. 67(3): 328–340. doi: 10.1016/j.scib.2021.10.013.

Lundberg, S.M., and S.-I. Lee. 2017. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. Curran Associates, Inc. doi: 10.48550/arXiv.1705.07874.

Mallarino, A.P., and D.J. Wittry. 2004. Efficacy of Grid and Zone Soil Sampling Approaches for Site-Specific Assessment of Phosphorus, Potassium, pH, and Organic Matter. Precis. Agric. 5(2): 131–144. doi: 10.1023/B:PRAG.0000022358.24102.1b.

Manzoni, S., and J.P. Schimel. 2024. Advances in modelling soil microbial dynamics. Soil Biol. Biochem. 197: 109535. doi: 10.1016/j.soilbio.2024.109535.

McKinney, W. 2010. Data structures for statistical computing in Python. In: S. van der Walt and J. Millman, eds. Proceedings of the 9th Python in Science Conference. p. 56–61. doi: 10.25080/Majora-92bf1922-00a.

Moore, J.M., S. Klose, and M.A. Tabatabai. 2000. Soil microbial biomass carbon and nitrogen as affected by cropping systems. Biol. Fertil. Soils 31(3): 200–210. doi: 10.1007/s003740050646.

NOAA (National Oceanic and Atmospheric Administration). 2023. U.S. Climate Normals (1991–2020). National Centers for Environmental Information, Asheville, NC. Available at https://www.ncei.noaa.gov/products/land-based-station/us-climate-normals (accessed 15 January 2026).

O'Sullivan, L., F. Bampa, K. Knights, and R.E. Creamer. 2017. Soil protection for a sustainable future: options for a soil monitoring network for Ireland. Soil Use Manag. 33(2): 346–363. doi: 10.1111/sum.12351.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al. 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12: 2825–2830.

Pellegrini, E., R. Nicola, Z. Stefano, I. Franco, N.M. De, et al. 2021. Artificial neural network (ANN) modelling for the estimation of soil microbial biomass in vineyard soils. Biol. Fertil. Soils 57(1): 145–151. doi: 10.1007/s00374-020-01498-1

Philippot, L., C. Chenu, A. Kappler, M.C. Rillig, and N. Fierer. 2024. The interplay between microbial communities and soil properties. Nat. Rev. Microbiol. 22(4): 226–239. doi: 10.1038/s41579-023-00980-5.

Philippot, L., K. Ritz, P. Pandard, S. Hallin, and F. Martin-Laurent. 2012. Standardisation of methods in soil microbiology: progress and challenges. FEMS Microbiol. Ecol. 82(1): 1–10. doi: 10.1111/j.1574-6941.2012.01436.x.

Python Software Foundation. 2024. Python Language Reference, version 3.12. Available at https://www.python.org (accessed 20 January 2026).

Saiya-Cork, K.R., R.L. Sinsabaugh, and D.R. Zak. 2002. The effects of long term nitrogen deposition on extracellular enzyme activity in an Acer saccharum forest soil. Soil Biol. Biochem. 34(9): 1309–1315. doi: 10.1016/S0038-0717(02)00074-3.

Semenov, M.V., A.D. Zhelezova, N.A. Ksenofontova, E.A. Ivanova, D.A. Nikitin, et al. 2025. Microbiological Indicators for Assessing the Effects of Agricultural Practices on Soil Health: A Review. Agronomy 15(2): 335. doi: 10.3390/agronomy15020335.

Sessitsch, A., A. Weilharter, M.H. Gerzabek, H. Kirchmann, and E. Kandeler. 2001. Microbial Population Structures in Soil Particle Size Fractions of a Long-Term Fertilizer Field Experiment. Appl. Environ. Microbiol. 67(9): 4215–4224. doi: 10.1128/AEM.67.9.4215-4224.2001.

Singh, J.S., and V.K. Gupta. 2018. Soil microbial biomass: A key soil driver in management of ecosystem functioning. Sci. Total Environ. 634: 497–500. doi: 10.1016/j.scitotenv.2018.03.373.

Sinsabaugh, R.L., C.L. Lauber, M.N. Weintraub, B. Ahmed, S.D. Allison, et al. 2008. Stoichiometry of soil enzyme activity at the global scale. Ecol. Lett. 11(11): 1252–1264. doi: https://doi.org/10.1111/j.1461-0248.2008.01245.x

Six, J., R.T. Conant, E.A. Paul, and K. Paustian. 2002. Stabilization mechanisms of soil organic matter: Implications for C-saturation of soils. Plant Soil 241(2): 155–176. doi: 10.1023/A:1016125726789.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8(1): 25. doi: 10.1186/1471-2105-8-25.

Sun, T., Y. Wang, M.E. Lucas-Borja, X. Jing, and W. Feng. 2021. Divergent vertical distributions of microbial biomass with soil depth among groups and land uses. J. Environ. Manage. 292: 112755. doi: 10.1016/j.jenvman.2021.112755.

Sünnemann, M., C. Alt, J.E. Kostin, A. Lochner, T. Reitz, et al. 2021. Low-intensity land-use enhances soil microbial activity, biomass and fungal-to-bacterial ratio in current and future climates. J. Appl. Ecol. 58(11): 2614–2625. doi: 10.1111/1365-2664.14004.

Tabatabai, M.A. 1994. Soil enzymes. In: R.W. Weaver, S. Angle, and P. Bottomley, eds. Methods of Soil Analysis. Part 2. Microbiological and Biochemical Properties. SSSA Book Ser. 5. SSSA, Madison, WI. p. 775–833. doi: https://doi.org/10.2136/sssabookser5.2.c37

Templer, P., S. Findlay, and G. Lovett. 2003. Soil microbial biomass and nitrogen transformations among five tree species of the Catskill Mountains, New York, USA. Soil Biol. Biochem. 35(4): 607–613. doi: 10.1016/S0038-0717(03)00006-3.

Tian, L., E. Dell, and W. Shi. 2010. Chemical composition of dissolved organic matter in agroecosystems: Correlations with soil enzyme activity and carbon and nitrogen mineralization. Appl. Soil Ecol. 46(3): 426–435. doi: 10.1016/j.apsoil.2010.09.007.

Tukey, J.W. 1993. Exploratory Data Analysis: Past, Present and Future. doi: https://apps.dtic.mil/sti/html/tr/ADA266775/ (accessed 8 January 2026).

USDA NRCS (Natural Resources Conservation Service). 2024. Web Soil Survey. United States Department of Agriculture, Natural Resources Conservation Service, Washington, DC. Available at https://websoilsurvey.nrcs.usda.gov/ (accessed 10 January 2026).

USDA. 2026. Soil health. USDA, Washington, DC. Available at https://www.usda.gov/about-usda/general-information/initiatives-and-highlighted-programs/peoples-garden/soil-health (accessed 16 Jan. 2026).

van der Heijden, M.G.A., R.D. Bardgett, and N.M. van Straalen. 2008. The unseen majority: Soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. Ecol. Lett. 11(3): 296–310. doi: 10.1111/j.1461-0248.2007.01139.x.

Wadoux, A.M.J.-C., B. Minasny, and A.B. McBratney. 2020a. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Sci. Rev. 210: 103359. doi: https://doi.org/10.1016/j.earscirev.2020.103359

Wadoux, A.M.J.-C., M. Román-Dobarco, and A.B. McBratney. 2021. Perspectives on data-driven soil research. Eur. J. Soil Sci. 72(4): 1675–1689. doi: 10.1111/ejss.13071.

Wadoux, A.M.J.-C., A. Samuel-Rosa, L. Poggio, and V.L. Mulder. 2020b. A note on knowledge discovery and machine learning in digital soil mapping. Eur. J. Soil Sci. 71(2): 133–136. doi: https://doi.org/10.1111/ejss.12909

Waldrop, M.P., D.R. Zak, R.L. Sinsabaugh, M. Gallo, and C. Lauber. 2004. Nitrogen deposition modifies soil carbon storage through changes in microbial enzymatic activity. Ecol. Appl. 14(4): 1172–1177. doi: 10.1890/03-5120.

Wan, X., X. Chen, Z. Huang, and H.Y.H. Chen. 2021. Global soil microbial biomass decreases with aridity and land-use intensification. Glob. Ecol. Biogeogr. 30(5): 1056–1069. doi: 10.1111/geb.13282.

Wardle, D.A. 1992. A Comparative Assessment of Factors Which Influence Microbial Biomass Carbon and Nitrogen Levels in Soil. Biol. Rev. 67(3): 321–358. doi: 10.1111/j.1469-185X.1992.tb00728.x.

Wardle, D.A. 1998. Controls of temporal variability of the soil microbial biomass: A global-scale synthesis. Soil Biol. Biochem. 30(13): 1627–1637. doi: 10.1016/S0038-0717(97)00201-0.

Wardle, D.A., and A. Ghani. 1995. Why is the strength of relationships between pairs of methods for estimating soil microbial biomass often so variable? Soil Biol. Biochem. 27(6): 821–828. doi: 10.1016/0038-0717(94)00229-T.

Warncke, D., and J.R. Brown. 1998. Potassium and other basic cations. In: J.R. Brown, ed. Recommended Chemical Soil Test Procedures for the North Central Region. NCR Publ. No. 221 (Revised). Missouri Agric. Exp. Stn., Columbia, MO. p. 31–33.

Waskom, M.L. 2021. seaborn: statistical data visualization. J. Open Source Softw. 6(60): 3021. doi: 10.21105/joss.03021

Welsch, J., C. Songling, H.L. Buckley, N.J. Lehto, E.E. Jones, et al. 2019. How many samples? Soil variability affects confidence in the use of common agroecosystem soil indicators. Ecol. Indic. 102: 401–409. doi: 10.1016/j.ecolind.2019.02.065.

White, P.M., and C.W. Rice. 2009. Tillage Effects on Microbial and Carbon Dynamics during Plant Residue Decomposition. Soil Sci. Soc. Am. J. 73(1): 138–145. doi: 10.2136/sssaj2007.0384.

Whitney, D.A. 1998. Micronutrients: Zinc, iron, manganese and copper. In: J.R. Brown, ed. Recommended Chemical Soil Test Procedures for the North Central Region. NCR Publ. No. 221 (Revised). Missouri Agric. Exp. Stn., Columbia, MO. p. 41–44.

Wilhelm, R.C., H.M. van Es, and D.H. Buckley. 2022. Predicting measures of soil health using the microbiome and supervised machine learning. Soil Biol. Biochem. 164: 108472. doi: 10.1016/j.soilbio.2021.108472.

Williams, G., R. Miller, and S. Deng. 2025. Dynamic relationships between microbial community, enzyme activity, and soil properties across global ecosystems. Appl. Soil Ecol. 206: 105843. doi: https://doi.org/10.1016/j.apsoil.2024.105843

Yeo, I.-K., and R.A. Johnson. 2000. A new family of power transformations to improve normality or symmetry. Biometrika 87(4): 954--959. doi: https://doi.org/10.1093/biomet/87.4.954

Zelles, L. 1999. Fatty acid patterns of phospholipids and lipopolysaccharides in the characterisation of microbial communities in soil: a review. Biol. Fertil. Soils 29(2): 111–129. doi: 10.1007/s003740050533.

Zhang, S., X. Zhang, E. Bai, X. Guo, X. Zhou, et al. 2025. Climate Change Accelerates Microbial Biomass Accumulation in Soils. Glob. Biogeochem. Cycles 39(12): e2025GB008511. doi: 10.1029/2025GB008511.

Zhu, Y., H. Zhang, Q. Wang, W. Zhu, and Y. Kang. 2024. Soil extracellular enzyme activity linkage with soil organic carbon under conservation tillage: A global meta-analysis. Eur. J. Agron. 155: 127135. doi: 10.1016/j.eja.2024.127135.

Zou, H., and T. Hastie. 2005. Regularization and Variable Selection Via the Elastic Net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67(2): 301–320. doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x

Figures

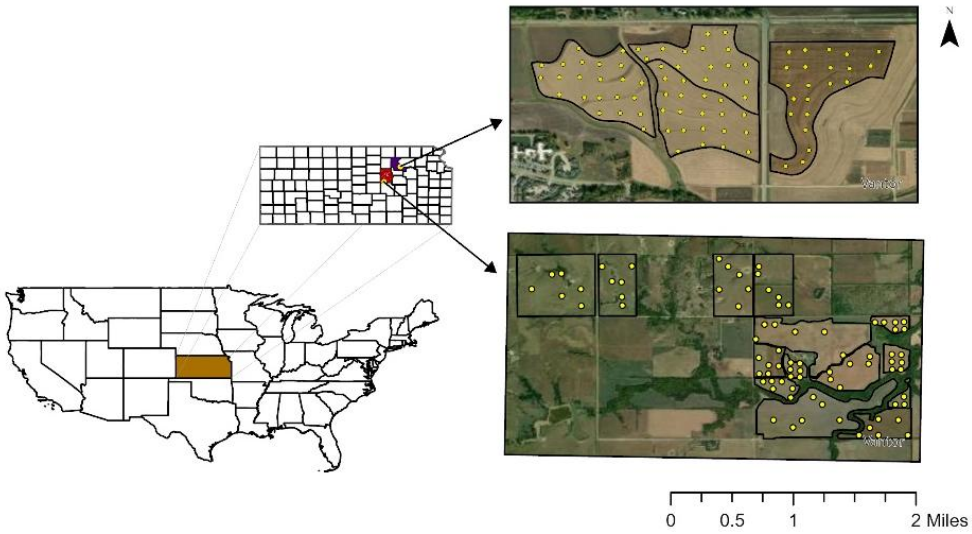

Figure 2.1. Study sites and soil sampling design across Kansas, USA. (A) Agronomy North Farm at Kansas State University, Manhattan, KS. (B) Dickinson County, KS.

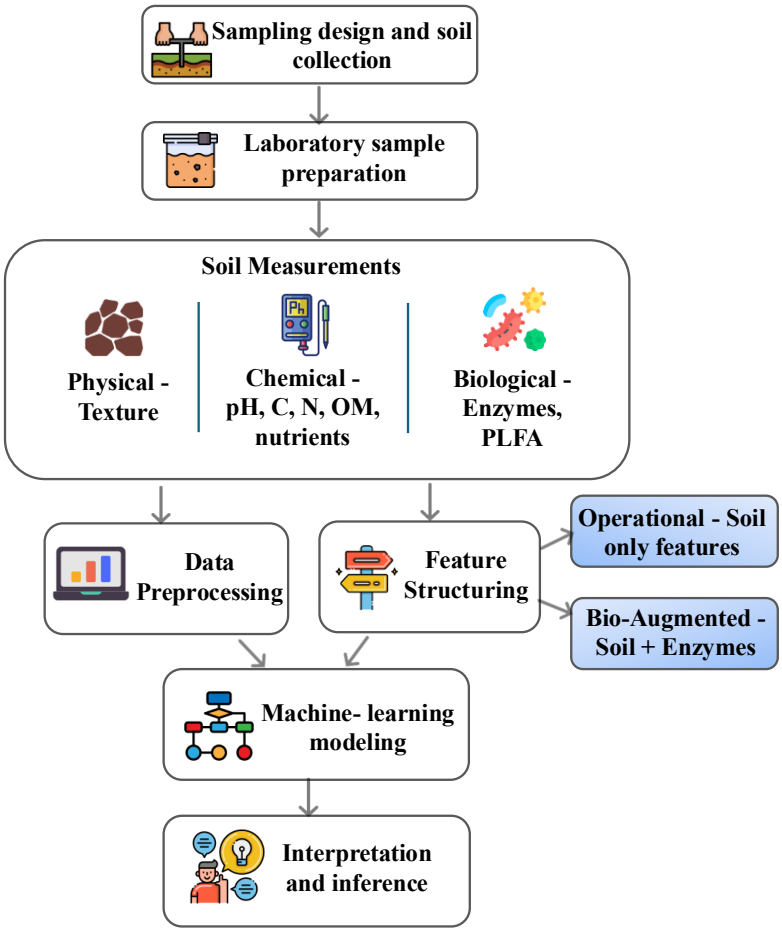

Figure 2.2. Conceptual workflow illustrating soil sampling, laboratory analyses, feature structuring, and machine-learning modeling.

