COMMUNICATIONS IN SOIL SCIENCE AND PLANT ANALYSIS
https://doi.org/10.1080/00103624.2025.2454015

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Pedotransfer Functions for Soil Protein Based on Random Forest Modeling for Routine Soil Health Analysis

Joseph P. Amsili, Harold M. van Es, and Robert R. Schindelbeck

Section of Soil and Crop Sciences, School of Integrative Plant Science, Cornell University, Ithaca, New York, USA

ABSTRACT

Autoclaved-citrate extractable soil protein (ACE protein, hereafter referred as "soil protein") is a novel biological soil health indicator that can indirectly capture a soil's capacity to supply nitrogen (N) but is relatively expensive to assess. To explore cost saving options, a dataset of 4,171 soil samples with texture, total carbon (C) and N, carbon-to-nitrogen ratio (C/N), soil protein, permanganate-oxidizable carbon (POXC), pH, and extractable magnesium (Mg) and iron (Fe), was used to develop three pedotransfer functions for soil protein. These included a full random forest (RF) model utilizing all variables, and a reduced RF model and a multiple linear regression model employing a subset of the variables. Models were validated using a US subset of the North American Project to Evaluate Soil Health Measurements dataset that contained 1,406 samples. The full RF model for soil protein reduced the root mean square error (RMSE) by 41.7 and 53.4% compared to reduced RF and multiple linear regression models, respectively. Total C was a more important variable in the model than total N. Additionally, POXC, sand, clay, and extractable Mg and Fe were found to be important in the model. Soil protein was sensitive to management at 36 of 57 long-term experiments. The full RF model was able to replicate 92% of those significant effects of management on soil protein. The new RF pedotransfer function for soil protein can improve prediction compared to traditional regression techniques and reduce the cost of comprehensive soil health assessment.

ARTICLE HISTORY
Received 4 January 2024
Accepted 7 January 2025

KEYWORDS
Carbon; nitrogen; pedotransfer function; random forest; soil health; soil protein

Introduction

A grand challenge in agriculture is to manage nitrogen (N) in a way that ensures adequate N for achieving high crop yields while minimizing N pollution. Soil health indicators that assess a soil's capacity to supply N may play an important role in improving N use efficiency. Autoclaved-citrate extractable soil protein (ACE protein, hereafter referred as "soil protein") is an important pool of soil organic N that is sensitive to management and amenable to high throughput commercial testing (Hurisso et al. 2018). Soil protein is an indicator of the size of the organic N pool and therefore provides an indirect measure of a soil's capacity to supply N to crops. As a result, soil protein has emerged as a prominent biological indicator of soil health (Stott 2019).

An extensive research base has demonstrated that soil protein is sensitive to management (van Es and Karlen 2019) and is positively correlated with important soil health indicators (Hurisso et al. 2018; Liptzin et al. 2023). The same protein extraction procedure was previously used under the names easily extractable glomalin, easily extractable glomalin related soil protein, and then easily extractable Bradford reactive soil protein, which is why there is a larger research base than expected for this

CONTACT Joseph P. Amsili jpa28@cornell.edu Section of Soil and Crop Sciences, School of Integrative Plant Science, Cornell University, 1001 Bradfield Hall, Ithaca, NY 14853, USA

Supplemental data for this article can be accessed online at https://doi.org/10.1080/00103624.2025.2454015

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

"novel" soil health indicator (Gillespie et al. 2011; Hurisso et al. 2018; Rillig 2004; Rosier, Hoye, and Rillig 2006; Wright and Upadhyaya 1998). These name changes came about as researchers determined that the procedure was extracting proteins from a wide range of sources and not just those related to glomalin, a putative glycoprotein of arbuscular mycorrhizal fungi origin (Rosier, Hoye, and Rillig 2006). Soil protein has been found to be sensitive to cropping system and specific management practices such as tillage, organic amendments, cover crops, and crop rotation (Amsili et al. 2023; Liptzin et al. 2023; Malone et al. 2023; Moebius-Clune et al. 2008; Naasko et al. 2023; Nunes et al. 2018; Sainju, Liptzin, and Stevens 2022; van Es and Karlen 2019). A few studies have further demonstrated that soil protein was more sensitive to management than soil organic matter (SOM; Nunes et al. 2018; van Es and Karlen 2019) and total N (Naasko et al. 2023; Sainju, Liptzin, and Stevens 2022). Martin, Culman, and Sprunger (2022) found that soil protein was more sensitive than permanganate-oxidizable carbon (POXC), another widely used soil health indicator related to phenolic compounds, in an experiment comparing corn–soybean and corn–perennial legume forage–perennial legume forage rotations. Nunes et al. (2018) also found that soil protein was more sensitive to tillage than POXC at 2 of 3 long-term tillage experiments in New York, USA. Soil protein has been shown to be positively correlated with several important soil health indicators, including wet aggregate stability, total N, and potentially mineralizable N (PMN). There is a long history of demonstrating the linkage between soil protein and wet aggregate stability across a range of climates and soils ($r = 0.53$–0.84; Amsili, van Es, and Schindelbeck 2021; Fine, van Es, and Schindelbeck 2017; Rillig 2004; Wright and Upadhyaya 1998). This linkage was originally hypothesized to be due to this soil protein fraction being an indicator for arbuscular mycorrhizal fungi, which are known to play an important role in building and maintaining stable aggregates (Rillig 2004). However, SOM also has strong positive correlations with aggregate stability ($r = 0.58$–0.77; Amsili, van Es, and Schindelbeck 2021; Fine, van Es, and Schindelbeck 2017), indicating that positive relationships between soil protein and aggregate stability should not be used to make statements about the effects of AMF on aggregate stability. Soil protein has also been shown to be positively correlated with total N ($r = 0.52$–0.93; Geisseler et al. 2019; Hurisso et al. 2018; Liptzin et al. 2023Naasko et al. 2023).

Soil protein has the potential to improve our quanitification of a soil's ability to supply N to crops. Past research found that soil protein is positively correlated with PMN ($r = 0.46$–0.69; Geisseler et al. 2019; Jha et al. 2022; Liptzin et al. 2023). However, despite this strong relationships, they were not stronger than the positive correlations between total N and PMN ($r = 0.65$–0.79; Geisseler et al. 2019; Liptzin et al. 2023).

Recent studies have found positive correlations between soil protein and non-legume crop yield for maize (*Zea mays*), wheat (*Triticum aestivum* L.), and switch grass (*Panicum virgatum* L.; Naasko et al. 2023; Roper, Osmond, and Heitman 2019; Sainju, Liptzin, and Stevens 2022; van Es and Karlen 2019). This is noteworthy because finding relationships between biological soil health indicators and crop yield is challenging due to other effects on crop yield that include weather variability, pest pressure, and compaction. Some of these studies either failed to find relationships or found weaker relationships between SOM or total N and crop yield (Naasko et al. 2023; Roper, Osmond, and Heitman 2019; van Es and Karlen 2019). Relationships between soil protein and soybean yield are generally weak (Faé et al. 2020; Malone et al. 2023; van Es and Karlen 2019), which is logical because soybean yield depends less on soil N status because it is a N fixer.

One advantage of soil protein compared to other soil health indicators is that it is less analytically variable (Hurisso, Culman, and Zhao 2018; Liptzin et al. 2023). Naasko et al. (2023) found that soil protein fluctuates less across the season than total N, which would have advantages for picking up treatment effects. Also, Hurisso, Culman, and Zhao (2018) found that soil protein was less analytically variable than SOM, POXC, and soil respiration, which also has advantages for assessing the effects of management practices. Although PMN more closely captures a soil's capacity to supply N than soil protein, its disadvantage is that it has a much higher coefficients of variation than soil protein (Liptzin et al. 2023), which makes it a less reliable indicator of soil health and predictor of non-legume crop yield.

Soil protein is being measured at several public and commercial soil health laboratories. Some offer soil health packages that feature soil protein and respiration in place of PMN because the latter is more costly to measure and more analytically variable (Liptzin et al. 2023) than soil protein or respiration. Yet soil protein analysis is still more costly than other labile carbon (C) indicators like POXC and soil respiration and is potentially redundant if total N is also being measured.

To improve the cost-effectiveness of soil health assessments, laboratories may predict soil protein using machine learning algorithms based on properties included in soil health assessment that are less expensive to measure. Predictive models that relate easy-to-measure soil properties to difficult-to-measure properties are known as pedotransfer functions. Soil scientists have a long history of developing predictive models for soil physical properties, but few have been developed for soil biological properties (Kinoshita et al. 2012; Mcbratney et al. 2002; Ramcharan et al. 2017; Saxton and Rawls 2006; Van Looy et al. 2017). However, in recent years, a large number of models have been developed for soil biological properties due to the growing interest in soil C storage to improve crop productivity and mitigate and adapt to climate (Chambers, Lal, and Paustian 2016; Karlen et al. 2019). Pedotransfer functions have been developed to predict soil organic carbon (SOC), total N, C saturation levels, mineral associated OC, N mineralization, crop yield, and various other biological soil health indicators (Deiss et al. 2023; Glendining et al. 2011; Hassink 1997; Heumann, Böttcher, and Springob 2003; Laurence et al. 2023; Sanderman et al. 2021; Sanderman, Savage, and Dangal 2020; Xiao et al. 2022). These commonly utilize some combination of the following predictor variables: SOC, total N, C:N, particle size fractions (i.e., % sand, silt, and clay), soil chemical properties (i.e., pH, oxalate-extractable iron and aluminum), diffuse reflectance spectroscopy (i.e., visible and near-infrared or mid-infrared), soil taxonomy, and non-soil covariates like climate (i.e., mean annual precipitation and temperature) and land cover. One recent study showed that bacterial microbiome 16S rRNA data could also be used in the future to predict biological soil health indicators (Wilhelm, van Es, and Buckley 2022).

Pedotransfer functions are commonly developed using linear regression or machine learning (ML) methods, the latter having grown in importance in recent years. Linear regression has advantages for ease of implementation and interpretation, but ML algorithms generally offer greater predictive power (Padarian, Minasny, and Mcbratney 2020). Soil scientists have used various ML algorithms to develop pedotransfer functions, including, neural networks, support vector machines, k-nearest neighbors, multivariate adaptive regression spline, regression trees, random forest, cubist, boosted regression trees (e.g., extreme gradient boosting), and deep learning (Padarian, Minasny, and Mcbratney 2020; Rubio et al. 2025; Van Looy et al. 2017).

The objectives of this research were to (i) develop pedotransfer functions using random forest and linear regression techniques for prediction of soil protein in routine soil health analyses, (ii) extract variable importance insights from pedotransfer functions to determine how predictor variables shape soil protein, (iii) validate pedotransfer functions using an independent dataset, and (iv) assess ability of models to be sensitive to management.

