PNAS RESEARCH ARTICLE | ECOLOGY

Reducing the uncertainty in estimating soil microbial-derived carbon storage

Han Hu[a,b,1], Chao Qian[c,d,1], Ke Xue[c,d,1], Rainer Georg Jörgensen[e], Marco Keiluweit[f], Chao Liang[g,h,2], Xuefeng Zhu[g,h], Ji Chen[i,j,k], Yishen Sun[a,b], Haowei Ni[a,b], Jixian Ding[a], Weigen Huang[a,b], Jingdong Mao[l], Rong-Xi Tan[c,d], Jizhong Zhou[m], Thomas W. Crowther[n], Zhi-Hua Zhou[c,d], Jiabao Zhang[a], and Yuting Liang[a,b,2]

Affiliations are included on p. 7.

Edited by James Brown, The University of New Mexico, Morro Bay, CA; received January 29, 2024; accepted July 22, 2024

Soil organic carbon (SOC) is the largest carbon pool in terrestrial ecosystems and plays a crucial role in mitigating climate change and enhancing soil productivity. Microbial-derived carbon (MDC) is the main component of the persistent SOC pool. However, current formulas used to estimate the proportional contribution of MDC are plagued by uncertainties due to limited sample sizes and the neglect of bacterial group composition effects. Here, we compiled the comprehensive global dataset and employed machine learning approaches to refine our quantitative understanding of MDC contributions to total carbon storage. Our efforts resulted in a reduction in the relative standard errors in prevailing estimations by an average of 71% and minimized the effect of global variations in bacterial group compositions on estimating MDC. Our estimation indicates that MDC contributes approximately 758 Pg, representing approximately 40% of the global soil carbon stock. Our study updated the formulas of MDC estimation with improving the accuracy and preserving simplicity and practicality. Given the unique biochemistry and functioning of the MDC pool, our study has direct implications for modeling efforts and predicting the land–atmosphere carbon balance under current and future climate scenarios.

soil carbon cycle | microbial derived carbon | methodology

Significance

Soil organic carbon (SOC) plays a crucial role in mitigating climate change and enhancing soil productivity, with microbial-derived carbon (MDC) being the main component of the persistent SOC pool. However, the current formulas for estimating MDC storage have several limitations, which reduce the reliability of our estimates of global MDC storage. By using a comprehensive dataset and machine learning approaches, we addressed the limitations of the current formulas and proposed unique formulas. Based on these unique formulas, we estimated that the global MDC contributed approximately 758 Pg. This study has direct significance for modeling efforts to predict total terrestrial carbon storage and has great implications for accurately parameterizing next-generation soil-atmospheric C models.

Soil organic carbon (SOC) is the largest carbon pool in terrestrial ecosystems (1), storing more carbon than vegetation and the atmosphere combined (2, 3). Given its pivotal role in mitigating climate change and preserving soil fertility, it is imperative to understand the intricate mechanisms underlying SOC formation and stabilization (4). Historically, it was believed that most carbon inputs to soil were directly derived from plants (5–7). However, over the past decade, a new understanding has emerged that easily degradable carbon inputs undergo a series of microbial transformations of both catabolism and anabolism (8, 9). Cell debris binds to minerals and stabilizes in soil, forming what is known as microbial-derived carbon (MDC) (3, 5–7, 10). Compared with plant-derived SOC, MDC has a more resistant chemical structure (11, 12) and a greater affinity for minerals and metal oxides (13, 14), making it an important component of the persistent SOC pool (10, 15). A quantitative assessment of MDC (16, 17) and its contribution to SOC is fundamental to understanding SOC stabilization mechanisms (10), with critical implications for predicting terrestrial carbon storage under current and future climate change scenarios (17, 18).

Amino sugar analysis has emerged as the most prevalent and widely accepted method for estimating MDC concentrations in soils (10, 15, 19, 20), with the potential for upscaling standardized in situ measurements to global scales (15, 20, 21). Amino sugars are important biomarkers in microbial cell walls (17) that persist and accumulate in soil after cell lysis (15). Different amino sugars are associated with specific microbial groups (10), allowing for the estimation of bacterial- and fungal-derived carbon (BDC and FDC) concentrations by multiplying the concentrations of muramic acid (MurA) and fungal-derived glucosamine (GlcN) in soils (10, 22), respectively, by conversion factors (23). It is generally accepted that the conversion factors for bacteria and fungi are 45 and 9, respectively (*SI Appendix*, Supporting Information Text 1). However, there remain considerable uncertainties (*SI Appendix*, Supporting Information Text 2 for details) (15, 19, 24, 25). In brief, the limited number of observations of key parameters used to calculate conversion factors introduces inherent uncertainty. More importantly, neglecting global variations in bacterial group composition can also contribute to uncertainties in the estimation of the conversion factors. Constant conversion factors in current formulas were major sources of uncertainty in MDC estimation. Moreover, other factors such as the number and presentiveness of amino sugar samples, climatic, vegetation, and edaphic factors, etc., can also introduce uncertainty. Thus, in addition to updating

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]H.H., C.Q., and K.X. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: cliang823@gmail.com or ytliang@issas.ac.cn.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2401916121/-/DCSupplemental.