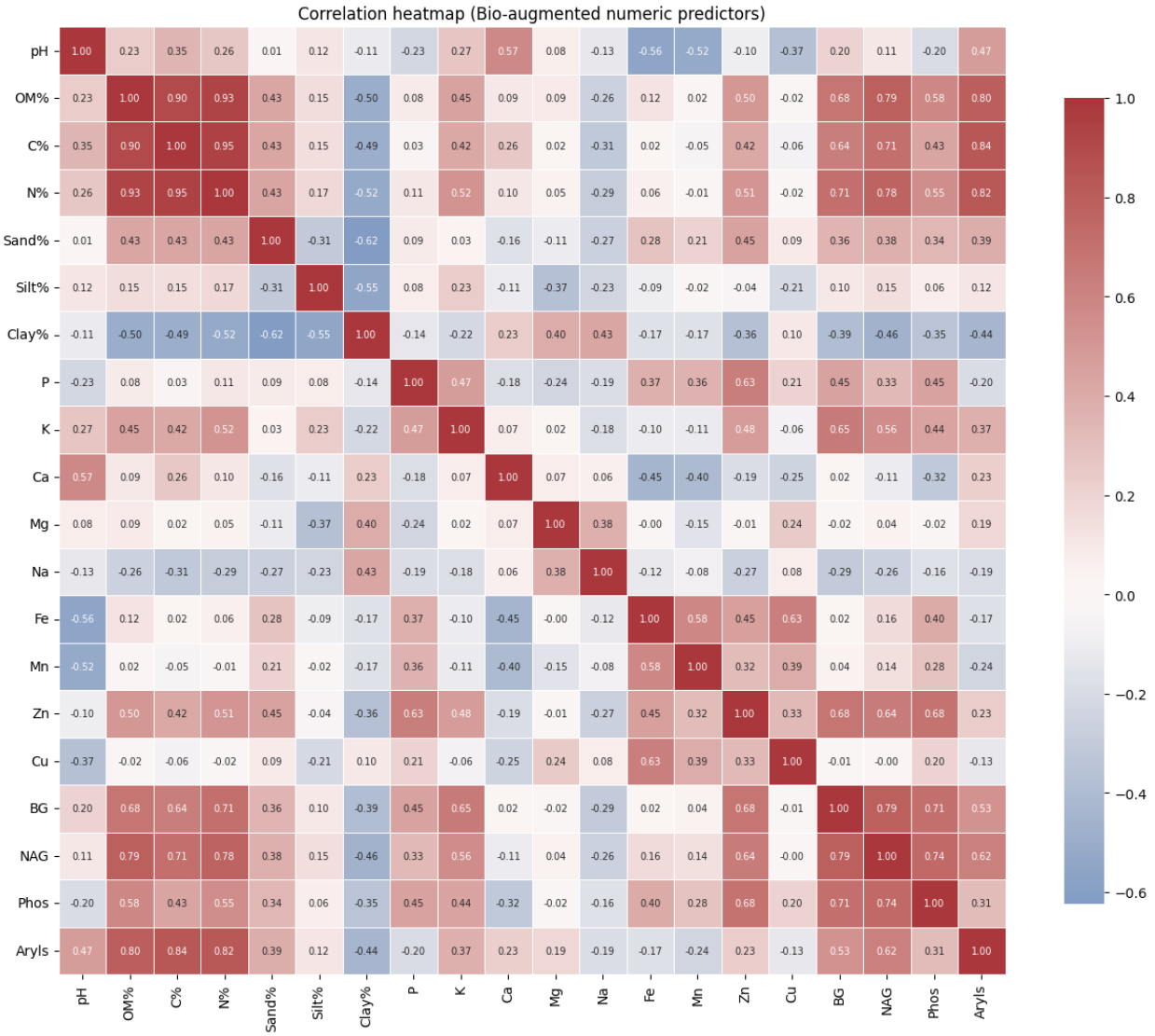

Figure 2.3. Correlation matrix of bio-augmented numeric predictor variables. Variables: pH, organic matter (%), carbon (%C), and nitrogen (%N). Texture: %Sand, %Silt, and %Clay. Chemical properties: phosphorus (P), potassium (K), magnesium (Mg), sodium (Na), calcium (Ca), iron (Fe), manganese (Mn), zinc (Zn), and copper (Cu). Biological indicators: β-glucosidase (βG), arylsulfatase (AS), acid phosphatase (AP), and N-acetyl-β-glucosaminidase (NAG).

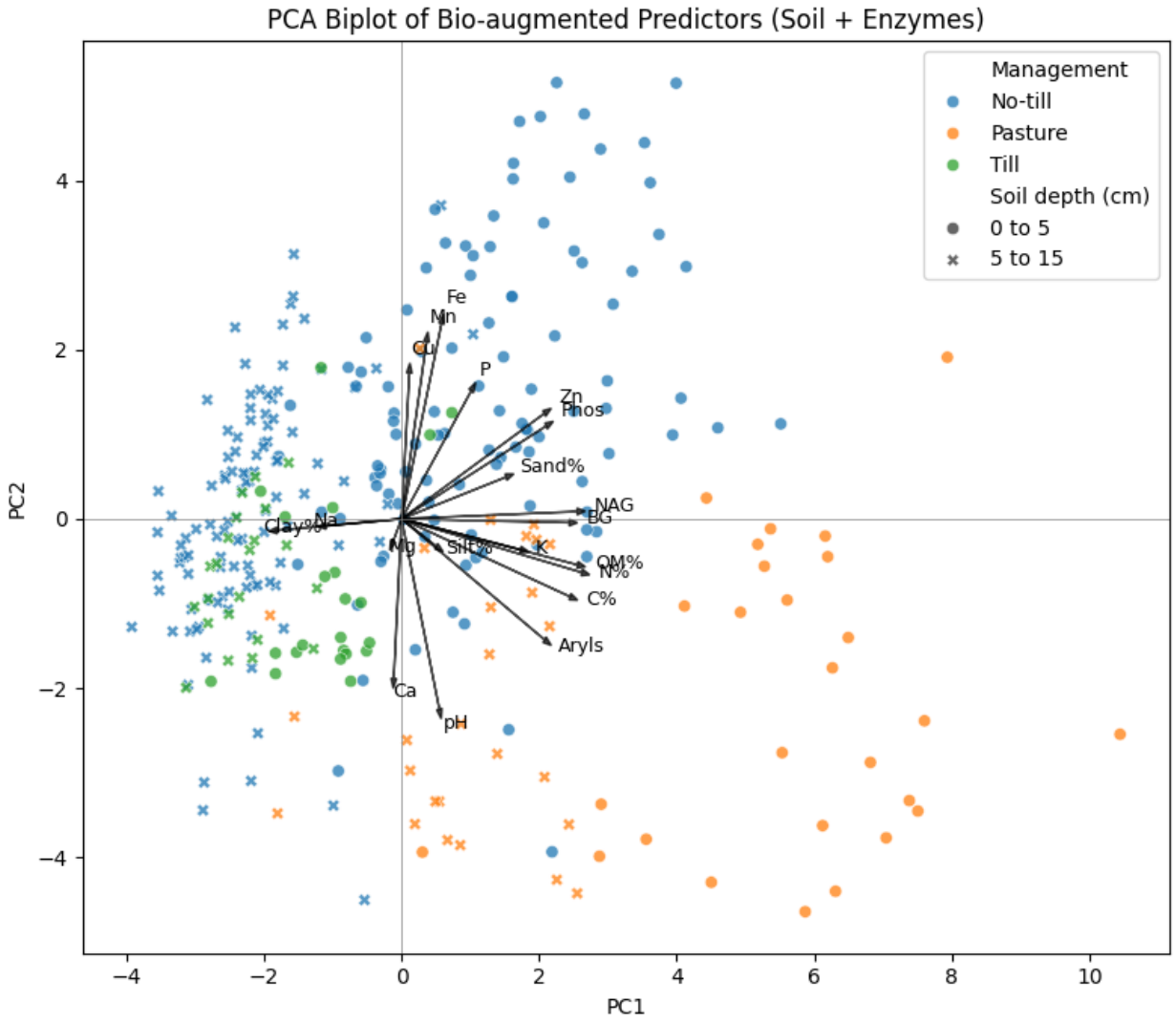

Figure 2.4. Principal component analysis (PCA) of bio-augmented predictor variables. Variables: pH, organic matter (%), carbon (%C), and nitrogen (%N). Texture: %Sand, %Silt, and %Clay. Chemical properties: phosphorus (P), potassium (K), magnesium (Mg), sodium (Na), calcium (Ca), iron (Fe), manganese (Mn), zinc (Zn), and copper (Cu). Biological indicators: β-glucosidase (βG), arylsulfatase (AS), acid phosphatase (AP), and N-acetyl-β-glucosaminidase (NAG).

(Biplot illustrates relationships among soil physicochemical properties, enzyme activities, management systems, and soil depths.)

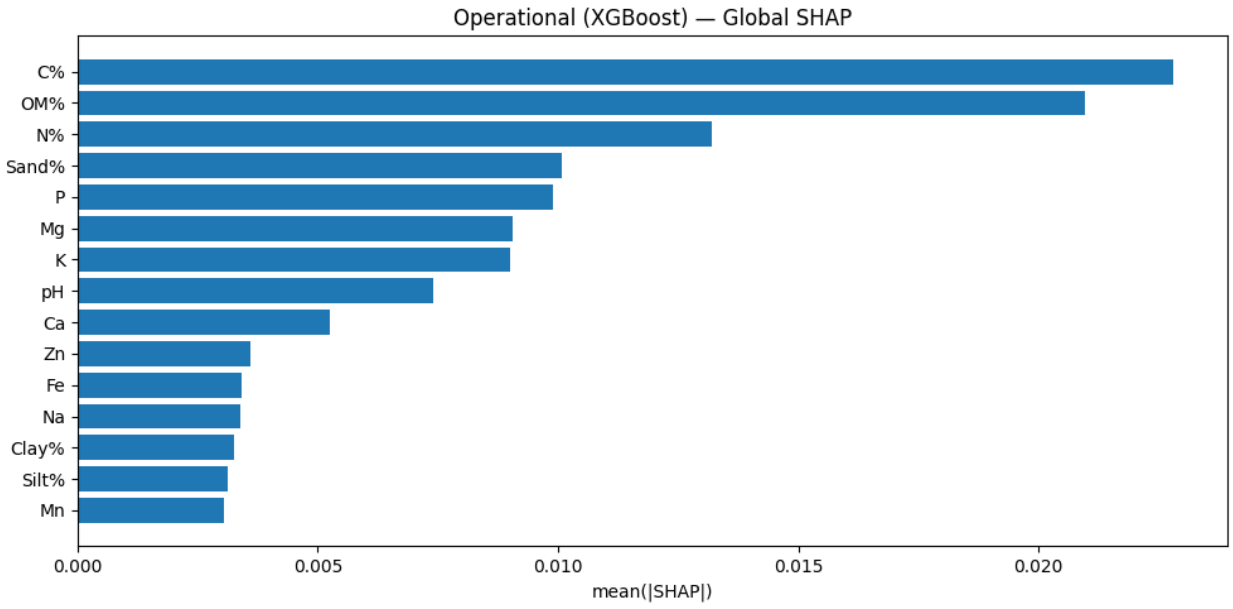

Figure 2.5. Global SHAP (SHapley Additive exPlanations) importance for the XGBoost model predicting soil microbial biomass. Variables: Organic matter (%), carbon (%C), and nitrogen (%N). Texture: %Sand, %Silt, %Clay. Chemical properties: phosphorus (P), magnesium (Mg), sodium (Na), zinc (Zn)

(Bars represent mean absolute SHAP values across cross-validation folds.)

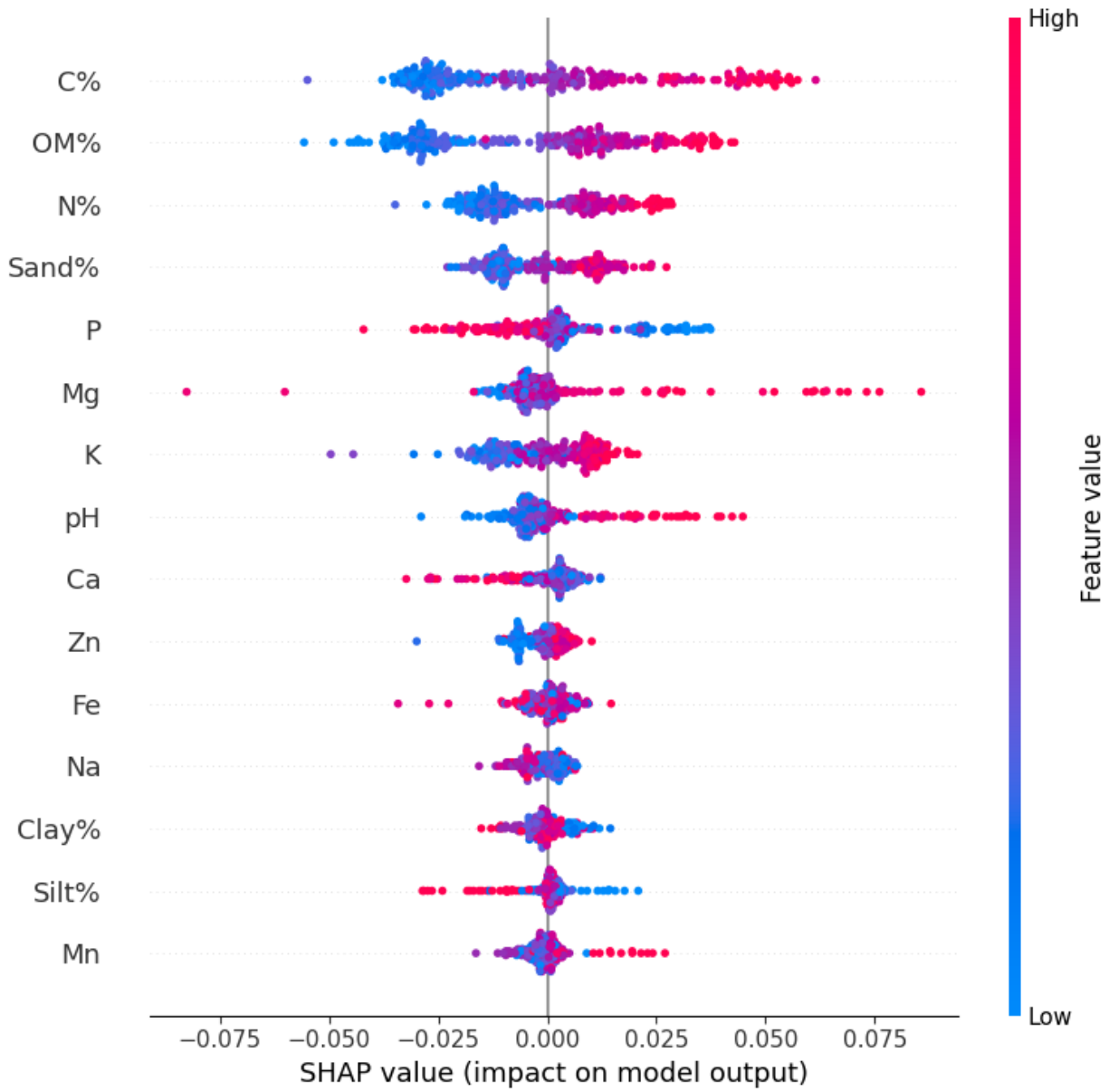

Figure 2.6. SHAP beeswarm plot of predictor effects in the XGBoost model for soil microbial biomass. Variables: pH, organic matter (%), carbon (%C), and nitrogen (%N). Texture: %Sand, %Silt, and %Clay. Chemical properties: phosphorus (P), potassium (K), magnesium (Mg), sodium (Na), iron (Fe), manganese (Mn), zinc (Zn), and copper (Cu), Categorical: Soil depth