Materials and methods

Training and validation datasets

A soil health dataset was compiled from 4,171 soil samples (0–15 cm depth) from across the continental U.S. that were collected and analyzed between 2014 and 2022 (Figure 1; Table S1). The dataset was derived from routine soil sample submission to the Cornell Soil Health Laboratory. While all regions of the continental U.S. were represented in the Training Dataset, most samples came from the Midwest and Northeast, followed by the Mid-Atlantic, Northern Plain State, Southeast, and Southwest regions (Figure 1; Table S1). Soil samples were, per guidelines, assumed to have been collected by compositing five or more soil slices (0–15 cm depth) from different locations in a field or plot (Cornell Soil Health Laboratory 2022). Samples were analyzed for soil texture and a suite of

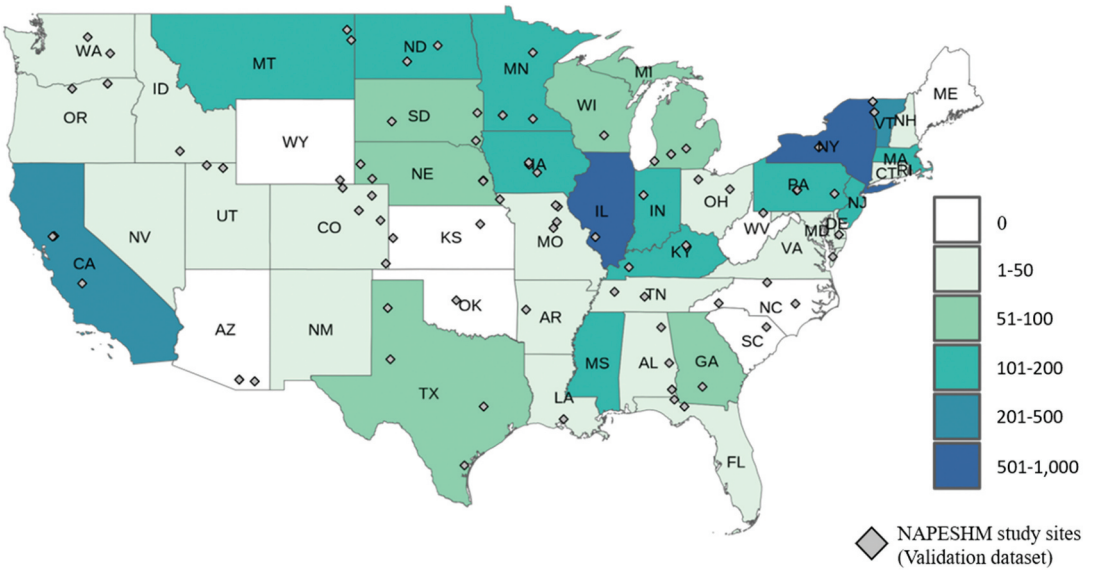

Figure 1. Distribution of samples by state across the continental U.S. within the Training Dataset for autoclaved-citrate extractable soil protein pedotransfer functions. Soil samples from NAPESHM study sites within the U.S. were used as the validation dataset to test models.

physical, biological, and chemical soil health (SH) indicators according to the Comprehensive Assessment of Soil Health (CASH) protocol (Moebius-Clune et al. 2017). Therefore, all samples contained measurements for the following properties that were used to develop models: sand, silt, clay, soil protein, total carbon (C), total nitrogen (N), permanganate oxidizable carbon (POXC), pH, and Modified Morgan extractable magnesium (Mg) and iron (Fe). All soil analyses procedures are described in Schindelbeck et al. (2022). Furthermore, methods for soil texture, SOM, POXC, and Modified Morgan nutrient extractions were described in a companion paper on pedotransfer function development for field capacity, permanent wilting point, and available water capacity and were therefore excluded here (Amsili, van Es, and Schindelbeck 2024a). Soil protein was extracted from 3.0 g soil samples, sieved to less than 8 mm, with 0.02 M sodium citrate at pH 7. The extract was run through a sequence of autoclaving and centrifugation steps, and then quantified by the bicinchoninic acid assay against a bovine serum albumin standard curve for soil protein concentration (Hurisso et al. 2018; Wright and Upadhyaya 1998). Total C and N in this dataset was measured with a Primacs SNC-100 Combustion Analyzer (Skalar, Buford, GA). Soil pH was measured in a 1:2 soil:water slurry.

Exploratory soil protein model development determined that 96-h soil respiration (also called C mineralization), wet aggregate stability, and Modified Morgan extractable phosphorus, potassium, manganese, and zinc added little predictability to the model and therefore were not included as predictor variables. Although SOM was used as a predictor variable previously, it was removed due to preference for total C and total N and evidence that it did not add any additional predictability to the model in exploratory analyses. Soils with SOM and soil protein values greater than 10.0% and 40 mg g^{-1} were removed from the dataset to develop models that are appropriate for agricultural soils.

The North American Project to Evaluate Soil Health Measurements (NAPESHM; Norris et al. 2020) data were used as an independent validation dataset to test the accuracy of developed pedotransfer models for soil protein. It involved a complete SH dataset from 1,406 composite soil samples from 0 to 15 cm depth that were collected in 2019 within the contiguous US (Figure 1). More details on this subset of the NAPESHM Dataset can be found in Amsili, van Es, and Schindelbeck (2024a) and Norris et al. (2020). This validation dataset contains all the same soil properties using the same methods as in the Training Dataset. The Training and NAPESHM Datasets had similar distributions

of sand, silt, clay, and soil protein/SOM (Figure S1; Table S2). However, the Training Dataset had slightly higher mean soil protein and total C values and a larger proportion of samples had soil protein and total C values that were higher than 10 mg/g and 4.0%, respectively, than the NAPESHM Dataset (Figure S1).

The NAPESHM Dataset was also used to determine if model predictions could be equally sensitive to management as the measured values. We selected 57 experiments with at least 12 experimental plots (1,178 observations) and assessed sensitivity of measured soil protein to management practices. In experiments where treatments had a significant effect on soil protein, we investigated whether modeled predictions could replicate those results.

Modeling, validation, and variable importance approaches

A multiple linear regression (MLR) model for soil protein was fit to the Training Dataset using total C, total N, and clay with the lm function in R (R Core Team 2022). Full and reduced random forest (RF) models were developed for soil protein. The full RF model used sand, silt, clay, total C, total N, carbon-to-nitrogen ratio (C/N; as determined by total C/total N), POXC, pH, and Modified Morgan Mg and Fe. Whereas the reduced RF model used sand, silt, clay, total C, and total N. Full and reduced RF models were compared to determine how new soil health variables contribute to prediction of soil protein. A detailed description of the RF modeling approach is discussed in Amsili, van Es, and Schindelbeck (2024a).

Initial validation of pedotransfer functions was done with the Training Dataset by plotting predictions against measured values. For the RF models, percent of variance explained (%Var) and out of bag (OOB) root mean square error (RMSE) were calculated. %Var is a metric with RF models and can be described as how well the OOB predictions match the observed values. Out of bag RMSE was calculated by comparing OOB model predictions to measured values. For the MLR model, adjusted R^2 and RMSE were calculated. A secondary validation step was carried out with the NAPESHM data. Regression equations, RMSE and mean absolute error (MAE), were all used as validation metrics in the secondary validation step. Equations for %Var, RMSE, and MAE can be found in Amsili, van Es, and Schindelbeck (2024a).

Variable importance analysis and visualization were carried out using the randomForestExplainer, kernalshap and shapviz packages in R (Mayer and Stando 2023; Mayer, Watson, and Prezemyslaw 2023; Palusynska, Biecek, and Jiang 2022). Distribution of min mean depth, mean square error (MSE) increase, and Node Purity increase were plotted for the full RF model using randomForestExplainer. Shapley Additive exPlanations (SHAP) values were calculated and visualized using the kernalshap and shapviz packages in R, respectively (Mayer and Stando 2023; Mayer, Watson, and Prezemyslaw 2023). Detailed descriptions of these metrics are available in Amsili, van Es, and Schindelbeck (2024a).

Partial dependence analysis with the pdp package was used to determine the influence of single predictor variables on soil protein in the full RF model while other variables are held constant. The four most important predictor variables for soil protein, as indicated by the MSE increase metric, were run as individual partial dependence plots. The pdp package also provides functionality to plot multi-predictor partial dependence plots (Greenwell 2017), which were constructed with total C and clay to visualize interaction among those variables. Variable importance and partial dependence plots were only made for the full RF model.

Results and Discussion

Model comparison and validation

To our knowledge, this is the first effort to develop pedotransfer functions to predict soil protein quantified by the ACE method. Pedotransfer functions were developed based on a full RF model and a reduced RF model, as well as the following MLR model (Eq. 1):

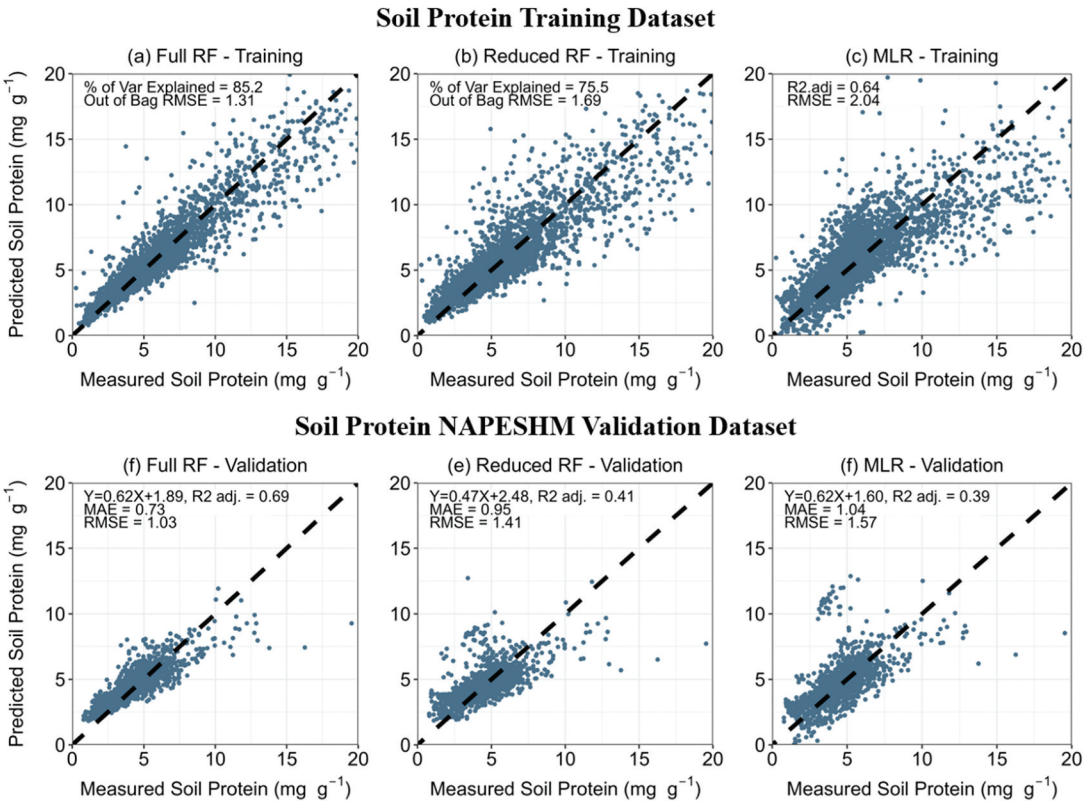

Figure 2. (a–c) Measured vs. predicted autoclaved-citrate extractable soil protein (soil protein) for full random forest (RF), reduced RF, and multiple linear regression models on the Training Dataset. (a, b) Full and reduced RF models show RF model out of bag predictions. (d–f) Measured vs. predicted soil protein for full RF, reduced RF, and multiple linear regression models for the NAPESHM Validation Dataset.

$$soil\ protein = 2.86 - 0.1297\ clay + 1.172\ total\ C + 17.562\ total\ N \quad (1)$$

Units of soil protein and predictor variables in Eq. 1 are in mg/g and %, respectively. The first validation step with the Training Dataset revealed that the full RF model had the highest percent variance explained (85.2%) and lowest RMSE (1.31) compared to the reduced RF model (75.5%; RMSE = 1.69) and MLR model ($R^2_{adj} = 0.64$; RMSE = 2.04; Figure 2a–c). The independent validation step with the NAPESHM Dataset reinforced that the full RF model was capable of much better predictions ($R^2_{adj} = 0.69$) than the reduced RF or MLR models ($R^2_{adj} = 0.41$ and 0.39, respectively; Figure 2 (d–f). It also demonstrated that each model had RMSEs that were on average 18.8% lower (reduced RF: 13.1%, full RF: 20.8%, and MLR: 22.5%) than the same model calculated for the Training Dataset (Figure 2). This is presumably because the Training Dataset covered a broader attribute space for soil protein and total C than the NAPESHM Validation Dataset (Figure S1). The full RF model had a 41.7 and 53.4% lower RMSE than the reduced RF or MLR model, respectively, which indicates that inclusion of additional variables beyond clay, total C, and total N can provide meaningful improvements to prediction of soil protein (Figure 2d–f). In contrast, we found that the reduced RF model had a similar R^2 as the MLR model but was still able to slightly reduce the RMSE, demonstrating modest benefits of RF compared to MLR for prediction. If the goal is prediction and access is available to many potential predictors of soil protein, machine learning algorithms like RF can utilize more predictor variables than MLR to improve prediction of soil protein.