Published August 22, 2024.

the conversion factors, we also need to leverage artificial intelligence technology to explore potential contributing factors and reduce the uncertainty of MDC estimation. Addressing these uncertainties is central to our ability to estimate the contributions of MDC to global SOC pool and better understand next-generation SOC studies.

To address the uncertainties in conversion factors and improve MDC quantification, we employed machine learning techniques to refine the formula based on a global data collection. This refinement reduced the uncertainty resulting from sample size and bacterial group composition effects, while preserving the simplicity and feasibility of the formula. Based on this, we estimated the concentration, contribution, and stock of global MDC. By addressing the uncertainty in global MDC estimates, our study provides critical information for modeling and predicting the atmosphere–soil C cycle under climate change.

Results and Discussion

Determination of Amino Sugar Concentrations in Bacterial and Fungal Strains. To address the limitations of previous formulas and reduce uncertainty, we calculated the mean MurA concentrations of the Gram-positive (GP) and Gram-negative (GN) bacterial strains using our expanded dataset (*SI Appendix*, Appendix 1). The results indicated that the mean MurA concentrations of the GP and GN strains were 24.1 mg g^{-1} [95% confidence interval (CI): 21.9 to 26.3, relative standard error (RSE) = 4.7%] and 3.3 mg g^{-1} (95% CI: 3.0 to 3.7, RSE = 5.4%), respectively (Fig. 1*A*). Notably, our estimated mean MurA concentration in the GP strains was significantly greater than that reported for the commonly used formula (23) ($P < 0.05$, *SI Appendix*, Fig. S1). This discrepancy may be attributed to the inclusion of several high MurA-containing GP strains (e.g., *Gaffkya*, *Aerococcus*, etc.) that were not part of the previous limited dataset (23) (*SI Appendix*, Table S1). Furthermore, our results showed that the mean MurA concentration of Actinobacteria (28.4 mg g^{-1}, 95% CI: 24.5 to 32.7) was significantly greater than that of Firmicutes (20.9 mg g^{-1}, 95% CI: 18.8 to 23.3, Fig. 1*A*). This suggests that when calculating the mean MurA concentration of bacteria, it is necessary to consider the concentration and weight of Firmicutes and Actinobacteria separately, as we did for GP and GN in the estimation.

We calculated the bacterial conversion factor by dividing the mean carbon content of the bacterial biomass [~460 mg C g^{-1} dry cell weight (26)] by the mean MurA concentration of the bacterial strains as follows (see details in *SI Appendix*, Supporting Information Text 1):

$$\textit{Bacterial conversion factor} = \frac{460}{20.9 \times a + 28.4 \times b + 3.3 \times (1-a-b)}, \quad [1]$$

where a and b represent the proportions of the Firmicutes and Actinobacteria phyla in the soil bacterial community, respectively.

The method employed to determine the fungal conversion factor followed a similar strategy to that used for bacteria. Our findings revealed that the mean GlcN concentration of fungal strains was 42.7 mg g^{-1} (95% CI: 40.3 to 45.2, RSE = 2.9%, Fig. 1*A*). Since soil GlcN is not solely derived from fungi (15, 27), the conventional method for estimating fungal-derived GlcN concentrations involves subtracting the bacterial GlcN concentration from the total GlcN concentration in soil (10, 15, 27). The former was estimated by assuming a molar ratio of 2 (95% CI: 1.32 to 3.05, RSE = 22.1%) for GlcN to MurA in bacterial cells, estimated from only eight observations in previous formulas (27). Here, we collected data on the molar ratios of GlcN to MurA from 352 bacterial strains (*SI Appendix*, Appendix 4). There were no significant differences in the molar ratios of Firmicutes, Actinobacteria, or GN bacteria ($P > 0.05$ in ANOVA), and the mean molar ratio of the bacterial cells was 1.63 (95% CI: 1.61 to 1.76, RSE = 2.3%, Fig. 1*B*). Therefore, the formula for estimating the FDC is as follows (details in *SI Appendix*, Supporting Information Text 1):

$$\textit{Fungal conversion factor} = \frac{460}{42.7} = 10.8,$$

$$\textit{Fungal necromass C} = \left(\frac{GlcN}{179.17} - 1.63 \times \frac{MurA}{251.23}\right) \times 179.17 \times 10.8$$