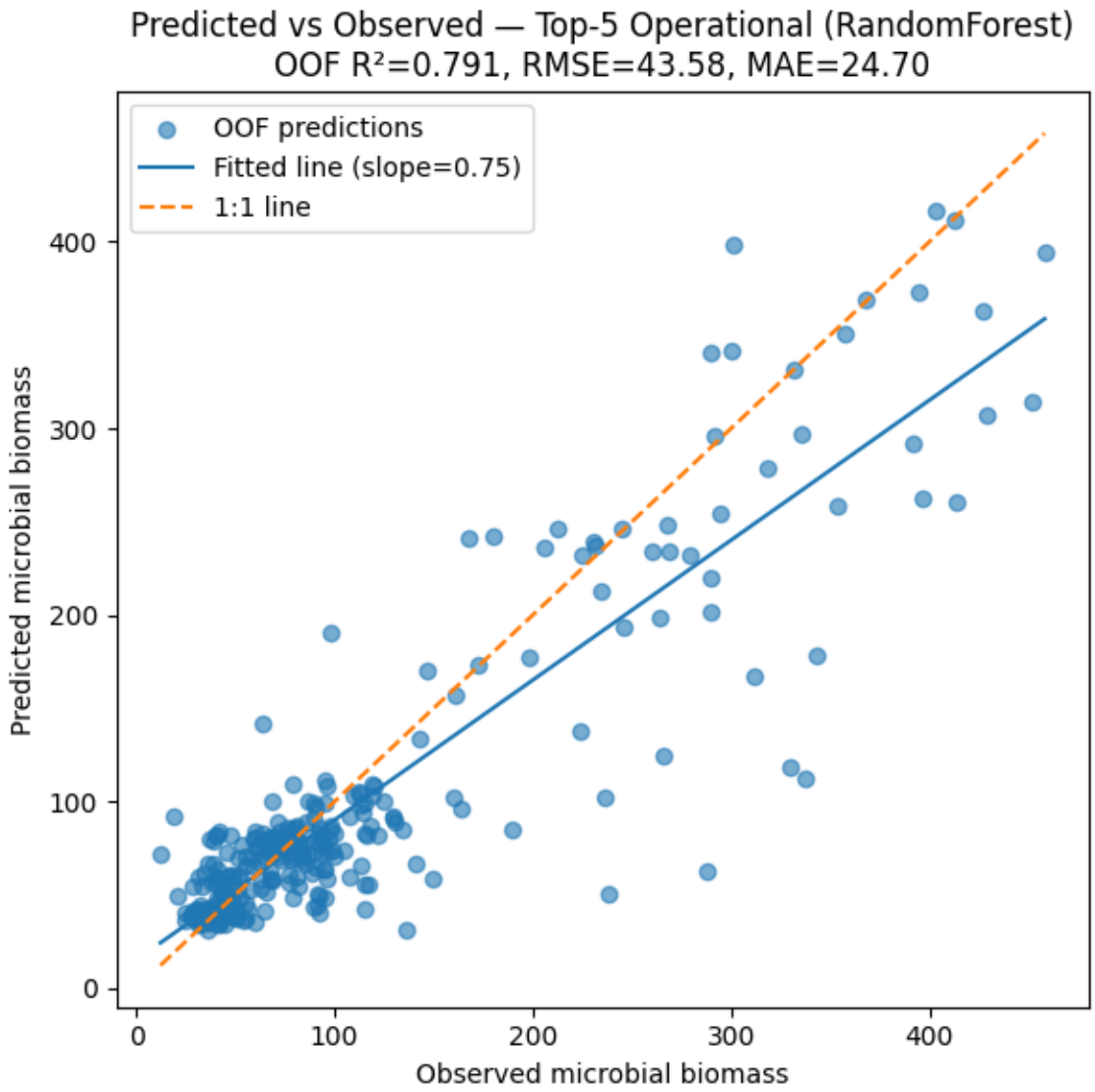

Figure 2.7. Relationship between observed and predicted soil microbial biomass for the Random Forest model using top predictors. Predictions across grouped cross-validation folds.

Tables

Table 2.1. Summary of study sites, management systems, and soil sampling design.

Site	Management System	Sampling Design	Years under management	Number of cores	Sampling depths (cm)
Kansas State University Ag Innovation Farm (North Farm), Riley County, KS	Continuous no-till (3 fields), Conventional tillage (1 field)	Grid-based sampling (1 composite sample per acre)	~22 years (no-till), variable for conventional field	81	0-5, 5-15
Dickinson County, KS	Pasture (continuous grazing), No-till cropland	Zone-based sampling (SOC and pH–derived management zones)	Long-term pasture, ~20 years (no-till cropland)	75	0-5, 5-15

Table 2.2. Soil physicochemical, biological, and categorical predictor variables used for machine-learning modeling.

Category	Variable	Units	Analytical method	Feature set
Chemical	Soil organic carbon (SOC)	g kg^{-1}	Dry combustion (elemental analyzer)	Soil-only, Bio-augmented
	Total nitrogen (TN)	g kg^{-1}	Dry combustion (elemental analyzer)	
	Soil pH	unitless	1:1 soil:water slurry	
	Electrical conductivity (EC)	dS m^{-1}	Conductivity meter (soil extract)	
	Available phosphorus (P)	mg kg^{-1}	Bray-1 / Mehlich-3 extraction	
	Exchangeable potassium (K)	mg kg^{-1}	Ammonium acetate extraction	
	Calcium (Ca)	mg kg^{-1}	ICP following extraction	
	Magnesium (Mg)	mg kg^{-1}	ICP following extraction	
Physical	Sand	%	Hydrometer method	Soil-only, Bio-augmented
	Silt	%		
	Clay	%		
Categorical	Soil Depth	Cm	Field sampling (0–5, 5–15)	Soil-only, Bio-augmented
Biological	β-glucosidase activity	μmol pNP g^{-1} h^{-1}	Colorimetric enzyme assay	Bio-augmented only
	N-Acetyl glucosaminidase	μmol pNP g^{-1} h^{-1}	Colorimetric enzyme assay	
	Acid Phosphatase	μmol pNP g^{-1} h^{-1}	Colorimetric enzyme assay	
	Aryl-sulphatase	μmol pNP g^{-1} h^{-1}	Colorimetric enzyme assay	

Table 2.3. Machine-learning models and optimized hyperparameter configurations used for soil microbial biomass prediction.

ML Model	Optimized Hyperparameters	Predictor Configuration
Random Forest (RF)	n_estimators = 500 max_depth = 15 min_samples_split = 5 min_samples_leaf = 2 random_state = 42	**Soil-only feature set:** operational soil physicochemical properties + soil depth **Bio-augmented feature set:** soil-only predictors plus enzyme activity variables
XGBoost	n_estimators = 500 max_depth = 6 learning_rate = 0.05 subsample = 0.8 colsample_bytree = 0.8	
Elastic Net	alpha = 0.1 l1_ratio = 0.5 max_iter = 10,000	
Support Vector Regression (SVR)	kernel = “rbf” C = 10 epsilon = 0.1 gamma = “scale”	
k-Nearest Neighbors (KNN)	n_neighbors = 15 weights = “distance” p = 2	

Table 2.4. Predictive performance of machine-learning models across operational features (Mean ± SD across 5-fold cross-validation).
Models: Random Forest (RF), XGBoost, Elastic Net, Support Vector Regression (SVR), and k-Nearest Neighbors (KNN).

Model	Mean R^2	R^2 SD	Mean RMSE	RMSE SD	Mean MAE	MAE SD
XGBoost	0.805	0.057	41.5	5.72	23.9	1.96
KNN	0.788	0.082	43.1	9.48	24.2	3.50
Random Forest	0.787	0.068	43.3	6.55	24.2	2.19
SVM	0.305	0.193	78.2	10.98	44.5	6.41
ElasticNet	0.296	0.056	79.6	3.65	44.9	2.10

R2: coefficient of determination, SD: standard deviation, RMSE: root mean square error, MAE: mean absolute error.

Table 2.5. Predictive performance metrics for all machine-learning models across bio-augmented features (Mean ± SD across 5-fold cross-validation). Models: Random Forest (RF), XGBoost, Elastic Net, Support Vector Regression (SVR), and k-Nearest Neighbors (KNN).

Model	Mean R^2	R^2 SD	Mean RMSE	RMSE SD	Mean MAE	MAE SD
Random Forest	0.874	0.036	33.4	5.21	20.2	1.76
XGBoost	0.863	0.040	34.8	5.28	20.2	2.09
KNN	0.815	0.059	40.4	7.30	22.6	2.66
ElasticNet	0.390	0.063	74.0	3.26	41.4	2.13
SVM	0.258	0.143	81.3	7.29	45.5	3.28

R2: coefficient of determination, SD: standard deviation, RMSE: root mean square error, MAE: mean absolute error.

Table 2.6. Predictive performance metrics for all machine-learning models across the top 5 operational features (Mean ± SD across 5-fold cross-validation). Models: Random Forest (RF), XGBoost, Elastic Net, Support Vector Regression (SVR), and k-Nearest Neighbors (KNN).

Model	Mean R^2	R^2 SD	Mean RMSE	RMSE SD	Mean MAE	MAE SD
Random Forest	0.788	0.072	43.0	7.01	24.7	2.02
XGBoost	0.771	0.083	44.7	8.38	27.0	3.12
KNN	0.770	0.085	44.7	8.20	26.6	3.03
SVM	0.403	0.212	71.7	10.80	42.5	4.21
ElasticNet	0.296	0.056	79.6	3.65	44.9	2.10

Table 2.7. Permutation feature importance for XGBoost model (for operational feature set)

Feature	Mean Importance%
C%	29.27
N%	21.59
P	11.32
OM%	9.44
pH	7.73
Sand%	7.09
K	4.53
Mg	4.10
Na	2.01
Zn	1.69
Mn	1.44
Clay%	1.00
Fe	0.74
Soil depth (0–5 cm)	0.44
Soil depth (5–15 cm)	0.01

Variables: pH, organic matter (%), carbon (%C), and nitrogen (%N). Texture: %Sand, %Silt, and %Clay. Chemical properties: phosphorus (P), potassium (K), magnesium (Mg), sodium (Na), iron (Fe), manganese (Mn), zinc (Zn), and copper (Cu), Categorical - depth.

Table 2.8. SHAP (SHapley Additive exPlanations) values for XGBoost model (operational feature set)

Feature	Mean absolute SHAP
C%	0.023
OM%	0.021
N%	0.013
Sand%	0.01
P	0.01
Mg	0.009
K	0.009
pH	0.007
Ca	0.005
Zn	0.004
Fe	0.003
Na	0.003
Clay%	0.003
Silt%	0.003
Mn	0.003

Variables: pH, organic matter (%), carbon (%C), and nitrogen (%N). Texture: %Sand, %Silt, and %Clay. Chemical properties: phosphorus (P), potassium (K), magnesium (Mg), sodium (Na), iron (Fe), manganese (Mn), zinc (Zn), and copper (Cu)

Chapter 3 - Optimizing field scale soil sampling for soil health assessment

Abstract

Integrating biological soil health indicators into routine agricultural monitoring requires sampling strategies that balance spatial resolution with practical feasibility. This study evaluated whether optimized, reduced-intensity sampling approaches can reliably capture field-scale variability in key soil health indicators while substantially reducing sampling effort.

Two complementary approaches were assessed across agricultural sites in Kansas. At a high-resolution reference site (North Farm), a dense 1-acre grid was systematically thinned to simulate coarser sampling intensities. At a second site (Dickinson County), management zone-based sampling was implemented using soil organic carbon (SOC) and pH gradients. Soil health indicators included microbial biomass (PLFA), enzyme activities (β-glucosidase and N-acetylglucosaminidase), SOC, pH, and aggregate stability. Spatial patterns were evaluated using variogram analysis, and reduced sampling designs were compared against high-resolution reference data.

Moderate reductions in sampling density, up to 1 point per 5 acres, preserved field-scale means and dominant spatial patterns for most indicators, with variogram analysis confirming that spatial dependence remained largely intact. A six-point zone-based design estimated field-scale microbial biomass with less than 2% bias and performed consistently across zoning approaches. Indicator-specific differences were evident: microbial biomass exhibited stable and interpretable spatial patterns aligned with SOC gradients, whereas enzyme activities showed greater fine-scale variability and uncertainty.

These results demonstrate that sampling design can be strategically aligned with the spatial behavior of individual soil health indicators. Reduced-intensity and zone-based approaches provide a practical and quantitative framework for integrating biological measurements into scalable soil health monitoring and precision agriculture systems.

3.1 Introduction

3.1.1 Spatial variability of soil health indicators

Agricultural soil exhibits substantial spatial variability arising from the interaction of soil-forming factors, environmental gradients, and management practices across nested spatial scales. (Wilding and Drees, 1983; Trangmar et al., 1986; Lin et al., 2005). This variability is not random; soil properties display structured spatial patterns driven by topography, texture, and landscape position, with additional fine-scale heterogeneity introduced by biological processes and microsite conditions. (Burrough, 1993; Webster and Oliver, 2007). Because these processes operate across multiple spatial scales, no single sampling resolution can fully capture the complexity of soil variability. (Trangmar et al., 1986; Franklin and Mills, 2003; Lin et al., 2005). Geostatistical tools, particularly semivariogram analysis, provide a quantitative framework for characterizing these spatial structures by quantifying the distance over which soil properties remain correlated (Burgess and Webster, 1980; Cambardella et al., 1994).

Among soil health indicators, biological properties, particularly microbial biomass and enzyme activities, exhibit especially pronounced spatial heterogeneity because they are acutely sensitive to substrate availability, moisture, temperature, and local environmental conditions. (Nunan et al., 2002; Franklin and Mills, 2003; Baldrian, 2014). Soil microbial biomass (SMB) plays a central role in organic matter decomposition and nutrient cycling, and has been established as a sensitive integrative indicator of soil health (Kallenbach and Grandy, 2011;

Lehmann et al., 2020). Critically, microbial biomass is not uniformly distributed across fields; it tends to mirror the spatial patterns of soil organic carbon, pH, and topography, which together regulate substrate supply, moisture retention, and microbial habitat quality (Fierer et al., 2009; Liu et al., 2010; Constancias et al., 2015). Extracellular enzymes, by contrast, respond more rapidly to short-term environmental fluctuations and can persist in the soil matrix independently of living microbial cells, conferring greater spatial and temporal variability than biomass (Baldrian, 2014; Nayar Afaq Kirmani et al., 2017). The differing spatial behavior of these two indicator classes, stable biomass versus dynamic enzyme activities, has direct implications for sampling design and is a central focus of this chapter.

3.1.2 Challenges in soil sampling design

Accurate characterization of soil health depends on sampling strategies capable of resolving spatial variability at ecologically relevant scales. Dense, uniform grid sampling offers high spatial resolution but is constrained by analytical costs, labor requirements, and the capacity limitations of biological assays, making it impractical for large-scale or routine monitoring programs. (Weindorf and Zhu, 2010; Lawrence et al., 2020a). This constraint is particularly acute for biological measurements; techniques such as phospholipid fatty acid (PLFA) analysis and enzyme assays require specialized laboratory infrastructure and are substantially more resource-intensive than standard chemical soil tests (Bünemann et al., 2018; Mann et al., 2019; Fan et al., 2022). As a result, biological soil health assessments are typically limited to small numbers of samples, reducing spatial coverage and limiting their utility for field-scale inference. Reduced-intensity sampling approaches, including systematic grid thinning and management zone (MZ) based allocation, have been proposed as practical alternatives that balance spatial coverage against logistical feasibility (Minasny and McBratney, 2007; Lawrence et al., 2020a).