Variable importance

Total C was the most influential variable for predicting soil protein as evidenced by partial dependence analysis and mean SHAP values, and was also the most important variable in the model for reducing the MSE (Figure 3; Figure S2). Total N and POXC had similar impact on soil protein and error reduction in the model (Figure 3; Figure S2). Total C, total N, and POXC were important predictors of soil protein due to their well-established and positive relationships with soil protein (Amsili, van Es, and Schindelbeck 2021; Fine, van Es, and Schindelbeck 2017; Geisseler et al. 2019; Hurisso et al. 2018; Malone et al. 2023). Past significant results were reinforced in the Training Dataset where linear regression coefficients between soil protein and total C, total N, and POXC were $R^2_{adj} = 0.48$, 0.47, and 0.45, respectively (Table S3). Total C and POXC both had strong positive relationships with soil protein through their range, whereas total N had a positive relationship on soil protein between 0.1 and 0.4% total N but then appeared to plateau between 0.4 and 0.5% total N (Figure 3). Although these variables were all expected to be important predictors of soil protein, total N was surprisingly less important in the model because soil protein is supposed to represent a bioavailable pool of organic N (Hurisso et al. 2018). There are two explanations. The tight positive relationship between total N and total C or POXC in the Training Dataset ($R^2_{adj} = 0.80$, $R^2_{adj} = 0.69$, respectively; Table S4) means that total C was also able to represent changes in total N on soil protein. Second, the soil protein procedure potentially co-extracts humic and lipid substances along with protein substances (Gillespie et al. 2011; Roberts and Jones 2008; Schindler, Mercer, and Rice 2007). Previous chemical characterization of glomalin related soil protein, the previous name, showed that the extracted soil protein fraction had a C/N ratio (13.3–15.6) intermediate between bovine serum albumin protein (4.1) and humic acids (18–23.8; Schindler, Mercer, and Rice 2007). If one assumes that proteins contain 16% N (Jones Factor; equivalent to a C/N ratio of 6.25; Jones 1941), then soil protein N makes up an average of 55.3 or 55.7% of total N (Figure S1). A recent study in California found that soil protein N accounted for 28% of total N across all sites, which is almost ½ of what was found across the datasets used in this study (Geisseler et al. 2019). The authors stated that this is likely because soil protein makes up a smaller percentage of total N in low N soils. A general consensus in soil science is

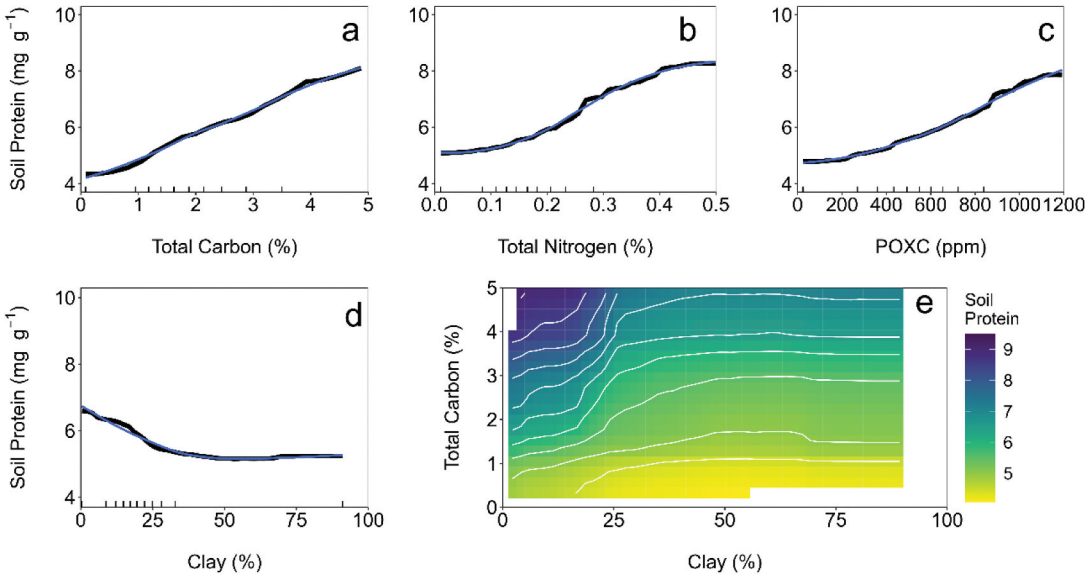

Figure 3. (a–d) Partial dependence plots for individual autoclaved-citrate extractable soil protein (soil protein) predictor variables: total carbon, total nitrogen, permanganate oxidizable carbon (POXC), and clay. (e) Multi-predictor partial dependence plot with the interaction between clay and total carbon. These partial dependence plots show the influence of individual predictor variables on soil protein within the full RF model while all other predictor variables are held constant.

that proteins in soil generally make up 30–45% of total N (Stevenson 1982). Therefore, assuming soil protein reflects total protein, soils with soil protein N/total N greater than 45% would indicate that the C/N ratio of that fraction is above 6.5, as suggested by Schindler, Mercer, and Rice (2007).

When successively removing each predictor included in the full RF model, total N became the most important variable in the model when total C was removed (Figure S5; OOB RMSE increased from 1.31 to 1.48). We found that removing total N led to increases in OOB RMSE from 1.31 to 1.38 (5.3% increase) in the Training Dataset and in RMSE from 1.03 to 1.10 (6.8% increase) in the Validation Dataset (Table S6). Exclusion of total N from the RF model led to the largest increase in RMSE in the Southern Plain States (14.9% increase) and Southwest (69.4% increase) regions presumably because total C is an inadequate predictor of soil protein in regions with significant concentrations of soil inorganic carbon (Table S6).

The poor relationship between total C and soil protein in the NAPESHM Dataset ($R^2_{adj} = 0.13$) due to the higher proportion of samples with significant amounts of inorganic carbon compared to the Training Dataset (Table S3), might lead one to believe that using SOC in the RF model might improve the soil protein prediction accuracy. However, when we created an RF model with SOC instead of total C and SOC/total N instead of total C/total N, using a slightly smaller dataset ($n = 2{,}985$ vs. $n = 4{,}171$), it did not improve prediction accuracy of the model across the NAPESHM Dataset (Figure S6). Apparently other variables in the RF model like POXC, total N, extractable Mg, and pH can address the weaker relations between total C and soil protein than SOC and soil protein.

Clay, sand, Mg, and Fe had intermediate importance in the model after total C, total N, and POXC. Clay had a negative relationship with soil protein, whereas sand had a positive relationship with soil protein in the model (Figure 3d; Figure S2; Table S3). These trends have been documented in past research and are hypothesized to be due to soil protein having a lower extraction efficiency with higher clay content (Amsili, van Es, and Schindelbeck 2021; Fine, van Es, and Schindelbeck 2017). Although correlations between sand or clay and soil protein were not found in the NAPESHM Dataset (Liptzin et al. 2023), they were present to a much smaller degree when we trained the same RF model using the NAPESHM Dataset (Figure S8). This indicates that non-linear effects can be taken into account in a machine learning model even if there is not a significant linear relationship. The greater extractability of soil protein in coarser textured soils is also expressed in the positive relationship between sand and soil protein/SOM ($R^2_{adj} = 0.41$) and the negative relationship between clay and soil protein/SOM ($R^2_{adj} = 0.44$; Training Dataset). The significant trends between sand or clay and soil protein/SOM were also found in the Validation Dataset ($R^2_{adj} = 0.32$, $R^2_{adj} = 0.45$, respectively). Related to the influence of soil texture on soil protein values, there was a clear interaction between sand or clay content and total C in the model, where total C had a larger effect on soil protein in coarser textured soils (soils with sand >30–40% and clay <20–25%; Figure 3e). To quantify this interaction, we found that changes in total C from 0 to 4% were associated with increases in soil protein of 9.4, 9.4, 8.8, and 5.1 mg g^{-1} soil in sandy loam, loam, silt loam, and clay loam, respectively (Figure S7). Removing sand, silt, and clay from the RF model led to increases in OOB RMSE from 1.31 to 1.43 (9.2% increase) in the Training Dataset and in RMSE from 1.03 to 1.05 (1.9% increase) in the Validation Dataset (Table S6). This indicates that these texture effects were more clearly seen in the Training Dataset than the NAPESHM Validation Dataset, which was also confirmed by Liptzin et al. (2023). Texture-related parameters, sand and clay, improved predictions of soil protein in the Northeast and Southern Plain State regions, impaired predictions in the Southwest and Southeast, and did neither in other regions (28% of the data came from Northeast). Extractable Mg and Fe had similar impact on soil protein and on reducing error in the model (Figure S2; Table S5). Mg had a negative impact on soil protein values and was slightly more important for improving prediction across the NAPESHM Dataset than it was across the Training Dataset (Table S6). Mg was particularly important for improving prediction of soil protein in the Southwest, Northwest, and North Plain State regions where RMSE increased by 30.6, 18.5, and 15.3%, respectively, when it was removed from the RF model. We hypothesize that extractable Mg helps to parameterize drier

regions that have higher concentrations of carbonates and higher pHs. Extractable Fe had a positive impact on soil protein, which could help parameterize pastures and hayfields, which tend to have high soil protein, but also lower soil pHs because they are less managed, and as a result have higher extractable Fe (Amsili, van Es, and Schindelbeck 2021).

Finally, pH, C/N, and silt had the smallest influence on soil protein in the model. pH values above 7 had a minor negative impact on ACE values (Figure S2c; Figure S3c) and may act like Mg in the model by helping to parameterize drier regions that tend to maintain higher soil pHs. Also, like Mg, soil pH was slightly more important for improving predictions across the NAPESHM Dataset than it was across the Training Dataset. Soil pH led to the greatest improvements in predictability in the Southwest, Southern, and Northern Plain State regions where removing it from the model led to increases in RMSE by 25.5, 20.7, and 16.7%, respectively. This is presumably because pH helps to correct predictions in areas where there are greater concentrations of soil inorganic carbon, which is included in total carbon values. Besides silt, C/N was the least important variable in the model. This was slightly muddled by calculating C/N as total C/total N with high C/N ratios mostly associated with high concentrations of soil inorganic carbon. Therefore, C/N helped to improve predictability of soil protein in the Southwest region. However, if SOC was used in the RF model instead of total C, and C/N was calculated as SOC/total N, then C/N takes on a clearer role in the model (Figure S6). In that alternative, we see a weak positive association between soils with higher SOC/N ratios and soil protein. This is counterintuitive, but perhaps provides evidence for the understanding that the soil protein methodology can include significant co-extraction of humic substances, which might explain this association (Schindler, Mercer, and Rice 2007).