Management zones are delineated using proxy variables, most commonly soil organic carbon (SOC), pH, and apparent electrical conductivity, that integrate long-term management effects and correlate with biological activity, thereby concentrating sampling effort in areas representing distinct soil conditions (Fleming et al., 2000; Mallarino and Wittry, 2004). For chemical properties such as phosphorus and potassium, MZ sampling has demonstrated competitive or superior performance relative to uniform grid approaches (Mallarino and Wittry, 2004; Valente et al., 2024). However, whether these strategies translate effectively to biologically mediated indicators, which exhibit greater spatial and temporal heterogeneity than chemical properties, remains insufficiently tested (Peigné et al., 2009; Gyawali et al., 2023).

3.1.3 Gaps in existing studies

Despite growing adoption of precision agriculture tools and geostatistical methods, several critical gaps persist in the literature. Most studies evaluating reduced sampling performance have focused on stable chemical properties (e.g., phosphorus, potassium, pH), which typically exhibit stronger spatial structure and lower temporal variability than biological indicators (Rodrigues et al., 2012; Lawrence et al., 2020). Studies that do include biological indicators seldom compare grid-thinning and zone-based approaches within the same framework or apply both statistical and geostatistical validation simultaneously (Peigné et al., 2009; Gyawali et al., 2023). Furthermore, the evidence base for optimal sampling of biologically mediated indicators, including whether field-scale mean values and spatial structure can both be preserved under reduced intensity, remains weak, particularly in the mixed agricultural landscapes typical of the central Great Plains (Smith and Halvorson, 2011; Nayar Afaq Kirmani et al., 2017).

There is therefore a need for systematic, empirically grounded evaluation of reduced-intensity sampling strategies for soil biological indicators using a rigorous multi-metric

framework that assesses both central tendency and spatial structure. The present study addresses this need using two Kansas agricultural sites that differ in size, land use, and sampling design, enabling comparison of grid-based and zone-based approaches for a suite of chemical, biological, and physical soil health indicators. To address these limitations, this study integrates statistical and geostatistical evaluation across contrasting field sites to systematically assess the performance of reduced-intensity sampling strategies for soil health indicators.

3.1.4 Objectives

The objectives of this chapter were

1. Quantify how a reduction in sampling density affects field-scale estimates of soil health indicators, testing whether mean values and spatial structure are preserved relative to a dense 1-acre grid reference.
2. Distinguish the sensitivity of biologically mediated indicators (microbial biomass, enzyme activities) to sampling intensity compared to more spatially stable chemical and physical properties.
3. Evaluate management zone-based sampling as a cost-effective alternative to uniform grid sampling, assessing whether strategically allocated samples capture dominant spatial gradients while reducing sampling effort.
4. Develop practical guidelines linking sampling design to trade-offs between cost, accuracy, and information retention for soil health monitoring programs incorporating biological indicators.

3.2 Materials and methods

3.2.1 Study sites and experimental design

This study evaluated soil sampling optimization strategies across two contrasting agricultural sites in Kansas. The first site, Kansas State University Regenerative Agriculture Innovation Farm (North Farm) in Riley County, encompasses approximately 80 acres under row-crop production. The field exhibits moderate topographic and soil variability typical of rainfed cropland in the central Great Plains. Its manageable size permitted implementation of a dense, uniform grid sampling design, 80 sampling points distributed at 1-acre intervals. This high-resolution data set served as our reference for evaluating the effects of reduced sampling density. The second site, Terrace Lane Farm in Dickinson County, spans several hundred acres, making dense grid sampling impractical. This landscape includes both pasture and no-till cropland, resulting in pronounced within-field heterogeneity. Here we applied a management-zone-based approach, strategically allocating samples within predefined zones rather than across a uniform grid.

Despite differences in size and land use, both sites lie within the same regional agroecological zone, allowing for across-site comparison while minimizing confounding climatic or pedogenic influences. Dominant soil series at North Farm includes Smolan silt loam, Wymore silty clay loam, and Reading silt loam. The Dickinson County site consists primarily of Irwin and Clime series with silt loam to silty clay loam textures (USDA NRCS, 2024). This dual-site design enables evaluation of sampling optimization under both high-resolution reference datasets and operational, large-field conditions, enhancing the generalizability of the findings.

3.2.2 Soil sampling

Grid based sampling at North Agronomy Farm

We established an 80-point sampling grid across the Regenerative Agriculture Innovation Farm, from now on just referred to as RA Farm, using georeferenced coordinates generated in ArcGIS Pro and uploaded them to RTK GPS for field navigation. At each location, multiple cores were collected within a small radius and composited by depth to minimize microscale variability while preserving field-scale spatial contrasts.

Samples were collected at three depth intervals (0-5, 5-15, and 15-30 cm); however, all analyses reported here focus on the surface 0-5 cm layer. This depth receives the strongest management influence and exhibits the greatest spatial variability in soil health indicators at the field scale. Restricting analyses to a single depth also avoided confounding vertical gradients when comparing sampling strategies.

Simulating sampling density reduction

To quantify information loss associated with reduced sampling intensity, we systematically thinned the RA Farm grid dataset. Rather than collecting new samples, we generated nested subsets of the original 80 points, effectively simulating coarser sampling densities while maintaining uniform spatial coverage. Locations were removed in a structured manner to avoid clustering and preserve spatial representativeness.

This approach assumes that the dense grid provides a sufficiently high-resolution approximation of the underlying spatial variability, allowing subsampling to serve as a proxy for reduced sampling designs without introducing additional measurement errors. Each reduced-density dataset retained original georeferenced coordinates, enabling direct spatial comparison against the full grid reference.

Management zone sampling at Dickinson County site

At the Dickinson County site, we combined proximal sensing with targeted soil testing to guide biological sampling at substantially lower intensity than grid-based approaches. A Veris scanning probe generated apparent electrical conductivity (ECa) maps across each field. Concurrently, we collected routine soil samples on an approximate 2.5-acre grid and analyzed soil organic carbon and pH.

These datasets informed the delineation of management zones that reflect spatial patterns in soil conditions. We selected soil organic carbon and pH as zoning variables because they capture long-term management effects and strongly correlate with biological activity. From each field, six locations were selected for biological sampling (microbial biomass and enzyme activities), positioned to capture the range of conditions across zones rather than to redundantly sample similar areas. Management zones were delineated using k-means clustering applied to the combined spatial layers of soil organic carbon and soil pH. The resulting clusters were interpreted as low, medium, and high-productivity zones to guide biological sampling.
Field and laboratory protocols matched those applied at North Farm, with analyses restricted to 0-5 cm depth for cross-site consistency.

3.2.3 Laboratory analyses

Following field collection, samples were transported to the laboratory and homogenized by depth. Visible plant materials and stones were removed. Subsamples for routine chemical analyses were air-dried, and those designated for biological analyses were refrigerated until processing.

Soil microbial biomass

Total microbial biomass was quantified via phospholipid fatty acid (PLFA) analysis using a modified Bligh and Dyer lipid extraction (Bligh and Dyer, 1959) following established

protocols (White and Rice, 2009). We selected PLFA-derived biomass because phospholipids degrade rapidly following cell death, providing an index of living microbial communities that remains sensitive to management-induced shifts in soil conditions (Frostegård et al., 1993; Zelles, 1999).

Freeze-dried subsamples were extracted, and phospholipids were separated using silicic acid chromatography, saponified with potassium hydroxide, and methylated to form fatty acid methyl esters (FAMEs). FAMEs were analyzed on a Thermo Scientific Trace GC-ISQ GC-MS equipped with a DB-5MS capillary column (30 m × 250 μm i.d. × 0.25 μm film thickness). Peaks were identified using bacterial acid methyl ester (BAME) standard mix (Matreya 1114) supplemented with authentic standards. Quantification employed methyl nonadecanoate (19:0 FAME) as an internal standard. Total microbial biomass (TMB), expressed as nmol PLFA g^{-1} dry soil, was calculated as the sum of identified biomarkers.

Soil chemical properties

Soil pH was measured by the Kansas State University Soil Testing Laboratory, while soil organic carbon (SOC) was determined by dry combustion using a LECO TruSpec CN Analyzer.

Enzyme activities

Potential activities of β-glucosidase (βG) and N-acetyl-glucosaminidase (NAG), extracellular enzymes involved in carbon and nitrogen cycling, were quantified using colorimetric assays with p-nitrophenyl (PNP) linked substrates (Tabatabai, 1994). Briefly, 0.5 g of soil was incubated with the appropriate buffer and substrate at 37°C for 1 h. Controls accounted for background absorbance. Reactions were terminated with $CaCl_2$ and alkaline stop solution, and absorbance was measured at 400 nm (Genesys 50 UV-visible spectrophotometer).

Activities were quantified using p-nitrophenol calibration curves and expressed as μg PNP g^{-1} soil h^{-1}.

Water stable aggregates

The aggregate size distribution was determined by wet sieving, following Mikha and Rice (2004). Air-dried soil was gently separated along natural breaks and passed through an 8-mm sieve. We placed 50 g samples on stacked sieves (2000 and 250 μm), submerged them for 10 min (slaking), then oscillated vertically (4 cm stroke, 0.5 Hz) for 10 min using a Yoder-type apparatus. Material retained on each sieve was collected, allowed to settle, and oven-dried (60°C, 72 h). Material passing through larger sieves was subsequently filtered through 53 and 20 μm sieves, with retained fractions similarly dried and weighed. Aggregate size fractions were calculated as percentages of total dry mass after sand correction.

3.2.4 Spatial data processing

All spatial datasets were processed in ArcGIS Pro using consistent projection, extent, and resolution. Apparent EC data were screened for outliers and mapped. Routine soil test data collected on the 2.5-acre grid were georeferenced and interpolated to continuous raster surfaces representing field-scale gradients in soil organic carbon and pH. These layers were aligned with EC maps to delineate management zones.

Raster values of EC, (SOC), and pH were extracted from all sampling locations: dense grid points, reduced-density subsets, and zone-based biological sampling sites, creating a standardized spatial framework for cross-strategy comparison.

3.2.5 Evaluation framework

Sampling optimization was evaluated using a combined statistical and spatial framework designed to assess both central tendency and spatial structure. The dense 1-acre grid at North

Farm was treated as an empirical reference for field-scale variability. Although not a true ground truth, its high spatial resolution provides the most complete representation of field conditions available in this study. For each soil health indicator (SOC, pH, TMB, enzyme activities, and aggregate stability), we calculated summary statistics, mean, standard deviation, and coefficient of variation across the dense grid reference, reduced-density grid subsets, and management-zone datasets.

Spatial representations were generated for each sampling strategy to evaluate preservation of dominant field-scale patterns. We visually compared reduced-density and zone-based datasets with the dense-grid reference to assess agreement in spatial gradients and identification of high- and low-value regions across the field.

Information retention under alternative sampling strategies was evaluated by examining deviations in summary statistics and spatial distributions relative to the dense grid reference. Emphasis was placed on the ability of reduced and zone-based approaches to capture overall variability and major spatial features while minimizing sampling intensity.

This combined statistical and spatial framework quantified trade-offs between sampling effort and representation of soil health variability, revealing whether reduced-intensity strategies could approximate dense grid characterization while offering practical advantages in cost and logistical feasibility.

Zone validation and sensitivity analysis

Zone-based sampling performance was validated using the North Farm dense grid dataset under a transparent and reproducible framework. RA Farm grid points were assigned to three zones (Low, Medium, High) based on SOC and pH at 0-5 cm. Two zoning methods were evaluated: (i) quantile-based zones, where a composite zone score was calculated as the sum of

standardized SOC and pH values and then divided into tertiles, and (ii) k-means clustering (k=3) applied to standardized SOC and pH, with clusters ordered and labeled Low-High by mean composite score.

To simulate a practical zone-sampling design, we repeatedly drew a 6-point sample from the dense grid using a fixed allocation of zones (2 High, 2 Medium, 2 Low). This resampling procedure was repeated 200 times for each zoning method using a fixed random seed to ensure reproducibility.

For each draw and soil health indicator, subset means were calculated, and percent bias was computed relative to the dense-grid reference mean as:

$$\text{Percent Bias} = 100 * \frac{(Subset\ mean - Reference\ mean)}{(Reference\ mean)} \quad \text{(Equation 1)}$$

Zone-sampling performance was summarized using the median percent bias and interquartile range (IQR) across resamples, thereby quantifying both the accuracy and uncertainty of the zone-based estimates.

3.2.6 Statistical and GIS analysis

Spatial processing and visualization were conducted in ArcGIS Pro (Esri, Redlands, CA). Geographic datasets were projected to a common coordinate system, and raster/vector layers were aligned to ensure consistent resolution and extent. Descriptive statistics and data organization were done using Python (Python Software Foundation) with associated scientific libraries. Figures and tables were generated using Python-based visualization libraries. Consistent workflows across sampling strategies ensured reproducibility and direct comparability.

3.3 Results

3.3.1 Sampling intensity analysis

Reducing sampling density from the 1-acre reference grid to coarser intensities resulted in a continuous decline in the number of sampling locations across the field (Table 3.1). The mean number of points decreased from 81 under the reference grid to approximately 25 sampling locations when intensity was reduced to one point per five acres. Despite this substantial reduction in sampling density, summary statistics of soil indicators remained stable across sampling intensities.

Across sampling intensities, mean values of most soil properties deviated only slightly from the 1-acre reference mean (Fig 3.2). Microbial biomass, SOC, organic matter, macroaggregate stability, and soil pH remained within a narrow range of reference values across all thinning scenarios. Enzyme activities exhibited somewhat larger fluctuations, particularly β-glucosidase, which had modest increases at intermediate sampling intensities. However, even in these cases, deviations remained relatively small compared with the variability observed within the dense grid.

The spatial structure of soil properties also remained largely consistent as sampling density decreased (Fig 3.3). Nugget-to-sill ratios derived from variogram models showed only moderate changes across sampling intensities. Soil pH consistently exhibited the lowest nugget-to-sill ratios, indicating strong spatial dependence across the field. In contrast, macroaggregate stability and enzyme activities displayed moderate spatial structure, with slightly higher nugget contributions. Importantly, reductions in sampling density did not substantially alter these spatial relationships.

These combined results indicate that moderate reductions in sampling density preserved both the central tendency and spatial structure of key soil health indicators. Even at lower sampling intensities, field-scale patterns remained broadly consistent with those observed under the dense grid reference. These findings suggest that sampling intervals remain within the effective spatial range of most soil properties, preserving spatial autocorrelation despite reduced sampling density.

3.3.2 Zone-based sampling

Soil biological indicators exhibited measurable variation across management zones delineated using soil organic carbon and pH gradients (Table 3.2). Mean microbial biomass increased from the Low zone (92.38 ± 32.01 nmol PLFA g^{-1} soil) to the High zone (102.15 ± 21.04 nmol PLFA g^{-1} soil), with the medium zone showing intermediate values. Although variability within zones remained substantial, this trend suggests that the zoning framework captured broad spatial differences in soil biological activity across the field.

Enzyme activities exhibited weaker and less consistent differentiation across zones compared with microbial biomass and macroaggregate stability. β-glucosidase activity was around 12% lower in the Medium zone compared to both the Low and High zones. Macroaggregate stability increased gradually from Low (0.647 ± 0.158) to High (0.694 ± 0.118), indicating modest improvements in soil structural stability across the zoning gradient.

The distribution of microbial biomass across zones further illustrates these patterns (Fig 3.4). Median biomass values tended to increase from Low to High zones, although substantial overlap among zones indicates that within-zone variability remained considerable. Despite this overlap, the zoning approach captured meaningful differences in central tendencies across the

field, suggesting that management zones reflect underlying gradients in soil biological and structural properties.

The zone-based sampling framework captured broad spatial variability in soil health indicators while considerably reducing the number of sampling locations required compared with dense-grid approaches.

3.3.3 Validation of zone-based sampling

The performance of the six-point zone-based sampling design was evaluated by comparing subset estimates against the dense-grid reference means. Across soil indicators, median per cent bias remained relatively small for most variables, indicating that zone-distributed sampling captured field-scale averages with reasonable accuracy (Table 3.3; Fig 3.5).

Microbial biomass estimates had minimal bias under both zoning approaches. Median bias was 1.86% using k-means zones and -0.22% under quantile zoning, showing that six zone-distributed samples closely approximated the dense-grid reference mean. Macroaggregate stability also exhibited low bias (-3.76% for k-means and -0.16% for quantile zoning).

Enzyme activities had greater deviations comparatively, especially for β-glucosidase. Under k-means zoning, β-glucosidase had a median bias of -8.27%, whereas the quantile approach produced a lower positive bias (1.6%). In contrast, N-acetylglucosaminidase showed minimal bias under both zoning methods, with median deviations remaining within ±2%.

The variability of subset estimates differed among soil indicators (Fig 3.6). β-glucosidase and N-acetylglucosaminidase had the greatest uncertainty, with interquartile ranges exceeding 20%. In comparison, macroaggregate stability had substantially lower variability (IQR ≈ 11-13%), while microbial biomass hadmoderate variability (IQR ≈ 15-19%).

The six-point sampling design reproduced dense-grid estimates with relatively low bias for most soil indicators. Differences between zoning methods were generally small, suggesting that the accuracy of zone-based sampling was robust to the specific approach used for zone delineation.

3.4 Discussion

3.4.1 Sampling density and the preservation of field-scale spatial structure

The findings of this study demonstrate that moderate reductions in sampling density from a 1-acre grid to coarser intensities (up to 1 point per 5 acres) preserved field-scale mean values for most soil health indicators, including microbial biomass, soil organic carbon, organic matter, microaggregate stability, and pH. Variogram-derived nugget-to-sill ratios remained largely stable across sampling intensities, indicating that the underlying spatial dependence of soil properties was not substantially disrupted by thinning the grid. These findings align with previous studies demonstrating that spatially structured soil properties can be reliably characterized with reduced sampling density, provided that sampling intervals remain within the range of spatial dependence (Burgess and Webster, 1980; Trangmar et al., 1986; Minasny and McBratney, 2007; Lawrence et al., 2020).

The contrasting behavior of soil pH, which consistently exhibited the lowest nugget-to-sill ratios, reflecting strong spatial dependence, versus macroaggregate stability and enzyme activities, which showed moderate spatial structure with higher nugget contributions, reflects fundamental differences in the scales at which controlling processes operate. Soil pH integrates long-term soil-forming processes and management history that vary smoothly across the landscape, whereas biological indicators and structural properties respond to more localized controls such as root distribution, residue inputs, and microbial activity (Katsalirou et al., 2010;

Baldrian, 2014; Nayar Afaq Kirmani et al., 2017).The relative stability of pH spatial structure under grid thinning is therefore expected, while the somewhat greater sensitivity of biological indicators to sampling intensity reflects their inherently finer-scale and more dynamic spatial organization (Franklin and Mills, 2003; Ritz et al., 2004)

Importantly, while field-scale means and broad spatial patterns were well preserved under reduced sampling intensity, the present study did not assess fine-scale spatial heterogeneity at resolutions below the 1-acre reference. Soil biological properties are known to exhibit strong variability at sub-meter to decimeter scales associated with microbial hotspots, root proximity, and aggregate structure (Nunan et al., 2002; Franklin and Mills, 2003; Stark et al., 2004). Grid thinning necessarily sacrifices resolution at these scales, and this trade-off should be explicitly considered when the objective is precision mapping rather than field-scale characterization.

3.4.2 Biological interpretation of zone patterns

The observed zonal patterns provide ecological insight into the drivers of spatial variability in soil biological indicators. The consistent increase in microbial biomass from Low to High management zones reflects the underlying gradient in soil organic carbon that delineates those zones. Soil organic carbon is the primary energy and substrate source for microbial growth, and its positive relationship with microbial biomass is among the most robust and generalizable patterns in soil ecology, documented across diverse ecosystems and management systems (Fierer et al., 2009; Kallenbach and Grandy, 2011; Constancias et al., 2015; Chen et al., 2016). The co-occurrence of higher macroaggregate stability in the High zone reinforces this interpretation. The water-stable aggregates protect organic matter from decomposition and provide a stable physical habitat for microbial communities, creating a positive feedback between soil structure, carbon retention, and microbial biomass (Six et al., 2004; Kallenbach and Grandy, 2011). The gradients

observed in this study, therefore, likely reflect real, management-mediated differences in soil condition rather than sampling artifacts, lending ecological validity to the zone delineation approach.

The weaker, less consistent differentiation of enzyme activities across zones is ecologically meaningful and deserves explicit interpretation. Unlike microbial biomass, which integrates accumulated substrate and community size over months to years, extracellular enzyme activities represent instantaneous functional expression that responds rapidly to temperature, moisture, and substrate pulse (Shi et al., 2013; Baldrian, 2014). A significant fraction of enzyme activity in soils can persist after microbial cell death and is associated with mineral surfaces and organic matter, effectively decoupling enzyme activity from living biomass on short timescales (Baldrian, 2014). The spatial pattern of β-glucosidase, was highest in the High zone but lowest in the Medium rather than Low zone, which may reflect non-linear relationships between carbon availability, substrate stoichiometry, and enzyme induction that are not captured by the zoning variables (organic carbon and pH). These results underscore the caution warranted when interpreting enzyme activities as direct proxies for microbial community conditions, particularly in spatially heterogeneous systems.

3.4.3 Reliability of reduced-intensity sampling for biological indicators

The six-point zone-based sampling design achieved low median per cent bias for microbial biomass (1.86% under k-means; -0.22% under quantile zoning) and macroaggregate stability (-3.76% and -0.16%, respectively), indicating that strategically distributed zone samples can closely approximate dense-grid estimates of field-scale means for these indicators. The consistency of results across both zoning methods further suggests that performance was robust to the specific algorithm used for zone delineation, an encouraging finding for practical

implementation, where the choice of zoning method often depends on data availability and local expertise (Mallarino and Wittry, 2004; Valente et al., 2024).

By contrast, β-glucosidase showed substantially higher uncertainty across both zoning approaches (IQR exceeding 28%), and its median bias under k-means zoning (-8.27%) exceeded the ±5% threshold often used as a benchmark for acceptable sampling error in precision agriculture. This result was consistent with the known sensitivity of extracellular enzymes to short-term environmental conditions, their fine-scale spatial heterogeneity, and the possibility that zone boundaries defined by organic carbon and pH do not adequately stratify enzyme activity distributions (Baldrian, 2014; Gyawali et al., 2023). These findings suggest that for applications requiring precise field-scale estimates of enzyme activities, the six-point zone design may be insufficient, and either higher sampling intensity or repeated measurements across seasonal windows would be needed to reduce uncertainty (Bardgett et al., 1999; Shi et al., 2013). Taken together, these results establish a clear hierarchy of sampling requirements among soil health indicators: stable chemical properties (pH, carbon) can be estimated with moderate grid thinning; microbial biomass and macroaggregate stability are well-represented by zone-based approaches; and highly dynamic enzyme activities require either higher density or temporally replicated sampling to achieve comparable accuracy. This hierarchy has direct practical implications for designing tiered soil health monitoring programs that allocate resources proportionally to the variability characteristics of each indicator (Rodrigues et al., 2012; Stone et al., 2016; Lehmann et al., 2020). This distinction highlights that sampling design must be indicator-specific rather than universally applied, particularly when incorporating biologically mediated variables into soil health frameworks.

3.4.4 Implications of soil health monitoring

The finding that zone-based sampling with as few as six points can approximate field-scale means for microbial biomass with a median bias of under 2% provides practical validation that the Chapter 2 predictor space can be captured cost-effectively in field settings. This represents a meaningful contribution to operationalizing predictive soil health assessment: accurate prediction of PLFA biomass from routine properties is only practically useful if those routine properties can themselves be measured efficiently and representatively.

From a broader monitoring perspective, the demonstrated feasibility of zone-based biological sampling supports the integration of microbial biomass measurements into precision agriculture workflows at costs competitive with those of dense grid approaches. Proximal sensing tools such as apparent electrical conductivity (ECa) mapping, combined with routine soil chemical data, provide a practical basis for zone delineation without requiring dense biological pre-sampling (Adamchuk et al., 2004; Valente et al., 2024). For regional or national monitoring programs seeking to include biological indicators, the zone-based approach offers a structured, reproducible, and scalable sampling framework that can be implemented without specialized geostatistical expertise (Bünemann et al., 2018; Lehmann et al., 2020). This linkage between sampling design and predictive modeling represents a critical step toward operationalizing soil biological indicators within precision agriculture and large-scale soil health monitoring frameworks.

3.4.5 Limitations and future directions

This study relied on a single sampling campaign, limiting the ability to assess temporal variability in soil biological properties. Soil microbial biomass and enzyme activities are known to fluctuate seasonally in response to temperature, moisture, and plant carbon inputs (Bardgett et

al., 1999; Shi et al., 2013), and the spatial patterns documented here represent a snapshot that may not fully capture temporal dynamics. Incorporating repeated sampling across seasons or years, an approach enabled by the reduced-intensity design validated here would substantially strengthen confidence in the generalizability of zone-based estimates.

The dense 1-acre grid at North Farm served as the empirical reference for this study, but it does not constitute an absolute ground truth: variability at sub-acre scales remains uncharacterized, and the reference itself is subject to the same temporal constraints as the other sampling designs. Future work incorporating independent validation datasets, finer-scale sampling windows, or high-resolution proximal sensing would enable more rigorous assessment of the true detection limits of reduced-intensity approaches.

Finally, this study was conducted in a single agroecological region (central Kansas), and the performance of reduced sampling strategies will vary with the strength and range of spatial dependence of local soil properties, which differ across landscapes with contrasting pedogenic histories, climates, and management intensities (Trangmar et al., 1986; Lin et al., 2005; Lawrence et al., 2020). Expanding this evaluation framework to contrasting agroecosystems, particularly those with more heterogeneous biological communities or under different tillage and cover-cropping regimes, would improve the generalizability of recommendations for sampling intensity and zone design.

3.5 Conclusion

This study evaluated the effectiveness of reduced sampling strategies for characterizing spatial variability in soil health indicators across agricultural landscapes. Results demonstrate that moderate reductions in sampling density preserved field-scale mean values and dominant spatial patterns for key soil properties, including microbial biomass, soil organic carbon, and pH.

Variogram analysis further confirmed that spatial dependence was largely maintained under reduced sampling intensity, indicating that essential field-scale structure can be captured with fewer observations.

Zone-based sampling further improved efficiency by strategically targeting areas representing distinct soil conditions. The six-point zone-based design produced low bias in estimating field-scale means for most soil indicators and performed consistently across zoning approaches, demonstrating its potential as a practical alternative to dense grid sampling in heterogeneous systems.

Importantly, the study revealed clear differences among soil health indicators in their sensitivity to sampling design. Microbial biomass exhibited relatively stable and interpretable spatial patterns, whereas enzyme activities showed greater variability and uncertainty, reflecting their dynamic and fine-scale nature. These findings highlight the need for indicator-specific sampling strategies, particularly when incorporating biologically mediated processes into soil health assessments.

Overall, the results demonstrate that optimized sampling approaches can substantially reduce sampling effort while maintaining reliable estimates of soil health indicators. By explicitly linking sampling design to the spatial behavior of different indicators, this study provides a quantitative framework for integrating biological measurements into scalable soil health monitoring systems, supporting more efficient and scientifically robust evaluation of sustainable agricultural practices.

References

Adamchuk, V.I., J.W. Hummel, M.T. Morgan, and S.K. Upadhyaya. 2004. On-the-go soil sensors for precision agriculture. *Comput. Electron. Agric.* 44(1): 71–91. doi: https://doi.org/10.1016/j.compag.2004.03.002

Baldrian, P. 2014. Distribution of extracellular enzymes in soils: Spatial heterogeneity and determining factors at various scales. *Soil Sci. Soc. Am. J.* 78(1): 11–18. doi: https://doi.org/10.2136/sssaj2013.04.0155dgs

Bardgett, R.D., R.D. Lovell, P.J. Hobbs, and S.C. Jarvis. 1999. Seasonal changes in soil microbial communities along a fertility gradient of temperate grasslands. *Soil Biol. Biochem.* 31(7): 1021–1030 doi: https://doi.org/10.1016/S0038-0717(99)00016-4

Bligh, E.G., and W.J. Dyer. 1959. A rapid method of total lipid extraction and purification. *Can. J. Biochem. Physiol.* 37(8): 911–917. doi: https://doi.org/10.1139/o59-099

Bünemann, E.K., G. Bongiorno, Z. Bai, R.E. Creamer, G. De Deyn, et al. 2018. Soil quality—A critical review. *Soil Biol. Biochem.* 120: 105–125. doi: https://doi.org/10.1016/j.soilbio.2018.01.030

Burgess, T.M., and R. Webster. 1980. Optimal interpolation and isarithmic mapping of soil properties: I. The semivariogram and punctual kriging. *J. Soil Sci.* 31(2): 315–331. doi: https://doi.org/10.1111/j.1365-2389.1980.tb02084.x

Burrough, P.A. 1993. Soil variability: A late 20th century view. *Soils Fert.* 56: 529–562

Cambardella, C.A., T.B. Moorman, J.M. Novak, T.B. Parkin, D.L. Karlen, et al. 1994. Field-scale variability of soil properties in central Iowa soils. *Soil Sci. Soc. Am. J.* 58(5): 1501–1511. doi: https://doi.org/10.2136/sssaj1994.03615995005800050033x

Chen, Y.-L., J.-Z. Ding, Y.-F. Peng, F. Li, G.-B. Yang, et al. 2016. Patterns and drivers of soil microbial communities in Tibetan alpine and global terrestrial ecosystems. *J. Biogeogr.* 43(10): 2027–2039. doi:: https://doi.org/10.1111/jbi.12806

Constancias, F., S. Terrat, N.P.A. Saby, W. Horrigue, J. Villerd, et al. 2015. Mapping and determinism of soil microbial community distribution across an agricultural landscape. *MicrobiologyOpen* 4(3): 505–517. doi: https://doi.org/10.1002/mbo3.255

Fan, Y., X. Wang, T. Funk, I. Rashid, B. Herman, et al. 2022. A critical review for real-time continuous soil monitoring: Advantages, challenges, and perspectives. *Environ. Sci. Technol.* 56(19): 13546–13564. doi: 10.1021/acs.est.2c03562

Fierer, N., M.S. Strickland, D. Liptzin, M.A. Bradford, and C.C. Cleveland. 2009. Global patterns in belowground communities. *Ecol. Lett.* 12(11): 1238–1249. doi: https://doi.org/10.1111/j.1461-0248.2009.01360.x

Fleming, K.L., D.G. Westfall, D.W. Wiens, and M.C. Brodahl. 2000. Evaluating farmer-defined management zone maps for variable rate fertilizer application. *Precis. Agric.* 2(2): 201–215. doi: https://doi.org/10.1023/A:1011481832064

Franklin, R.B., and A.L. Mills. 2003. Multi-scale variation in spatial heterogeneity for microbial community structure in an agricultural field. *FEMS Microbiol. Ecol.* 44(3): 335–346. doi: https://doi.org/10.1016/S0168-6496(03)00074-6

Frostegård, Å., A. Tunlid, and E. Bååth. 1993. Phospholipid fatty acid composition and microbial biomass in soils exposed to heavy metals. *Appl. Environ. Microbiol.* 59(11): 3605–3617. doi: https://doi.org/10.1128/aem.59.11.3605-3617.1993