Sensitivity of soil protein and RF models to agronomic management

Laboratory-measured soil protein was sensitive to one-way treatment effects ($p < .05$) from the NAPESHM study at 36 of 57 sites analyzed. This compares to SOC, total N, 96-h soil respiration, and POXC, which were significant at 30, 36, 36, and 42 sites out of 57, respectively. This demonstrates that soil protein can provide greater sensitivity to management than SOC at some sites but was equivalent to total N, which corroborates past evidence showing that soil protein can be more sensitive to management than SOM (Nunes et al. 2018; van Es and Karlen 2019). See Liptzin et al. (2022, 2023) for a more complete discussion of the effects of management practices on SOC, total N, POXC, 96-h soil respiration, and soil protein. Tillage, nutrient type, cropping system, residue return level, and cover crops significantly affected soil protein at 53.7, 50.0, 30.0, 28.6, and 23.8% of sites that tested those management practices. This reinforces previous findings that soil protein is relatively more sensitive to tillage and organic amendments compared to other practices (Hurisso et al. 2018; Malone et al. 2023; Nunes et al. 2018, 2020; Rubio, Sawchik, and van Es 2022). Malone et al. (2023) demonstrated that soil protein was more sensitive to manure application than SOM, POXC, or C mineralization. Prediction of soil protein using the full RF model was able to replicate significant ($p < .05$) effects of management on soil protein at 33/36 (92%) sites. The consistent success of the full RF model to predict these significant management effects indicates that it is reasonable to predict rather than directly measure soil protein to achieve analytical cost savings in soil health testing.

Soil texture appeared to influence the likelihood that soil protein was sensitive to management as it was sensitive to management at 85% of coarse-textured sites but was only at 55 and 62% of medium and fine-textured sites. Modeled results were also more likely to be significant to management at coarse-textured sites (77%) than medium (48%) or fine-textured sites (62%). This is potentially linked to the negative influence of increasing clay content on soil protein values in soils (Figure 3; Figure S2), presumably a result of lower extraction efficiency (Giagnoni et al. 2013). Sainju, Liptzin, and Stevens (2022) also hypothesized that the relationship between soil protein and crop yield might be more likely on coarse-textured soils than medium-textured soils. The duration of the experiment also influenced the likelihood that agronomic management had a significant effect on soil protein, but to a lesser extent than for

AWC (Amsili, van Es, and Schindelbeck 2024a). For example, 55% of experiments running for 0–15 years showed significant results, indicating that soil protein can capture the relatively "short-term" effects of management practices. Whereas 73% of experiments running for >45 years had significant results. Experiment duration had similar influences on modeled soil protein results.

In all, soil protein is a biological SH indicator that is more sensitive to management than SOC and can be reliably predicted using an RF modeling approach. The full RF model was most precise (lowest RMSE) in the Northern Plain States and Northwest and least precise in the Southeast (Figure S4). Overall, the full RF model performed quite well across regions but had some difficulty predicting two groups of samples from Arkansas (Southeast) and Wisconsin (Midwest) with higher soil protein values. Both groups came from long-term experiments that included perennial pasture treatments with measured soil protein values above 10 mg g^{-1}. It appears that the predictability of soil protein declines above approximately 10 mg g^{-1} (Figure 2). This was verified in an exercise where we removed soil protein values above 10 mg g^{-1} before calculating RMSE values with full RF model predictions, which led to decreases in OOB RMSE from 1.31 to 0.92 (29.8% decrease) in the Training Dataset and decreases in RMSE from 1.03 to 0.90 (12.6% decrease) in the Validation Dataset. The decline in predictability above 10 mg/g is also less important because they are above the 80th–90th percentile of soil protein values in CASH and SHAPE frameworks and are therefore interpreted as "very high" values (Moebius-Clune et al. 2017; Nunes et al. 2024).

Model availability and use

Full and reduced RF models for soil protein are available for download at the Harvard Dataverse (Amsili, van Es, and Schindelbeck 2024b). The Training Dataset used to build the RF models is also available at the Harvard Dataverse.

Conclusions

Pedotransfer functions can reduce the cost of comprehensive soil health assessments by predicting certain soil health indicators that are relatively expensive to measure. Soil protein is an emerging biological soil health indicator that can assess the size of the soil organic N pool and indirectly a soil's capacity to supply N to crop but is more expensive to measure than soil respiration and POXC and is potentially redundant if total N is also measured. Random forest modeling can effectively predict soil protein from other soil health measurements and provided improved prediction across the contiguous US compared to MLR modeling. Interestingly, total C was more important in the model than total N. POXC, sand, clay, and extractable Mg and Fe were also important variables in the random forest model. NAPESHM experiments demonstrated that soil protein was sensitive to management at more US long-term experiments than SOC, an equal number compared to total N and soil respiration, and less sites compared to POXC. The full random forest model was able to replicate these significant effects of management on soil protein at a clear majority of sites (92%).

Abbreviations

ACE protein or soil protein, autoclaved-citrate extractable soil protein; C, carbon; N, nitrogen; C/N, carbon-to-nitrogen ratio; SOM, soil organic matter; SOC, soil organic carbon; POXC, permanganate-oxidizable carbon; Mg, magnesium; Fe, iron; PMN, potentially mineralizable N; SH, soil health; CASH, Comprehensive Assessment of Soil Health; NAPESHM, North American Project to Evaluate Soil Health Measurements; ML, machine learning; RF, random forest; OOB, out-of-bag; MLR, multiple linear regression; RMSE, root mean square error; MAE, mean absolute error.

Acknowledgements

The authors acknowledge the staff of the Cornell Soil Health Laboratory who ensured high quality soil health assessment results: Kirsten Kurtz, Zach Batterman, Brianna Binkerd-Dale, Nate Baker, Bamidaaye Sinon, Akossiwoa Sinon, and Steven Dunn. The authors also acknowledge the Soil Health Institute for carrying out the North American Project to Evaluate Soil Health Measurements, which provided a robust and independent validation dataset to test pedotransfer functions for soil protein.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by a contract with the New York State Environmental Protection Fund, which was administered by the New York State Department of Agriculture and Markets.

ORCID

Joseph P. Amsili http://orcid.org/0000-0002-8293-5824
Harold M. van Es http://orcid.org/0000-0001-9822-9476

References

Amsili, J. P., H. M. van Es, D. M. Aller, and R. R. Schindelbeck. 2023. Empirical approach for developing production environment soil health benchmarks. *Geoderma Regional* 34:e00672. doi: 10.1016/j.geodrs.2023.e00672.

Amsili, J. P., H. M. van Es, and R. R. Schindelbeck. 2021. Cropping system and soil texture shape soil health outcomes and scoring functions. *Soil Security* 4:100012. doi: 10.1016/j.soisec.2021.100012.

Amsili, J. P., H. M. van Es, and R. R. Schindelbeck. 2024a. Pedotransfer functions for field capacity, permanent wilting point, and available water capacity based on random forest models for routine soil health analysis. *Communications in Soil Science & Plant Analysis* 55 (13):1967–84. doi: 10.1080/00103624.2024.2336573.

Amsili, J. P., H. M. van Es, and R. R. Schindelbeck. 2024b. Autoclaved-citrate extractable soil protein pedotransfer functions based on random forest: Harvard Dataverse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HGBPCW.

Chambers, A., R. Lal, and K. Paustian. 2016. Soil carbon sequestration potential of us croplands and grasslands: Implementing the 4 per thousand initiative. *Journal of Soil and Water Conservation* 71 (3):68A–74A. doi: 10.2489/jswc.71.3.68A.

Cornell Soil Health Laboratory. 2022. Sample collection. https://soilhealthlab.cals.cornell.edu/testing-services/soil-sample-collection/.

Deiss, L., M. S. Demyan, A. Fulford, T. Hurisso, and S. W. Culman. 2023. High-throughput soil health assessment to predict corn agronomic performance. *Field Crops Research* 297:108930. doi: 10.1016/j.fcr.2023.108930.

Faé, G. S., A. R. Kemanian, G. W. Roth, C. White, and J. E. Watson. 2020. Soybean yield in relation to environmental and soil properties. *The European Journal of Agronomy* 118:126070. doi: 10.1016/j.eja.2020.126070.

Fine, A. K., H. M. van Es, and R. R. Schindelbeck. 2017. Statistics, scoring functions, and regional analysis of a comprehensive soil health database. *Soil Science Society of America Journal* 81 (3):589–601. doi: 10.2136/sssaj2016.09.0286.

Geisseler, D., K. Miller, M. Leinfelder-Miles, and R. Wilson. 2019. Use of soil protein pools as indicators of soil nitrogen mineralization potential. *Soil Science Society of America Journal* 83 (4):1236–43. doi: 10.2136/sssaj2019.01.0012.

Giagnoni, L., A. Migliaccio, P. Nannipieri, and G. Renella. 2013. High montmorillonite content may affect soil microbial proteomic analysis. *Applied Soil Ecology* 72:203–06. doi: 10.1016/j.apsoil.2013.07.010.

Gillespie, A. W., R. E. Farrell, F. L. Walley, A. R. S. Ross, P. Leinweber, K.-U. Eckhardt, and R. I. R. Blyth. 2011. Glomalin-related soil protein contains non-mycorrhizal-related heat-stable proteins, lipids and humic materials. *Soil Biology and Biochemistry* 43 (4):766–77. doi: 10.1016/j.soilbio.2010.12.010.

Glendining, M. J., A. G. Dailey, D. S. Powlson, G. M. Richter, J. A. Catt, and A. P. Whitmore. 2011. Pedotransfer functions for estimating total soil nitrogen up to the global scale. *European Journal of Soil Science* 62 (1):13–22. doi: 10.1111/j.1365-2389.2010.01336.x.

Greenwell, B. M. 2017. Pdp: An R package for constructing partial dependence plots. *The R Journal* 9 (1):421–36. doi: 10.32614/RJ-2017-016.

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

Heumann, S., J. Böttcher, and G. Springob. 2003. Pedotransfer functions for the pool size of slowly mineralizable organic N in sandy arable soils. *Journal of Plant Nutrition and Soil Science* 166 (3):308–18. doi: 10.1002/jpln.200390048.

Hurisso, T. T., S. W. Culman, and K. Zhao. 2018. Repeatability and spatiotemporal variability of emerging soil health indicators relative to routine soil nutrient tests. *Soil Science Society of America Journal* 82 (4):939–48. doi:10.2136/sssaj2018.03.0098.

Hurisso, T. T., D. J. Moebius-Clune, S. W. Culman, B. N. Moebius-Clune, J. E. Thies, and H. M. van Es. 2018. Soil protein as a rapid soil health indicator of potentially available organic nitrogen. *Agricultural & Environmental Letters* 3 (1). doi: 10.2134/ael2018.02.0006.

Jha, P., K. M. Hati, R. C. Dalal, Y. P. Dang, P. M. Kopittke, B. A. Mckenna, and N. W. Menzies. 2022. Effect of 50 years of no-tillage, stubble retention, and nitrogen fertilization on soil respiration, easily extractable glomalin, and nitrogen mineralization. *Agronomy* 12 (1):151. doi: 10.3390/agronomy12010151.