Gyawali, A.J., H. Neely, J. Foster, C. Neely, K. Lewis, et al. 2023. Sampling for biological indicators of soil health: How does sampling methodology affect results? *Geoderma* 435: 116513. doi: https://doi.org/10.1016/j.geoderma.2023.116513

Kallenbach, C., and A.S. Grandy. 2011. Controls over soil microbial biomass responses to carbon amendments: A meta-analysis. *Agric. Ecosyst. Environ.* 144(1): 241–252. doi: https://doi.org/10.1016/j.agee.2011.08.020

Katsalirou, E., S. Deng, D.L. Nofziger, A. Gerakis, and S.D. Fuhlendorf. 2010. Spatial structure of microbial biomass in prairie soils. *Eur. J. Soil Biol.* 46(3): 181–189. doi: https://doi.org/10.1016/j.ejsobi.2010.04.005

Lawrence, P.G., W. Roper, T.F. Morris, and K. Guillard. 2020. Guiding soil sampling strategies using classical and spatial statistics: A review. *Agron. J.* 112(1): 493–510. doi: https://doi.org/10.1002/agj2.20048

Lehmann, J., D.A. Bossio, I. Kögel-Knabner, and M.C. Rillig. 2020. The concept and future prospects of soil health. *Nat. Rev. Earth Environ.* 1(10): 544–553. doi: https://doi.org/10.1038/s43017-020-0080-8

Lin, H., D. Wheeler, J. Bell, and L. Wilding. 2005. Assessment of soil spatial variability at multiple scales. *Ecol. Model.* 182(3): 271–290. doi: https://doi.org/10.1016/j.ecolmodel.2004.04.006

Liu, S., Y. Li, J. Wu, D. Huang, Y. Su, et al. 2010. Spatial variability of soil microbial biomass in subtropical soils. *Soil Sci. Plant Nutr.* 56(5): 693–704. doi: https://doi.org/10.1111/j.1747-0765.2010.00510.x

Mallarino, A.P., and D.J. Wittry. 2004. Efficacy of grid and zone soil sampling approaches for site-specific assessment of phosphorus, potassium, pH, and organic matter. *Precis. Agric.* 5(2): 131–144. doi: https://doi.org/10.1023/B:PRAG.0000022358.24102.1b

Mann, C., D. Lynch, S. Fillmore, and A. Mills. 2019. Relationships between field management, soil health, and microbial community composition. *Appl. Soil Ecol.* 144: 12–21. doi: https://doi.org/10.1016/j.apsoil.2019.06.012

Minasny, B., and A.B. McBratney. 2007. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. *Geoderma* 140(4): 324–336. doi: https://doi.org/10.1016/j.geoderma.2007.04.028

Nunan, N., K. Wu, I.M. Young, J.W. Crawford, and K. Ritz. 2002. Spatial patterns of soil bacteria. *Microb. Ecol.* 44(4): 296–305. doi: 10.1007/s00248-002-2021-0

Peigné, J., J.-F. Vian, M. Cannavacciuolo, B. Bottollier, and R. Chaussod. 2009. Soil sampling based on the spatial variability of microbial indicators. *Eur. J. Soil Biol.* 45(5): 488–495. doi: https://doi.org/10.1016/j.ejsobi.2009.09.002

Python Software Foundation. 2024. Python Language Reference, version 3.12. Available at https://www.python.org (accessed 20 January 2026).

Ritz, K., J.W. McNicol, N. Nunan, S. Grayston, P. Millard, et al. 2004. Spatial structure in soil chemical and microbiological properties in an upland grassland. *FEMS Microbiol. Ecol.* 49(2): 191–205. doi: https://doi.org/10.1016/j.femsec.2004.03.005

Rodrigues, M.S., J.E. Corá, and C. Fernandes. 2012. Soil sampling intensity and spatial distribution pattern of soil attributes and corn yield in a no-tillage system. *Eng. Agríc.* 32: 852–865. doi: https://doi.org/10.1590/S0100-69162012000500005

Shi, Y., R. Lalande, C. Hamel, N. Ziadi, B. Gagnon, et al. 2013. Seasonal variation of microbial biomass, activity, and community structure under different tillage and phosphorus management practices. *Biol. Fertil. Soils* 49(7): 803–818. doi: https://doi.org/10.1007/s00374-013-0773-y

Six, J., H. Bossuyt, S. Degryze, and K. Denef. 2004. Aggregates, soil biota, and soil organic matter dynamics. *Soil Tillage Res.* 79(1): 7–31. doi: https://doi.org/10.1016/j.still.2004.03.008

Smith, J.L., and J.J. Halvorson. 2011. Spatial variability of soil quality indicators. *Appl. Environ. Soil Sci.* 2011: 198737. doi: https://doi.org/10.1155/2011/198737

Stark, C.H.E., L.M. Condron, et al. 2004. Small-scale spatial variability of soil properties. *Soil Biol. Biochem.* 36(4): 601–608. doi: https://doi.org/10.1016/j.soilbio.2003.12.005

Stone, D., K. Ritz, B.G. Griffiths, A. Orgiazzi, and R.E. Creamer. 2016. Selection of biological indicators appropriate for European soil monitoring. *Appl. Soil Ecol.* 97: 12–22. doi: https://doi.org/10.1016/j.apsoil.2015.08.005

Tabatabai, M.A. 1994. Soil enzymes. In: Weaver, R.W., et al. (eds.), Methods of Soil Analysis, Part 2. SSSA Book Series No. 5. Soil Science Society of America, Madison, WI. p. 775–833.

Trangmar, B.B., R.S. Yost, and G. Uehara. 1986. Application of geostatistics to spatial studies of soil properties. *Adv. Agron.* 38: 45–94

USDA NRCS (Natural Resources Conservation Service). 2024. Web Soil Survey. United States Department of Agriculture, Natural Resources Conservation Service, Washington, DC. Available at https://websoilsurvey.nrcs.usda.gov/ (accessed 10 January 2026).

Valente, D.S.M., et al. 2024. Accuracy of sampling techniques for precision agriculture. *Agriculture* 14(12)

Webster, R., and M.A. Oliver. 2007. Geostatistics for environmental scientists. 2nd ed. Wiley, Chichester, UK.

Weindorf, D.C., and Y. Zhu. 2010. Spatial variability of soil properties and sampling strategy implications. *Pedosphere* 20(2): 185–197. doi: https://doi.org/10.1016/S1002-0160(10)60006-9

White, P.M., and C.W. Rice. 2009. Tillage effects on microbial and carbon dynamics. *Soil Sci. Soc. Am. J.* 73(1): 138–145. doi: https://doi.org/10.2136/sssaj2007.0384

Wilding, L.P., and L.R. Drees. 1983. Spatial variability and pedology. In: *Pedogenesis and Soil Taxonomy*. Elsevier

Zelles, L. 1999. Fatty acid patterns for microbial community characterization. *Biol. Fertil. Soils* 29(2): 111–129. doi: 10.1007/s003740050533.

Figures

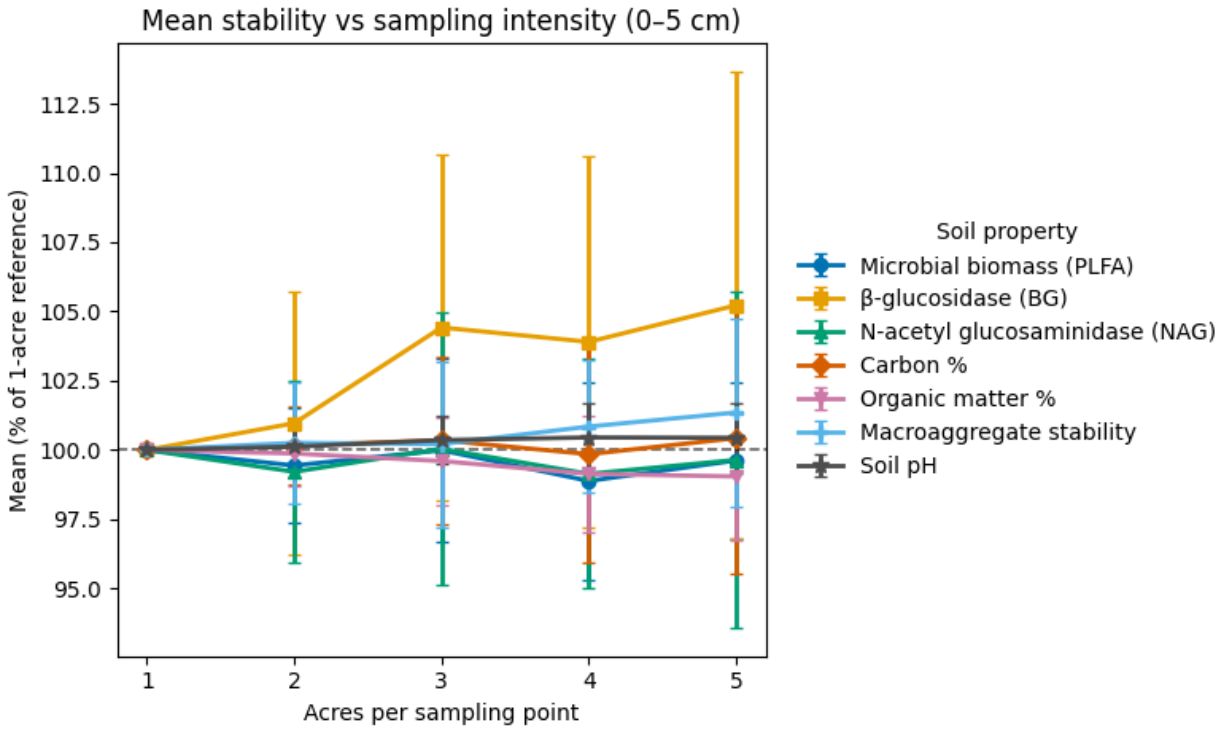

Figure 3.1. Stability of soil property means across sampling intensities (0–5 cm depth). Mean values of soil indicators estimated at different sampling intensities expressed as a percentage of the 1-acre reference grid. Error bars represent the standard deviation of resampled subsets at each intensity level. Soil properties include microbial biomass (PLFA; nmol PLFA g^{-1} soil), β-glucosidase (BG; µg PNP g^{-1} soil h^{-1}), N-acetyl glucosaminidase (NAG; µg PNP g^{-1} soil h^{-1}), soil carbon (%), organic matter (%), macroaggregate stability (%), and soil pH.

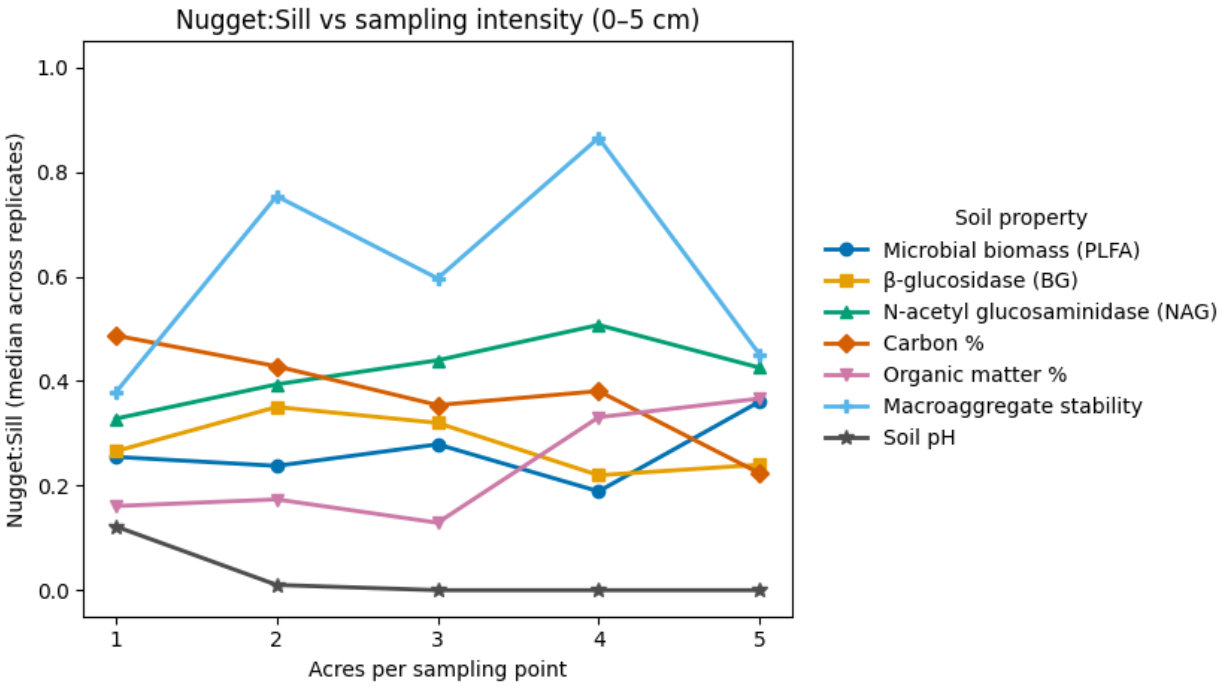

Figure 3.2. Nugget-to-sill ratio of variogram models across sampling intensities (0–5 cm depth). Nugget-to-sill ratios derived from variogram models for soil properties across sampling intensities ranging from 1 to 5 acres per sampling point. Lower nugget-to-sill values indicate stronger spatial structure, while higher values indicate greater random variability.

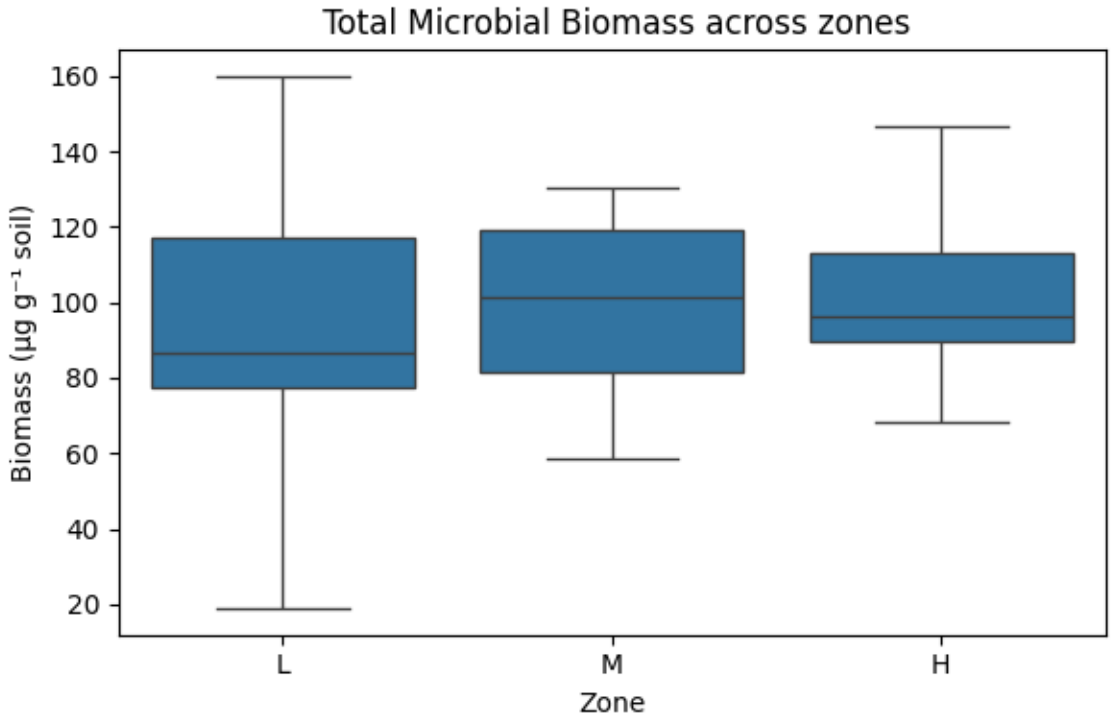

Figure 3.3. Distribution of total microbial biomass across management zones.
Boxplots showing the distribution of PLFA-derived microbial biomass across Low (L), Medium (M), and High (H) management zones. Boxes represent the interquartile range, the horizontal line indicates the median, and whiskers represent the range of observations.