Jones, D. B. 1941. *Factors for converting percentages of nitrogen in foods and feeds into percentages of protein*. Washington, DC: USDA.

Karlen, D. L., K. S. Veum, K. A. Sudduth, J. F. Obrycki, and M. R. Nunes. 2019. Soil health assessment: Past accomplishments, current activities, and future opportunities. *Soil and Tillage Research* 195:104365. doi: 10.1016/j.still.2019.104365.

Kinoshita, R., B. N. Moebius-Clune, H. M. Van Es, W. D. Hively, and A. V. Bilgilis. 2012. Strategies for soil quality assessment using visible and near-infrared reflectance spectroscopy in a western Kenya chronosequence. *Soil Science Society of America Journal* 76 (5):1776–88. doi: 10.2136/sssaj2011.0307.

Laurence, L., B. Heung, H. Strom, K. Stiles, and D. Burton. 2023. Towards a cost-effective framework for estimating soil nitrogen pools using pedotransfer functions and machine learning. *Geoderma* 440:116692. doi: 10.1016/j.geoderma.2023.116692.

Liptzin, D., C. E. Norris, S. B. Cappellazzi, G. M. Bean, M. Cope, K. L. H. Greub, and C. W. Honeycutt. 2022. An evaluation of carbon indicators of soil health in long-term agricultural experiments. *Soil Biology and Biochemistry* 172:108708. doi: 10.1016/j.soilbio.2022.108708.

Liptzin, D., E. L. Rieke, S. B. Cappellazzi, G. M. Bean, M. Cope, K. L. H. Greub, and C. W. Honeycutt. 2023. An evaluation of nitrogen indicators for soil health in long-term agricultural experiments. *Soil Science Society of America Journal* 87 (4):868–84. doi: 10.1002/saj2.20558.

Malone, L. C., M. D. Ruark, C. J. Kucharik, T. Whitman, and S. P. Conley. 2023. Linking soil health indicators to management history and soybean yield. *Field Crops Research* 297:108951. doi: 10.1016/j.fcr.2023.108951.

Martin, T., S. Culman, and C. D. Sprunger. 2022. Quality or quantity? Determining the impact of fine root traits on soil health in row crop agriculture. *Journal of Soil Science and Plant Nutrition* 22 (2):2322–33. doi: 10.1007/s42729-022-00811-1.

Mayer, M., and A. Stando. 2023. Shapviz: Shap visualizations. R package version 0.8.0: https://github.com/ModelOriented/shapviz.

Mayer, M., D. Watson, and B. Prezemyslaw. 2023. Kernalshap: Kernal shap. R package version 0.3.7: https://github.com/ModelOriented/kernelshap.

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: 10.1016/S0016-7061(02)00139-8.

Moebius-Clune, B. N., D. J. Moebius-Clune, B. K. Gugino, O. J. Idowu, R. R. Schindelbeck, A. J. Ristow, and G. S. Abawi. 2017. Comprehensive assessment of soil health - the Cornell framework. Edition 3.2. Geneva, NY. Cornell University. http://soilhealth.cals.cornell.edu/training-manual/.

Moebius-Clune, B. N., H. M. van Es, O. J. Idowu, R. R. Schindelbeck, D. J. Moebius-Clune, D. W. Wolfe, and R. Lucey. 2008. Long-term effects of harvesting maize stover and tillage on soil quality. *Soil Science Society of America Journal* 72 (4):960–69. doi: 10.2136/sssaj2007.0248.

Naasko, K., T. Martin, C. Mammana, J. Murray, M. Mann, and C. Sprunger. 2023. Soil protein: A key indicator of soil health and nitrogen management. *Soil Science Society of America Journal* 88 (1):89–108. doi: 10.1002/saj2.20600.

Norris, C. E., G. M. Bean, S. B. Cappellazzi, M. Cope, K. L. H. Greub, D. Liptzin, and C. W. Honeycutt. 2020. Introducing the north American project to evaluate soil health measurements. *Agronomy Journal* 112 (4):3195–215. doi: 10.1002/agj2.20234.

Nunes, M. R., D. L. Karlen, K. S. Veum, T. B. Moorman, and C. A. Cambardella. 2020. Biological soil health indicators respond to tillage intensity: A US meta-analysis. *Geoderma* 369:114335. doi: 10.1016/j.geoderma.2020.114335.

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

Nunes, M. R., K. S. Veum, P. A. Parker, S. H. Holan, J. P. Amsili, H. M. van Es, and D. L. Karlen. 2024. Shapev1.0 scoring curves for dynamic soil properties. *Soil Science Society of America Journal* 88 (3):858–75. doi: 10.1002/saj2.20668.

Padarian, J., B. Minasny, and A. B. Mcbratney. 2020. Machine learning and soil sciences: A review aided by machine learning tools. *SOIL* 6 (1):35–52. doi: 10.5194/soil-6-35-2020.

Palusynska, A., P. Biecek, and Y. Jiang. 2022. Randomforestexplainer: Explaining and visualizing random forests in terms of variable. R package version 0.10.1. https://github.com/ModelOriented/randomForestExplainer/tree/v0.10.1.

Ramcharan, A., T. Hengl, D. Beaudette, and S. Wills. 2017. A soil bulk density pedotransfer function based on machine learning: A case study with the ncss soil characterization database. *Soil Science Society of America Journal* 81 (6):1279–87. doi: 10.2136/sssaj2016.12.0421.

R Core Team. 2022. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rillig, M. C. 2004. Arbuscular mycorrhizae, glomalin, and soil aggregation. *Canadian Journal of Soil Science* 84 (4):355–63. doi: 10.4141/s04-003.

Roberts, P., and D. L. Jones. 2008. Critical evaluation of methods for determining total protein in soil solution. *Soil Biology and Biochemistry* 40 (6):1485–95. doi: 10.1016/j.soilbio.2008.01.001.

Roper, W. R., D. L. Osmond, and J. L. Heitman. 2019. A response to "reanalysis validates soil health indicator sensitivity and correlation with long-term crop yields". *Soil Science Society of America Journal* 83 (6):1842–45. doi: 10.2136/sssaj2019.06.0198.

Rosier, C. L., A. T. Hoye, and M. C. Rillig. 2006. Glomalin-related soil protein: Assessment of current detection and quantification tools. *Soil Biology and Biochemistry* 38 (8):2205–11. doi: 10.1016/j.soilbio.2006.01.021.

Rubio, V., J. P. Amsili, D. G. Rossiter, A. J. Mcdonald, and H. M. van Es. 2025. Disentangling soil health drivers and predicting spatial land use effects at regional scale. *Geoderma*. In Revision.

Rubio, V., J. Sawchik, and H. M. van Es. 2022. Soil health benefits from sequence intensification, fertilization, and no-tillage in annual cropping systems. *Soil Security* 9:100074. doi: 10.1016/j.soisec.2022.100074.

Sainju, U. M., D. Liptzin, and W. B. Stevens. 2022. Autoclaved citrate-extractable protein as a soil health indicator relates to soil properties and crop production. *Nutrient Cycling in Agroecosystems* 124 (3):315–33. doi: 10.1007/s10705-022-10230-4.

Sanderman, J., K. Savage, and S. R. S. Dangal. 2020. Mid-infrared spectroscopy for prediction of soil health indicators in the United States. *Soil Science Society of America Journal* 84 (1):251–61. doi: 10.1002/saj2.20009.

Sanderman, J., K. Savage, S. R. S. Dangal, G. Duran, C. Rivard, M. A. Cavigelli, and C. Stewart. 2021. Can agricultural management induced changes in soil organic carbon be detected using mid-infrared spectroscopy? *Remote Sensing* 13 (12):2265. doi: 10.3390/rs13122265.

Saxton, K. E., and W. J. Rawls. 2006. Soil water characteristic estimates by texture and organic matter for hydrologic solutions. *Soil Science Society of America Journal* 70 (5):1569–78. doi: 10.2136/sssaj2005.0117.

Schindelbeck, R. R., B. N. Moebiues-Clune, D. J. Moebius-Clune, K. S. Kurtz, M. Ozturk, B. Binkerd-Dale, and H. M. van Es. 2022. Cornell university comprehensive assessment of soil health laboratory standard operating procedures. Ithaca, NY. Cornell University. https://soilhealthlab.cals.cornell.edu/resources/standard-operating-procedures-sops/.

Schindler, F. V., E. J. Mercer, and J. A. Rice. 2007. Chemical characteristics of glomalin-related soil protein (grsp) extracted from soils of varying organic matter content. *Soil Biology and Biochemistry* 39 (1):320–29. doi: 10.1016/j.soilbio.2006.08.017.

Stevenson, F. J. 1982. Organic forms of soil nitrogen. In *Nitrogen in agricultural soils*, ed. F. J. Stevenson, 67–122. Madison, WI: American Society of Agronomy, Inc. Crop Science Society of America, Inc. Soil Science Society of America, Inc.

Stott, D. E. 2019. Recommended soil health indicators and associated laboratory procedures. *Soil Health Technical Note* (450–03. U.S. Department of Agriculture, Natural Resources Conservation District. https://directives.sc.egov.usda.gov/OpenNonWebContent.aspx?content=44475.wba.

van Es, H. M., and D. L. Karlen. 2019. Reanalysis validates soil health indicator sensitivity and correlation with long-term crop yields. *Soil Science Society of America Journal* 83 (3):721–32. doi: 10.2136/sssaj2018.09.0338.

Van Looy, K., J. Bouma, M. Herbst, J. Koestel, B. Minasny, U. Mishra, and H. Vereecken. 2017. Pedotransfer functions in earth system science: Challenges and perspectives. *Reviews of Geophysics* 55 (4):1199–256. doi: 10.1002/2017RG000581.

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 Biology and Biochemistry* 164:108472. doi: 10.1016/j.soilbio.2021.108472.

Wright, S. F., and A. Upadhyaya. 1998. A survey of soils for aggregate stability and glomalin, a glycoprotein produced by hyphae of arbuscular mycorrhizal fungi. *Plant and Soil* 198 (1):97–107. doi: 10.1023/a:1004347701584.

Xiao, Y., J. Xue, X. Zhang, N. Wang, Y. Hong, Y. Jiang, and S. Chen. 2022. Improving pedotransfer functions for predicting soil mineral associated organic carbon by ensemble machine learning. *Geoderma* 428:116208. doi: 10.1016/j.geoderma.2022.116208.

Supplemental figures and tables for:

Pedotransfer Functions for Soil Protein based on Random Forest Models for Routine Soil Health Analysis

Table S1. Number of soil samples in the U.S. Training Dataset by state, region, and soil texture class.