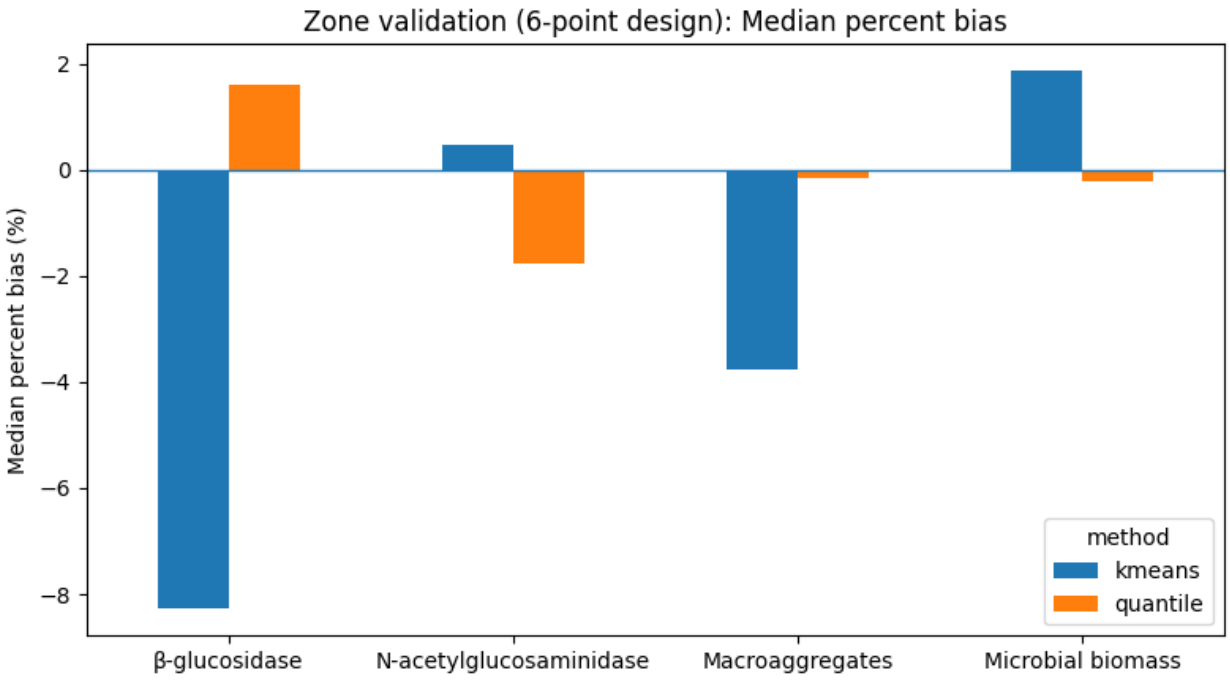

Figure 3.4. Median percent bias of soil indicators estimated from simulated six-point zone sampling designs. Median percent bias of subset means relative to the dense-grid reference mean across 200 resampled six-point zone sampling simulations. Results are shown for both k-means and quantile-based zoning approaches.

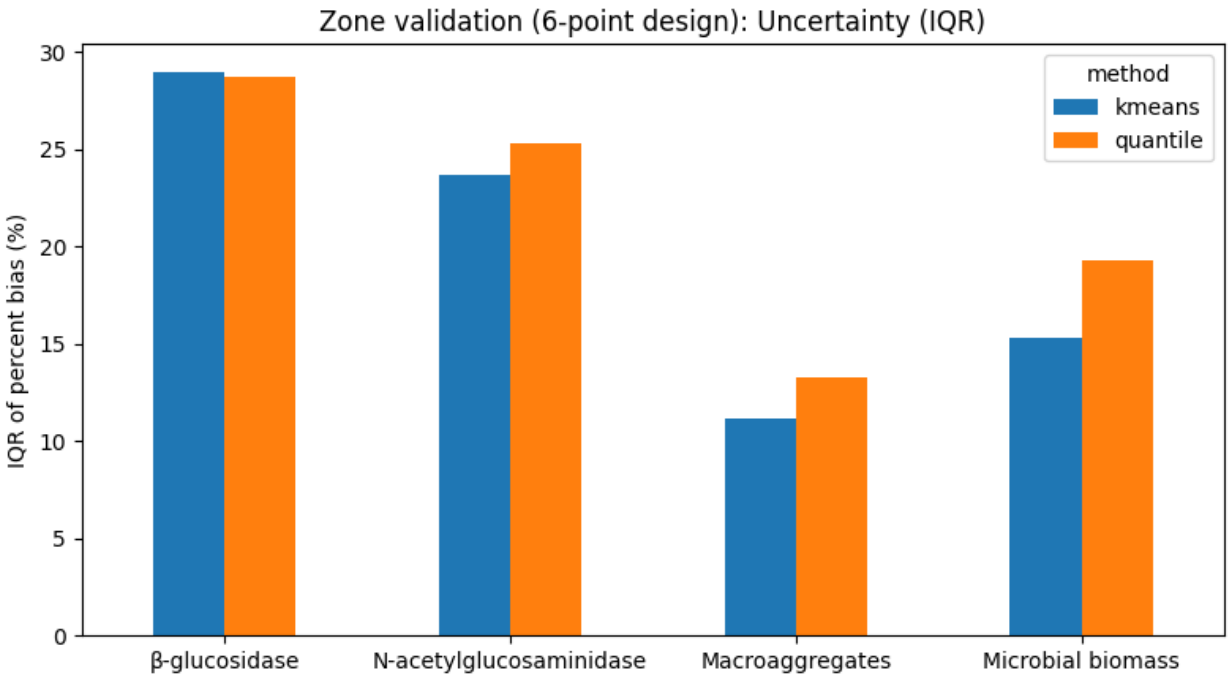

Figure 3.5. Uncertainty of soil indicator estimates under six-point zone sampling designs. Interquartile range (IQR) of percent bias across 200 resampled six-point zone sampling simulations. Lower values indicate greater stability of sampling estimates relative to the dense-grid reference.

Tables

Table 3.1. Summary statistics of resampled grid configurations representing sampling intensities ranging from one sampling point per acre to one sampling point per five acres. Values represent the number of sampling points retained in each resampled subset across replicate simulations. (number of replications = 50)

Acres per Point	Mean number of points	SD	min	25%	50%	75%	max
1	81	-	81	81	81	81	81
2	47	2.76	42	45	47	49	53
3	36	2.38	31	34	36	38	40
4	29	1.26	27	28	29	30	32
5	25	1.67	20	24	25	26	28

Table 3.2. Soil health indicators across management zones at the Dickinson County site (0–5 cm depth). Mean values (± standard deviation) of soil biological indicators and aggregate stability within management zones classified as Low, Medium, and High based on soil organic carbon and pH gradients. Values represent variability among sampling locations within each zone.

Zone	Microbial biomass	BG	NAG	Macroaggregate stability
Low	92.38 ± 32.01	326.07 ± 159.77	90.17 ± 29.23	0.647 ± 0.158
Medium	99.43 ± 24.33	288.82 ± 135.12	92.96 ± 28.19	0.686 ± 0.131
High	102.15 ± 21.04	329.21 ± 77.13	81.82 ± 28.48	0.694 ± 0.118

BG: β-glucosidase activity. NAG: N-acetyl glucosaminidase. Units – Enzyme activity - μg PNP g^{-1} soil h^{-1} , Microbial biomass – nmol PLFA g^{-1} dry soil

Table 3.3. Accuracy and uncertainty of zone-based sampling estimates relative to the dense-grid reference. Median per cent bias and interquartile range (IQR) of soil indicator estimates obtained from simulated six-point zone sampling designs. Validation was performed using both k-means clustering (the operational zoning method) and quantile-based zoning to assess the sensitivity of the results to the zone delineation approach.

Variable	Method	Median bias (%)	IQR (%)
BG	K-means	-8.27	28.95
	Quantile	1.6	28.73
Macroaggregates	K-means	-3.76	11.18
	Quantile	-0.16	13.3
NAG	K-means	-0.48	23.66
	Quantile	-1.77	25.33
Microbial biomass	K-means	1.86	15.32
	Quantile	-0.22	19.32

BG: β-glucosidase activity. NAG: N-acetyl glucosaminidase

Appendix A – Soil microbial biomass as an emergent property of soil health: A predictive modeling approach

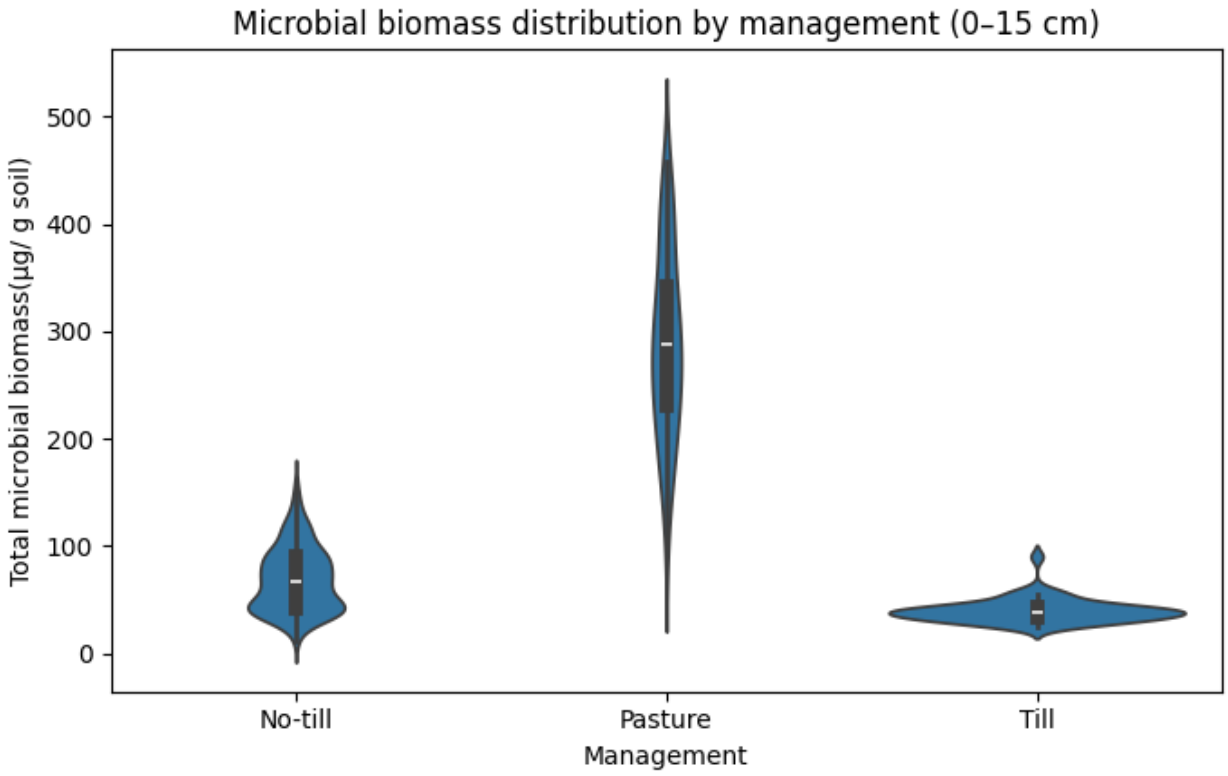

Figure A1. Raw-scale distribution of soil microbial biomass across management systems and depths (0–15 cm).

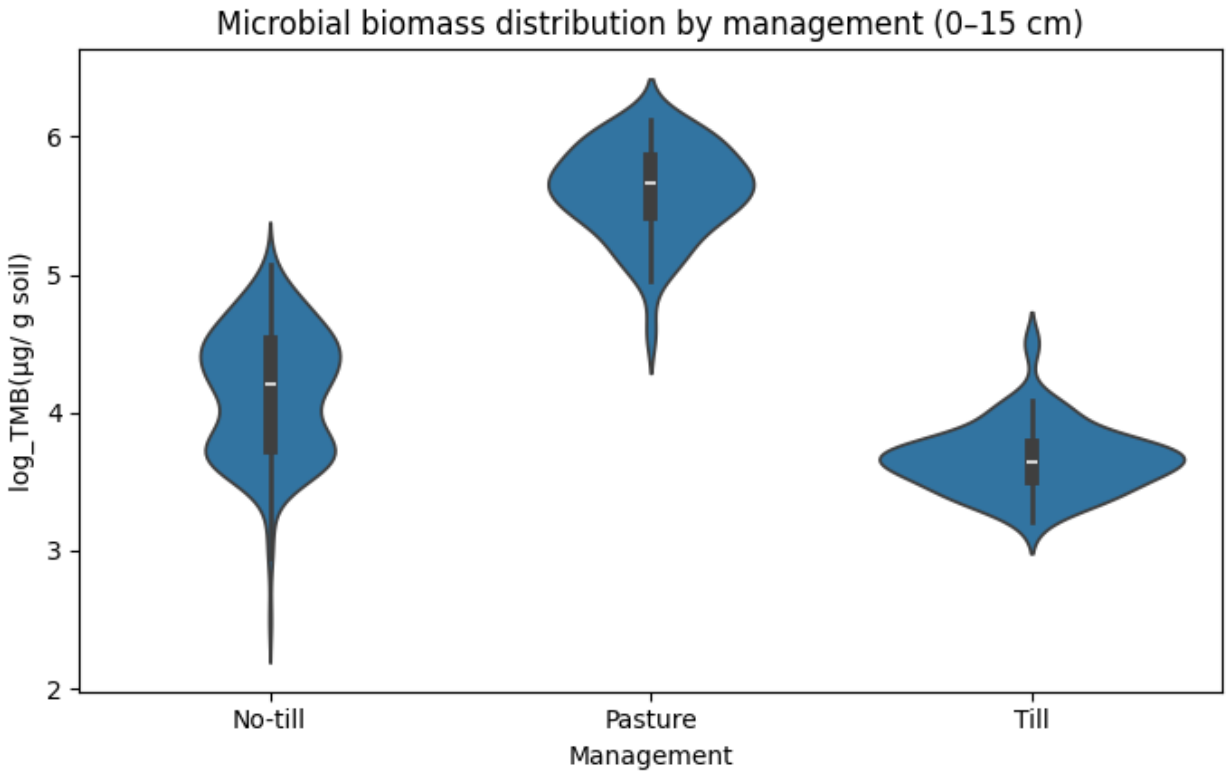

Figure A2. Distribution of soil microbial biomass after Yeo–Johnson transformation across management systems and depths (0–15 cm).

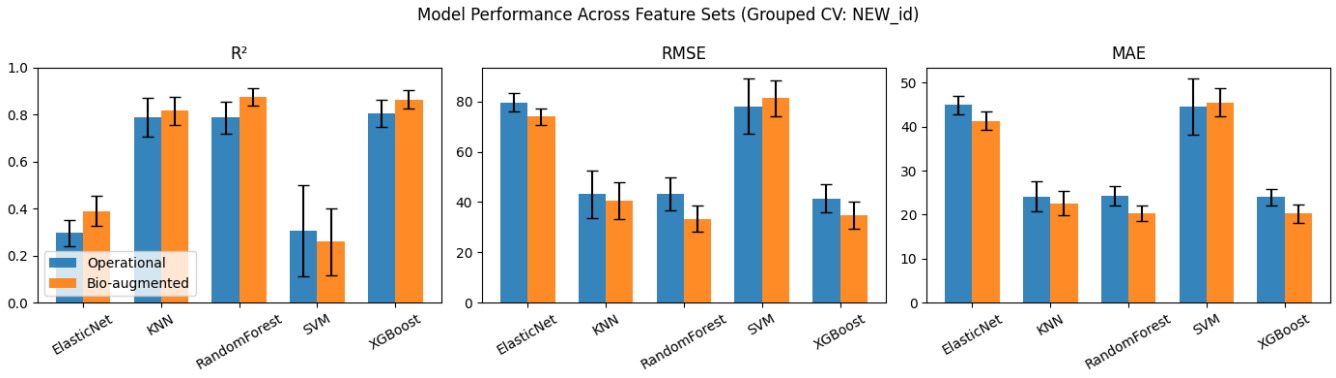

Figure A3. Predictive performance of machine-learning models across operational and bio-augmented feature sets.
Model evaluation was conducted using grouped cross-validation based on soil core identifiers (NEW_id).

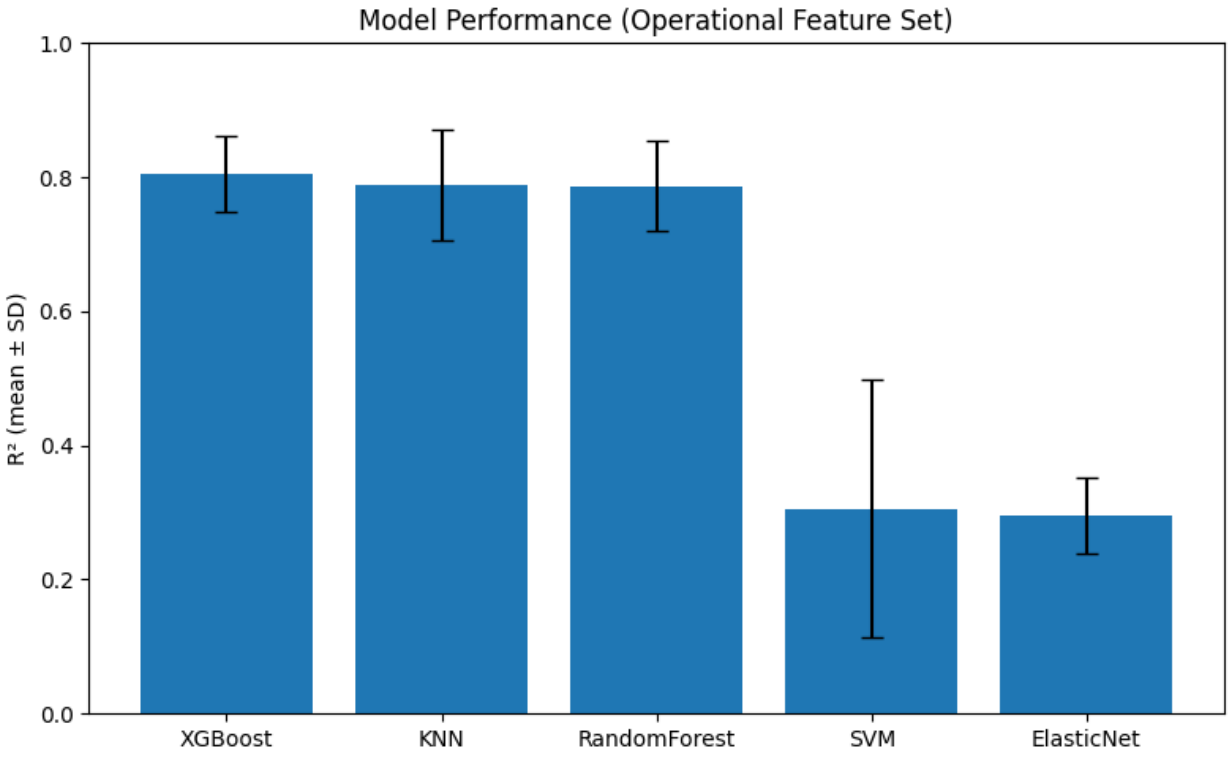

Figure A4. Residual distribution of model predictions for soil microbial biomass across cross-validation folds, illustrating model bias and variance.

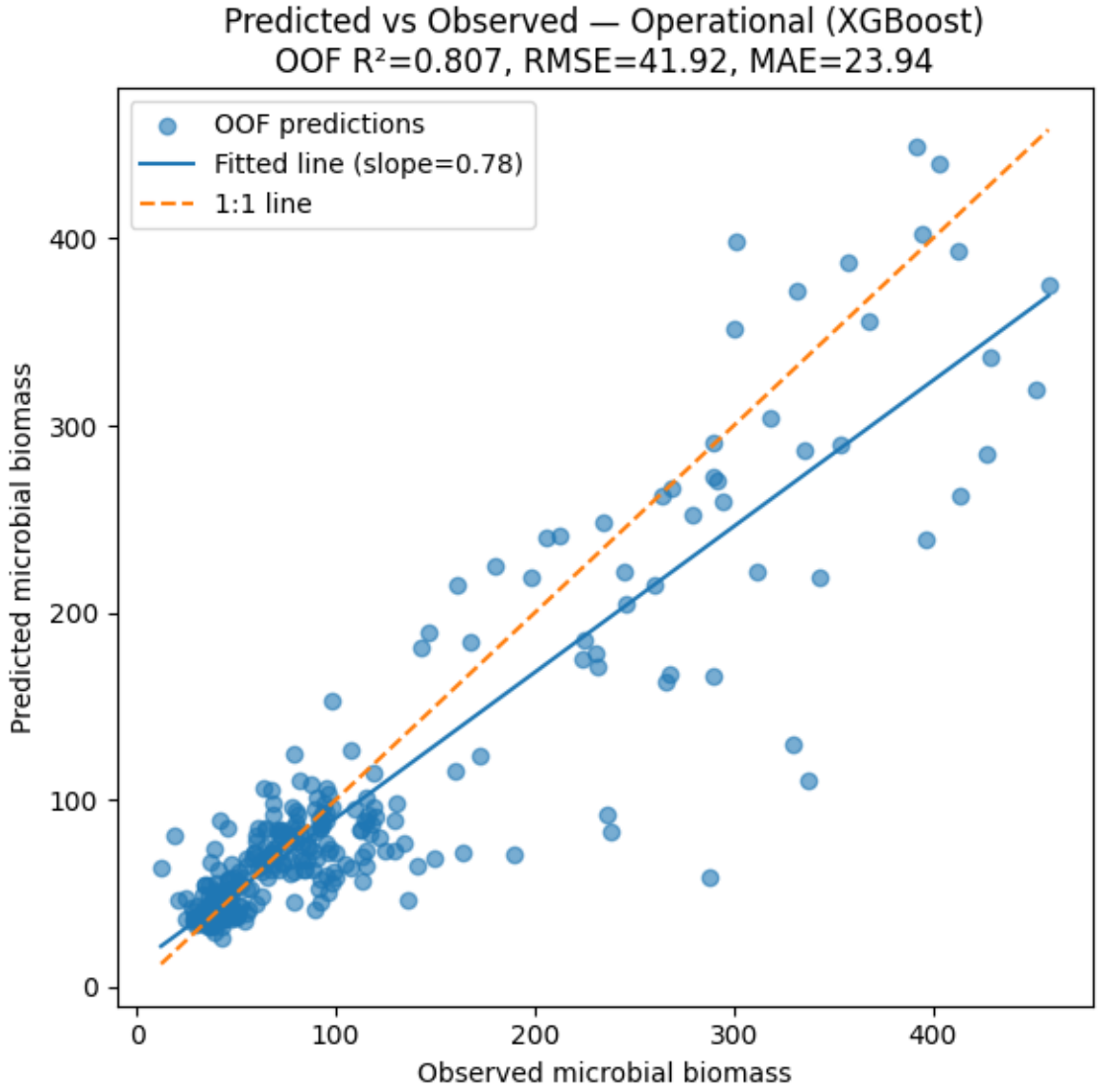

Figure A5. Relationship between observed and predicted soil microbial biomass for the XGBoost model using operational soil properties. Predictions across grouped cross-validation folds.

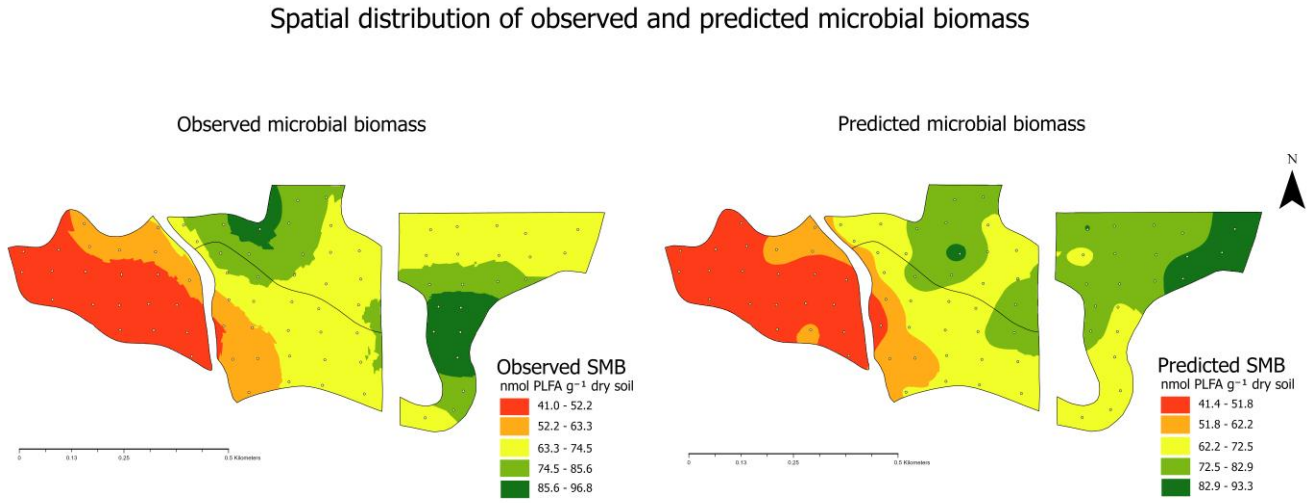

Figure A6. Spatial distribution of observed and predicted microbial biomass (0-5 cm) from XGBoost operational model through kriging. Predicted microbial biomass values were generated using the best-performing operational model. Maps illustrate the extent to which predicted values preserved broad spatial patterns observed in measured microbial biomass.

Table A1. Descriptive summary of soil physicochemical, biological, and response variables used in predictive modeling.

Variable	Mean	Median	SD	Min	Max
Total Microbial Biomass	102.98	68.07	95.50	12.16	458.17
pH	6.57	6.60	0.91	4.5	8.40
OM%	3.58	3.10	1.42	1.70	9.20
C%	2.06	1.72	1.08	0.43	6.93
N%	0.17	0.15	0.07	0.04	0.54
Sand%	18.52	17	6.42	8	55
Silt%	50.04	50	6.02	27	90
Clay%	31.45	32	7.32	10	49
P (ppm)	42.08	20.30	44.45	0	210
K (ppm)	310.07	298.55	94.30	110	668
Ca (ppm)	2732.22	2621.1	915.04	274	8401
Mg (ppm)	567.34	529.5	207.93	123	1337
Na (ppm)	24.05	20.65	12.89	3.6	97
Fe (ppm)	55.5	43.8	38.77	0	231
Mn (ppm)	30.89	27.8	18.07	0	115.1
Zn (ppm)	1.54	0.95	1.4	0	7.2
Cu (ppm)	1.43	1.4	0.54	0	5.2
BG	185.22	124.97	167.38	5.82	979.41
NAG	58.2	45.84	42.5	9.08	220.76
Phos	529.81	413.29	322.24	58.48	1960.65
Aryls	118.58	61.37	148.91	12.61	1098.72

BG - β-glucosidase activity; NAG - N-acetyl glucosaminidase; Phos - acid phosphatase; Aryls - arylsulfatase. Enzyme activities are expressed as μg PNP g^{-1} soil h^{-1}. Total microbial biomass is expressed as nmol PLFA g^{-1} dry soil.

Appendix B – Optimizing field scale soil sampling for soil health assessment

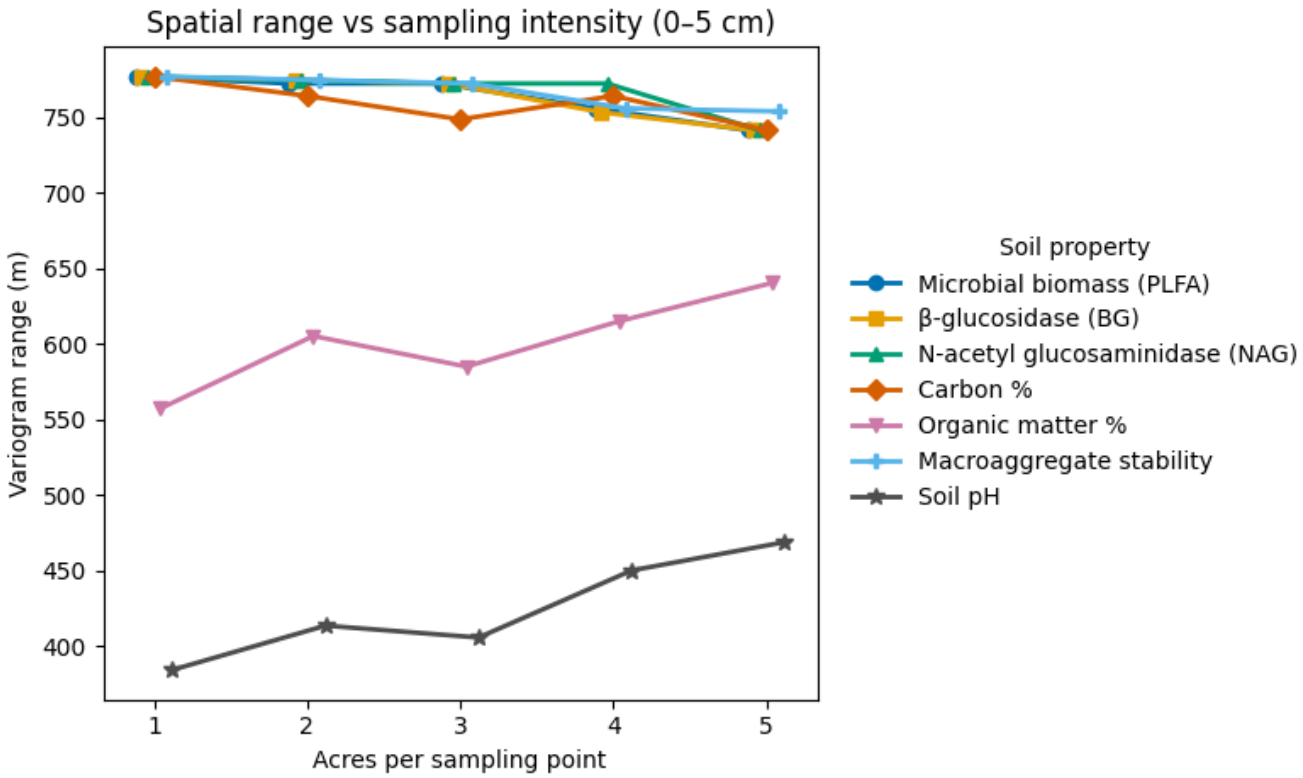

Figure B1. Variogram range estimates across simulated sampling intensities for selected soil health indicators at North Farm (0–5 cm depth).

Table B1. Semivariogram parameters for selected soil health indicators at the reference sampling intensity (1 sample acre^{-1}) at North Farm (0–5 cm depth).

Variable	Nugget	Sill	Range (m)	Nugget : Sill
Total microbial biomass	117.13	459.31	776.92	0.255
BG	6294.37	23679.34	776.92	0.266
NAG	277.11	844.72	776.92	0.328
C%	0.102	0.210	776.92	0.487
OM%	0.07	0.436	557.17	0.161
Macroaggregate stability	0.009	0.024	776.92	0.378
pH	0.07	0.575	384.14	0.122

BG - β-glucosidase activity, NAG - N-acetyl glucosaminidase activity, C% - soil organic carbon, OM% - organic matter