State	n	State	n
AL	3	**MT**	106
AR	45	**ND**	136
CA	329	**NE**	82
CO	1	**NH**	16
CT	45	**NJ**	101
DC	8	**NM**	43
DE	5	**NV**	8
FL	6	**NY**	626
GA	95	**OH**	20
IA	154	**OR**	1
ID	3	**PA**	198
IL	636	**RI**	10
IN	168	**SD**	52
KY	104	**TN**	36
LA	34	**TX**	100
MA	121	**UT**	15
MD	35	**VA**	8
MI	89	**VT**	278
MN	106	**WA**	47
MO	38	**WI**	85
MS	178		

Region	n (%)
Mid-Atlantic/ East Central/ Midwest	484 (12)
Northeast	1,208 (29)
Northwest	1,185 (28)
N. Plain States	51 (1)
S. Plain States	379 (9)
Southeast	100 (2)
Southwest	369 (9)
	395 (9)

Texture	n
Clay	62
Clay loam	236
Loam	991
Loamy sand	131
Sand	42
Sandy loam	720
Sandy clay	3
Sandy clay loam	37
Silt	10
Silt loam	1,292
Silty clay	85
Silty clay loam	562

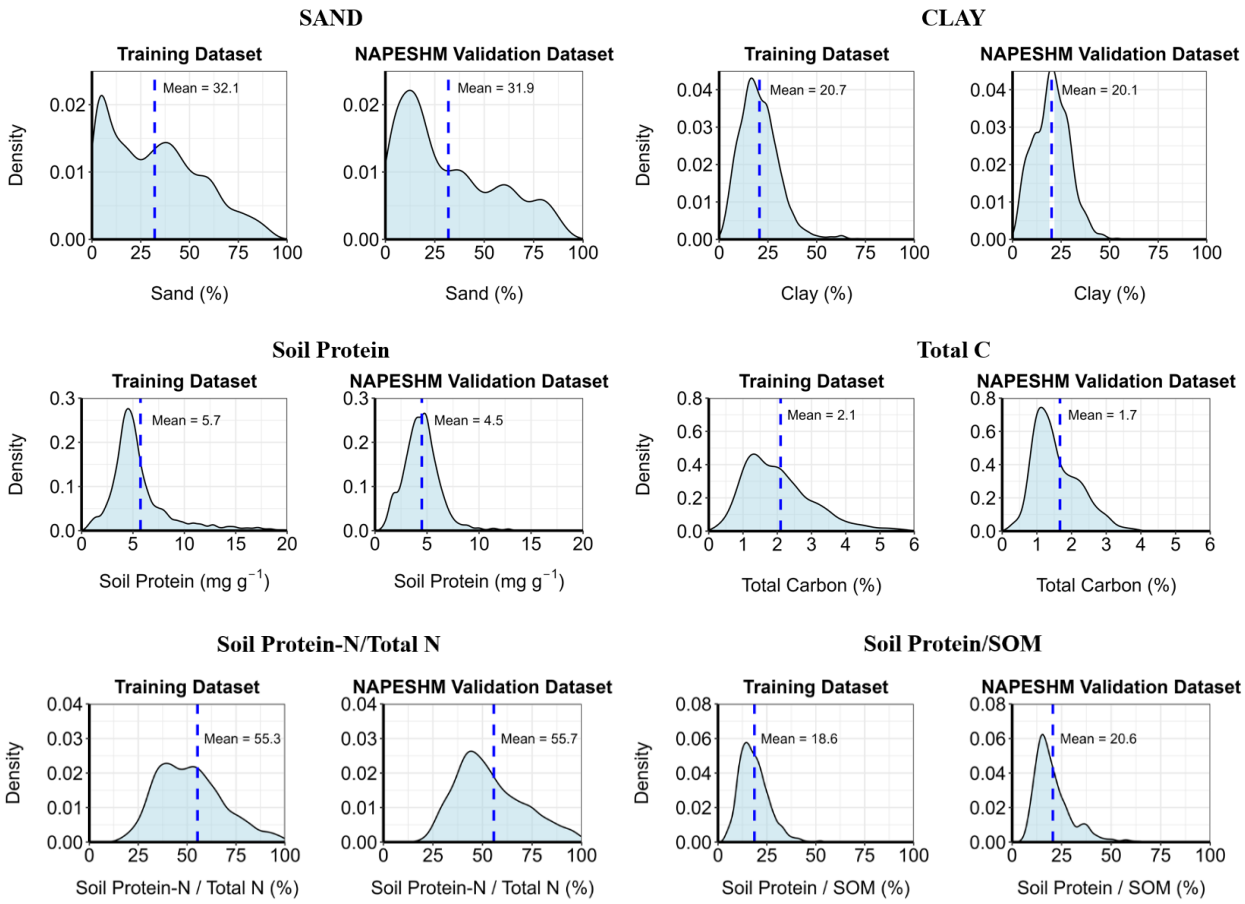

Figure S1. Comparison of attribute space of Training and NAPESHM Validation datasets for sand, clay, autoclaved-citrate extractable soil protein (soil protein), total carbon (C), soil protein-nitrogen (N)/total N, and soil protein/soil organic matter (SOM). Kernal density plots with dashed lines indicating mean values were constructed for sand, clay, soil protein, total C, soil protein-N/total N, and soil protein/SOM in both datasets.

Table S2. Mean values (SD) of autoclaved-citrate extractable soil protein/soil organic matter across ten soil texture classes. Mean values followed by different letters are significantly different at the 0.05 error level.

Soil Texture Class	Training Dataset	NAPESHM Validation Dataset
Sand	47.8a (15.8)	45.3a (7.6)
Loamy sand	33.8b (10.4)	34.2b (9.1)
Sandy loam	24.7c (6.9)	27.9c (12.2)
Loam	18.9d (5.7)	20.6d (7.6)
Silt loam	17.7e (4.6)	18.9d (4.4)
Sandy clay loam	17.9de (6.2)	11.9e (4.3)
Clay loam	12.1f (4.1)	13.7e (4.4)
Silty clay loam	12.0f (3.0)	13.9e (3.3)
Silty clay	8.9g (2.4)	12.7e (3.1)
Clay	8.0g (5.2)	NA

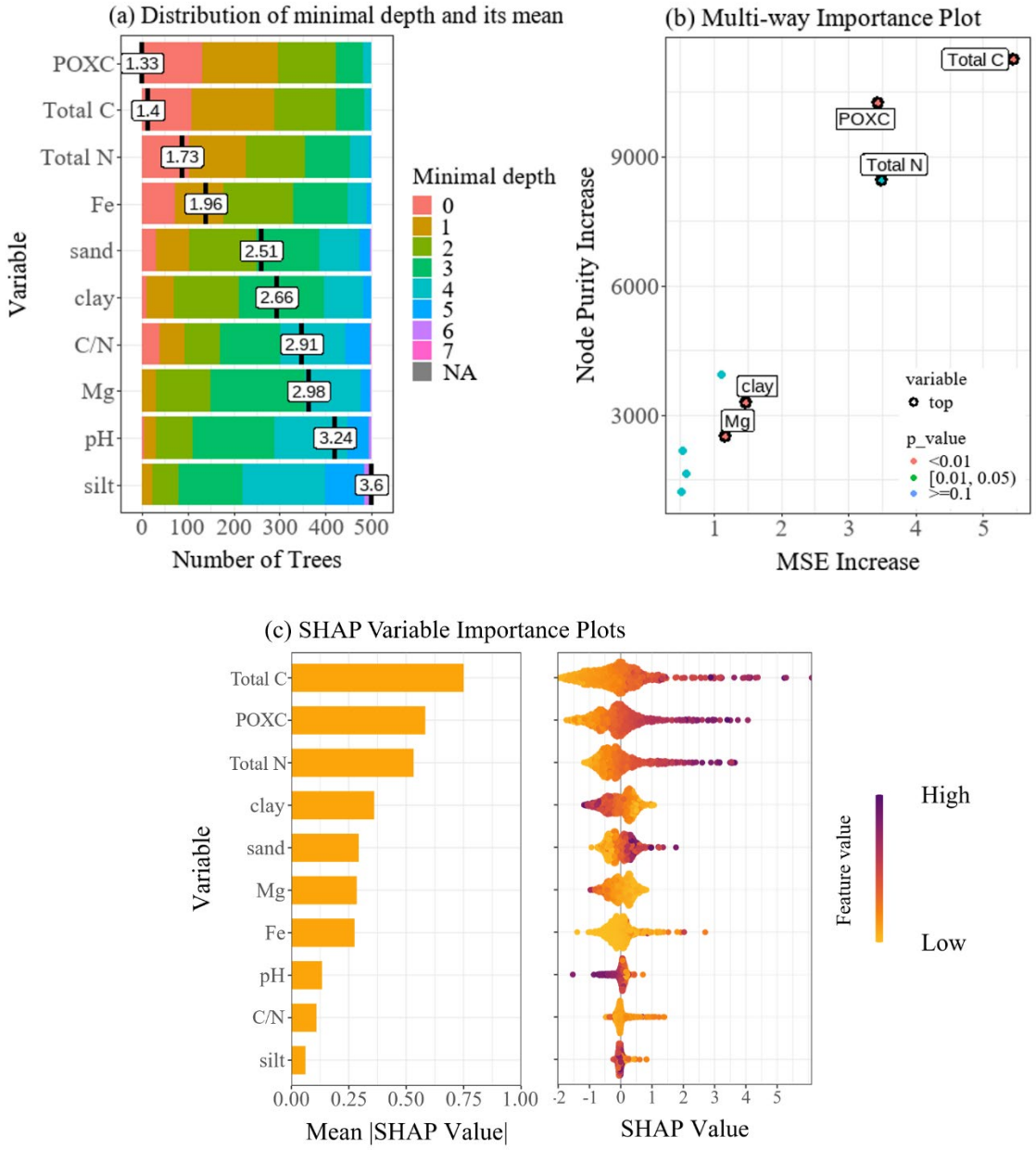

Figure S2. Variable importance plots for the Full RF model for autoclaved-citrate extractable soil protein (soil protein) created using the (a-b) randomForestExplainer and (c) shapviz packages in R. (a) Mean depth distribution of predictor variables in the 500 regression trees that make up the random forest model. (b) Mean squared error (MSE) increase vs. Node purity increase for the predictor variables in the RF model. Higher MSE increase is associated with more important variables in the model (i.e., removing that variable would result in a larger increase in MSE). (c) SHAP variable importance plots; (left panel) shows the ranked variable importance and its mean absolute SHAP value; (right panel) The beeswarm plot shows the relationship between the value of the predictor variable and its positive or negative impact on soil protein.

Table S3. Simple linear regression models between autoclaved-citrate extractable soil protein (soil protein) and soil health indicators used in ACE Protein RF model. R^2_{adj} values (+ indicates positive relationship; - indicates negative relationship) are presented for significant predictors of soil protein at alpha=0.05. *Regression coefficient for soil protein~SOC in the Training Dataset was calculated with a smaller subset of the Training Dataset (n=2,985).

Indicator	Training Dataset	NAPESHM Validation Dataset
Total C	0.48 (+)	0.13 (+)
SOC*	0.54 (+)	0.43 (+)
Total N	0.47 (+)	0.46 (+)
POXC	0.45 (+)	0.52 (+)
C/N	0.03 (+)	0.03 (-)
Clay	0.03 (-)	NS
Mg	0.007 (-)	0.008 (-)
Sand	0.022 (+)	0.400 (-)
Fe	0.014 (+)	0.014 (+)

Table S4. Simple linear regression models between total nitrogen (N) and soil health indicators used in autoclaved-citrate extractable soil protein (soil protein) RF model. R^2_{adj} values (+ indicates positive relationship; - indicates negative relationship) are presented for significant predictors of soil protein at alpha=0.05. *Regression coefficient for total N~SOC in the Training Dataset was calculated with a smaller subset of the Training Dataset (n=2,985).

Indicator	Training Dataset	NAPESHM Validation Dataset
Total C	0.80 (+)	0.42 (+)
SOC*	0.82 (+)	0.88 (+)
POXC	0.69 (+)	0.74 (+)
C/N	NS	0.01 (-)
Clay	0.07 (+)	0.25 (+)
Mg	0.12 (+)	0.07 (+)
Sand	0.07 (-)	0.22 (-)
Fe	0.002 (+)	NS

Table S5. Mean squared error (MSE) Increase and Node Purity Increase values for soil health indicators used in the Full RF model for autoclaved-citrate extractable soil protein. Values correspond to Figure S2b.

Indicator	MSE Increase	Node Purity Increase
Total C	5.44	11,251
Total N	3.49	8,442
POXC	3.43	10,243
Sand	1.52	2,790
Clay	1.49	3,311
Mg	1.17	2,510
Fe	1.11	3,941
pH	0.59	1,642
C/N	0.53	2,182
Silt	0.52	1,234

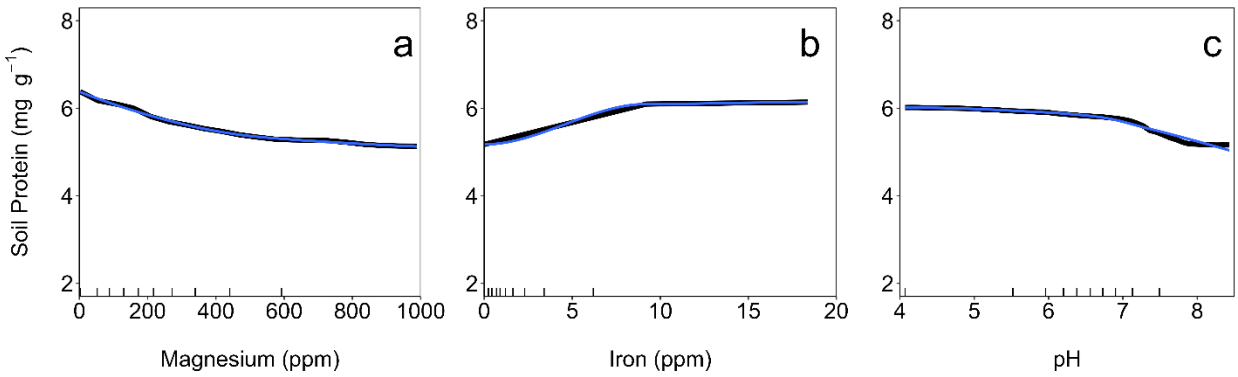

Figure S3. (a-c) Partial dependence plots for individual autoclaved-citrate extractable soil protein (soil protein) predictor variables: Modified Morgan extractable magnesium and iron and pH, within the Full random forest model for soil protein. These partial dependence plots show the influence of individual predictor variables on soil protein within the Full RF model while all other predictor variables are held constant.

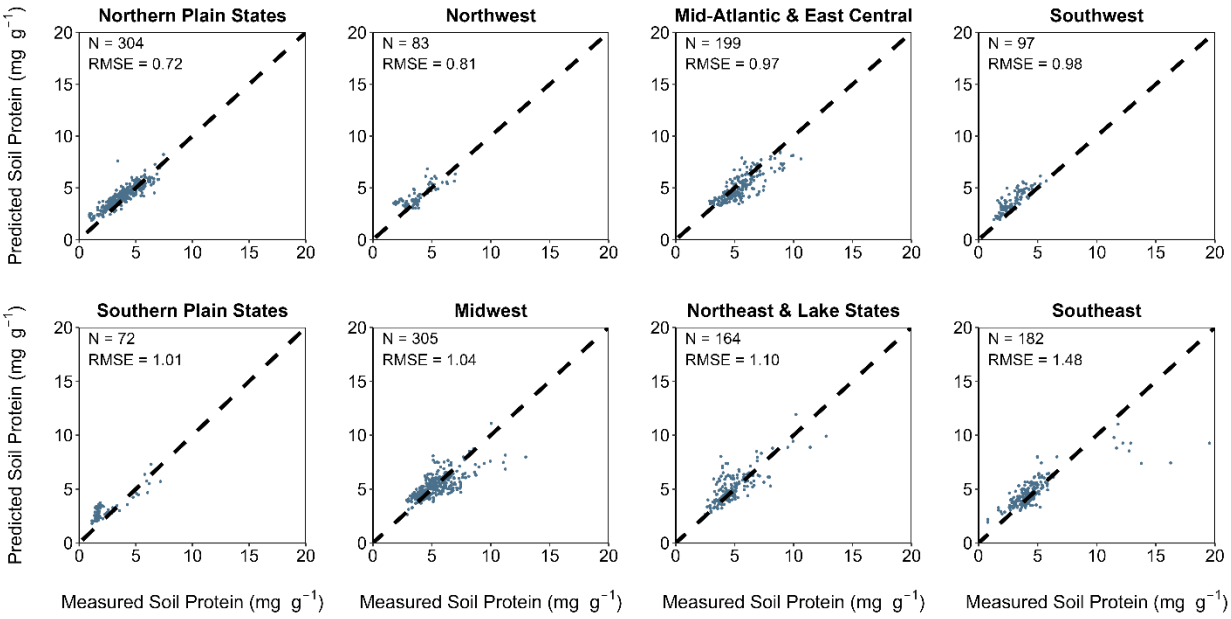

Figure S4. Measured vs. Predicted autoclaved-citrate extractable soil protein (soil protein) for the NAPESHM dataset in eight different regions of the contiguous US.

Table S6. Percent difference in root mean squared error (RMSE) if individual soil health indicators were removed from the Full RF model for autoclaved-citrate extractable soil protein (+ indicates RMSE increased when variable was removed; - indicates RMSE decreased when variable was removed). Values for Training Dataset correspond to Figure S2b and Table S5.

Indicator	Training Dataset	NAPESHM Dataset	NAPESHM Northeast	NAPESHM Southeast	NAPESHM S. Plain Stat	NAPESHM Southwest
Total C	13.0 (+)	2.9 (+)	10.9 (+)	13.5 (+)	18.8 (-)	4.1 (-)
Total N	5.3 (+)	6.8 (+)	0.9 (+)	0.7 (+)	14.9 (+)	69.4 (+)
POXC	4.6 (+)	8.7 (+)	8.2 (+)	0.7 (+)	22.8 (+)	32.7 (+)
Sand, Silt, Clay	9.2 (+)	1.9 (+)	21.8 (+)	10.1 (-)	36.6 (+)	21.4 (-)
Mg	3.1 (+)	6.8 (+)	12.7 (-)	4.1 (+)	3.0 (+)	30.6 (+)
Fe	2.6 (+)	1.0 (+)	2.3 (-)	0.0	27.7 (-)	6.1 (-)
pH	3.7 (+)	6.8 (+)	13.5 (+)	0.3 (-)	20.7 (+)	25.5 (+)
C/N	1.5 (+)	1.7 (-)	1.2 (-)	2.7 (-)	2.0 (-)	5.4 (+)

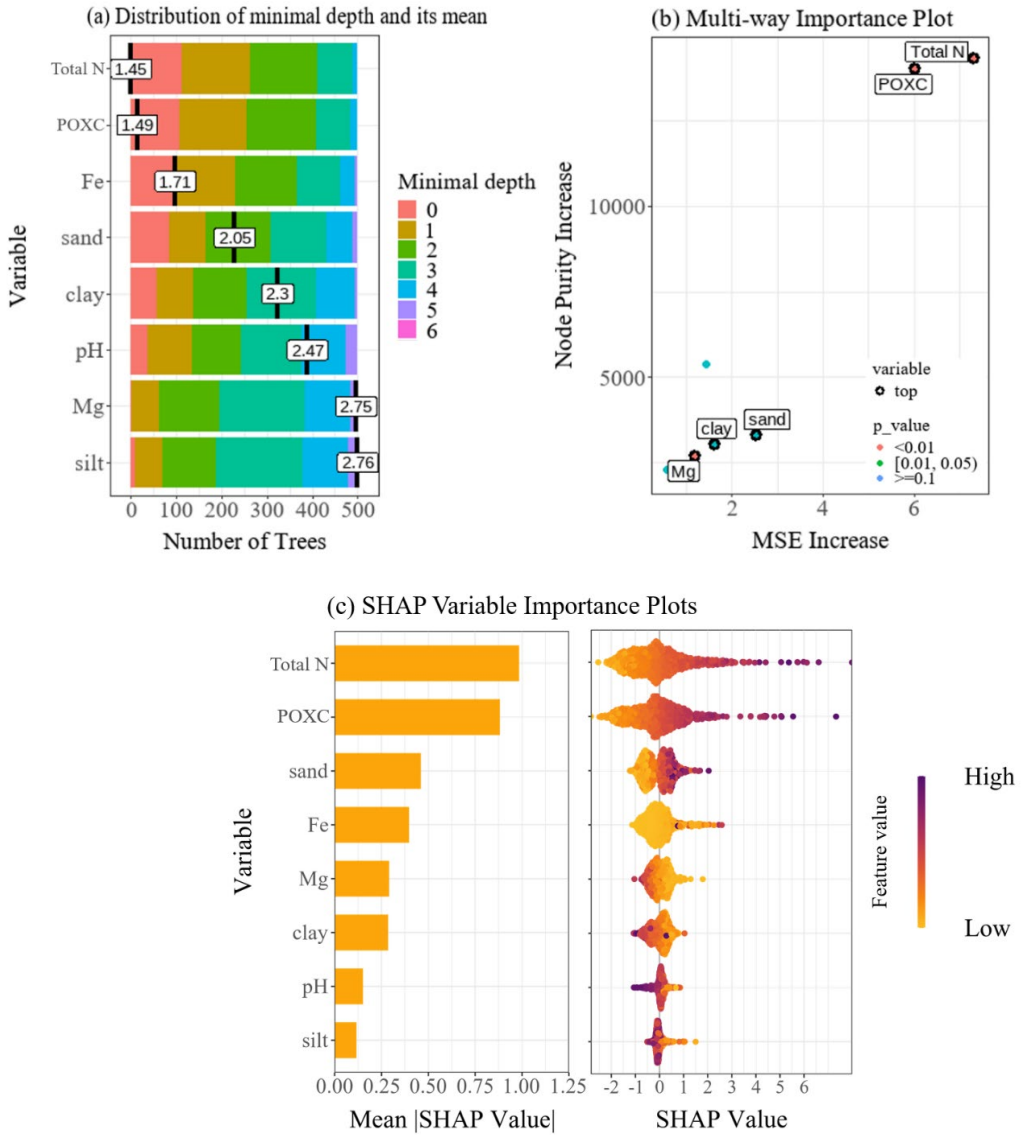

Figure S5. Variable importance plots for Alternative Full RF model for autoclaved-citrate extractable soil protein (soil protein) with no total C created using the (a-b) randomForestExplainer and (c) shapviz packages in R. This model had % of variance explained = 81.23, out of bag root mean square error (OOB RMSE) = 1.48. (a) Mean depth distribution of predictor variables in the 500 regression trees that make up the random forest model. (b) Mean squared error (MSE) increase vs. Node purity increase for the predictor variables in the RF model. Higher MSE increase is associated with more important variables in the model (i.e., removing that variable would result in a larger increase in MSE). (c) SHAP variable importance plots; (left panel) shows the ranked variable importance and its mean absolute SHAP value; (right panel) The beeswarm plot shows the relationship between the value of the predictor variable and its positive or negative impact on soil protein.

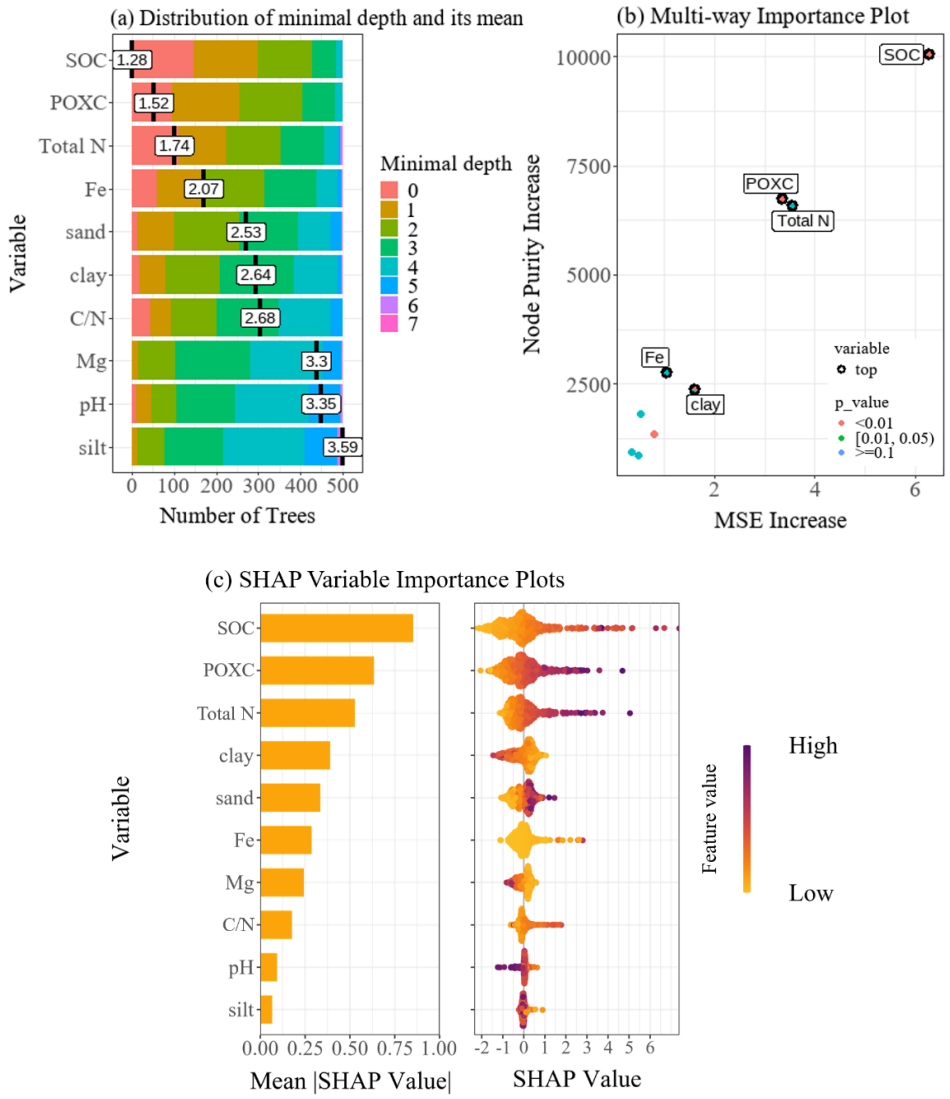

Figure S6. Variable importance plots for Alternative Full RF model for autoclaved-citrate extractable soil protein (soil protein) with soil organic carbon (SOC) instead of total C created using the (a-b) randomForestExplainer and (c) shapviz packages in R. This model had % of variance explained = 84.45, out of bag root mean square error (OOB RMSE) = 1.38. (a) Mean depth distribution of predictor variables in the 500 regression trees that make up the random forest model. (b) Mean squared error (MSE) increase vs. Node purity increase for the predictor variables in the RF model. Higher MSE increase is associated with more important variables in the model (i.e., removing that variable would result in a larger increase in MSE). (c) SHAP variable importance plots; (left panel) shows the ranked variable importance and its mean absolute SHAP value; (right panel) The beeswarm plot shows the relationship between the value of the predictor variable and its positive or negative impact on soil protein.

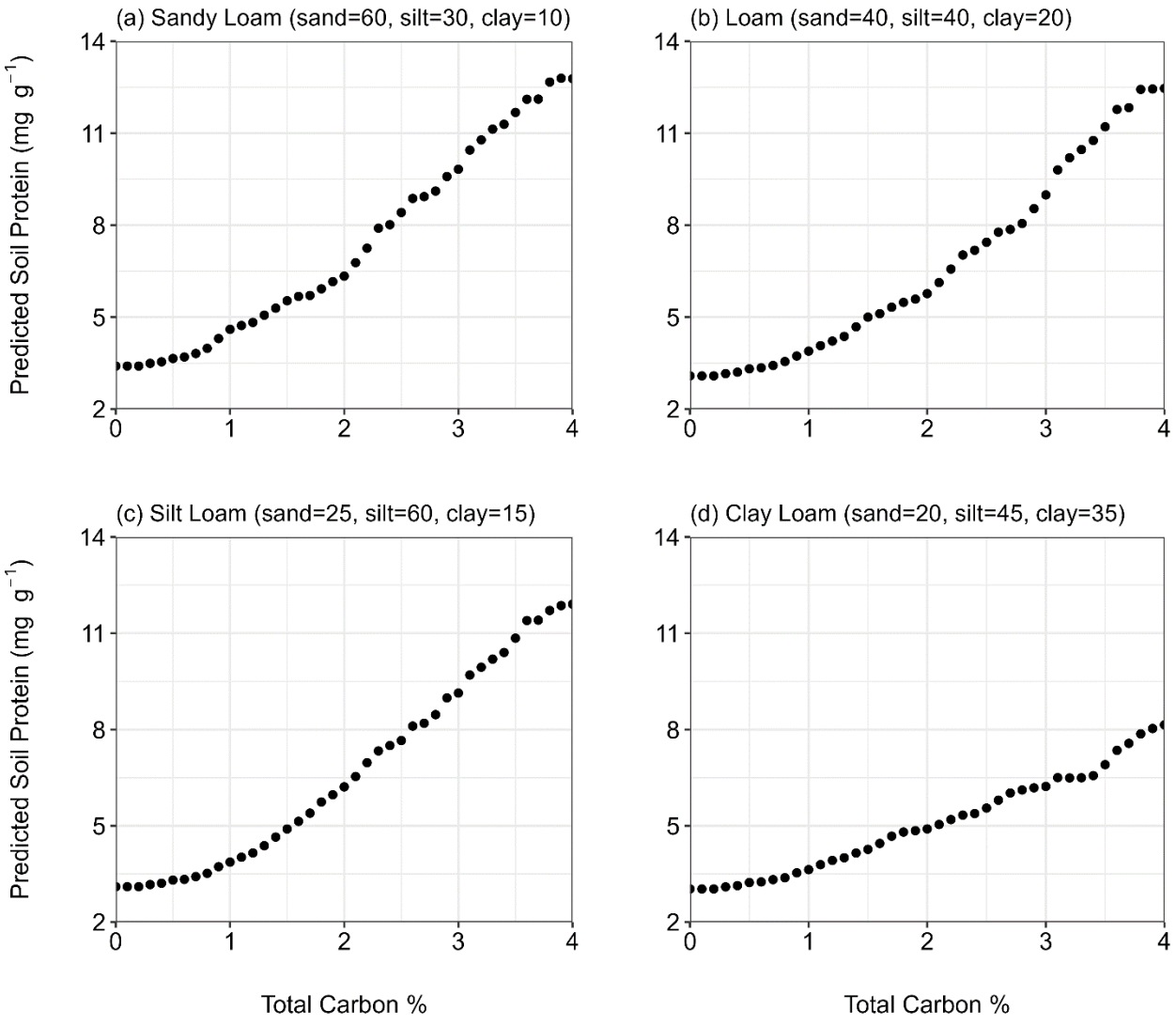

Figure S7. (a-d) Manual partial dependence plots for total carbon within the Full random forest model for autoclaved-citrate extractable soil protein (soil protein) for four different texture classes (sandy loam, loam, silt loam, and clay loam/silty clay loam).

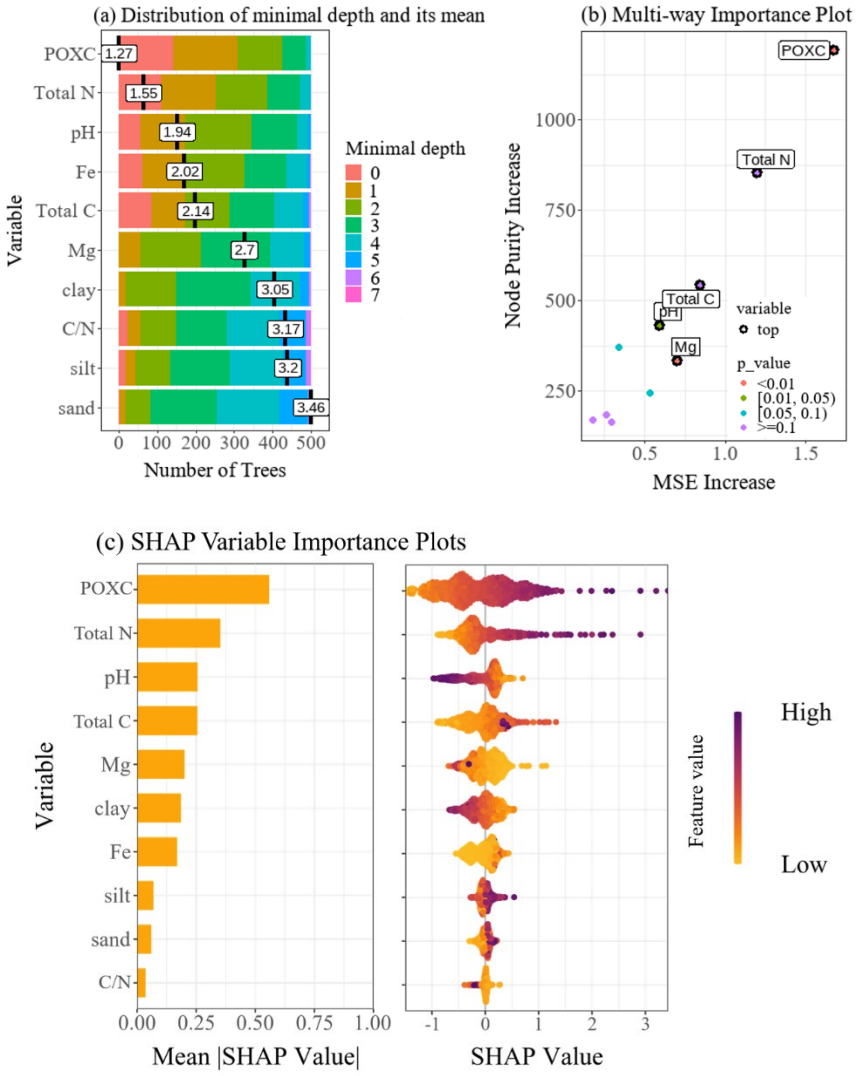

Figure S8. Variable importance plots for Alternative Full RF model for autoclaved-citrate extractable soil protein (soil protein) that was trained on the US NAPESHM dataset using the (a-b) randomForestExplainer and (c) shapviz packages in R. (a) Mean depth distribution of predictor variables in the 500 regression trees that make up the random forest model. (b) Mean squared error (MSE) increase vs. Node purity increase for the predictor variables in the RF model. Higher MSE increase is associated with more important variables in the model (i.e., removing that variable would result in a larger increase in MSE). (c) SHAP variable importance plots; (left panel) shows the ranked variable importance and its mean absolute SHAP value; (right panel) The beeswarm plot shows the relationship between the value of the predictor variable and its positive or negative impact on soil protein.