$$= (GlcN - 1.16 \times MurA) \times 10.8, \quad [2]$$

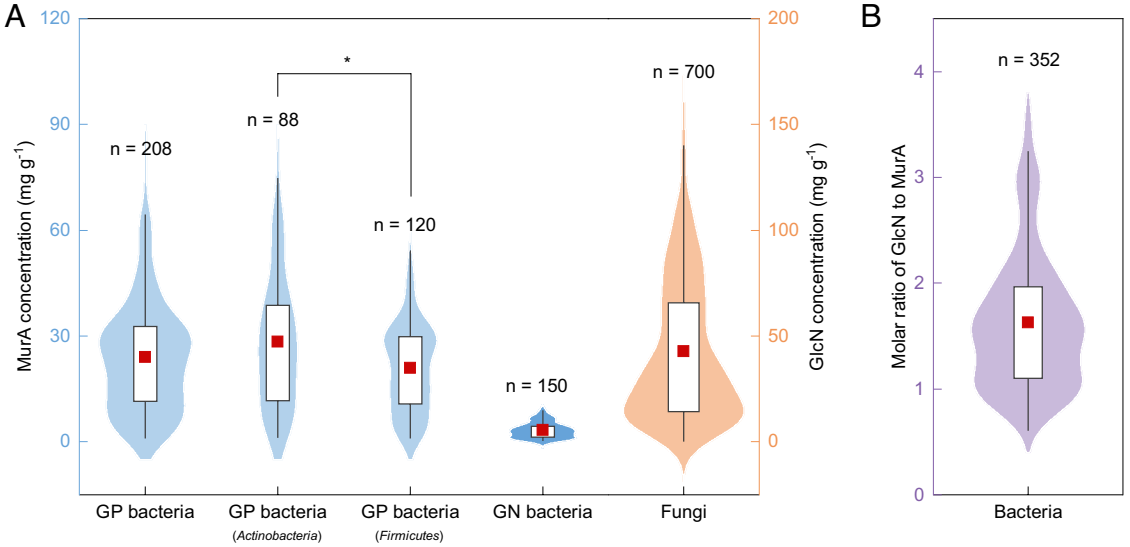

Fig. 1. Muramic acid (MurA) concentrations in bacterial strains, glucosamine (GlcN) concentrations in fungal strains (*A*), and the molar ratio of GlcN to MurA in bacteria (*B*). GP and GN bacteria refer to Gram-positive and Gram-negative bacteria, respectively. The red points represent the mean values, with the box limits indicating the upper and lower quartiles and whiskers extending to 1.5 times the interquartile ranges. The symbol * denotes a significant difference between their averages since their 95% CI do not overlap. Only bacteria possess muramic acid; hence, there are no fungi represented in panel *B*.

where GlcN and MurA are the concentrations of GlcN and MurA in the soil, respectively, and 179.17 and 251.23 are the relative molecular masses of GlcN and MurA, respectively.

Determination of the Optimal Ratio of Bacterial Group Composition. To minimize the impact of global variations in bacterial group composition on predicting BDC concentrations, we utilized machine learning approaches to estimate the global distributions of Firmicutes:GN (R^2 = 0.82, *SI Appendix*, Fig. S2*A*), Actinobacteria:GN (R^2 = 0.76, *SI Appendix*, Fig. S2*B*), MurA (R^2 = 0.81, *SI Appendix*, Fig. S3*A*), and GlcN concentrations (R^2 = 0.94, *SI Appendix*, Fig. S3*B*) in soils. Based on these global maps, we estimated the global distributions of BDC concentrations using Eq. **1**. Sliding window analysis was employed (see *Materials and Methods*), and the R^2 between the estimated BDC map and the predicted BDC maps reached a maximum when the assumed Firmicutes:Actinobacteria:GN ratio was 0.48:0.12:0.40 (Fig. 2*A*). The results indicated that when this assumed Firmicutes:Actinobacteria:GN ratio was used, the global average error in the predicted BDC was 7.5%, and only 27% of the area had a relatively large error (>10%) (*SI Appendix*, Fig. S4*A*). In contrast, when using the current prevailing ratio (GP:GN = 0.65:0.35) (23), the predicted BDC had an average error of 12.5% globally, and more than 50% of the area had a large error (*SI Appendix*, Fig. S4*B*). Furthermore, we found that the predicted values could explain 97.6% and 99.9% of the global changes in BDC and MDC, respectively, assuming a Firmicutes:Actinobacteria:GN ratio of 0.48:0.12:0.40 (*SI Appendix*, Fig. S5). Therefore, we suggest replacing the current prevailing ratio with a new ratio (0.48:0.12:0.40) to estimate the bacterial conversion factor to improve confidence in estimating the BDC concentration. Overall, the formula for estimating the BDC is as follows:

$$\textit{Bacterial conversion factor} = \frac{460}{20.9 \times 0.478 + 28.4 \times 0.119 + 3.3 \times 0.403} = 31.3,$$

$$\textit{Bacterial necromass C} = \textit{MurA} \times 31.3. \quad [3]$$

Albert Einstein once said, "Everything should be made as simple as possible, but not simpler". Although the constant bacterial conversion factor (31.3) we reported here was shown to predict the global distribution of BDC concentrations well, our growing understanding of microbial biogeography (28, 29) highlights the need to consider regional differences in conversion factors in some cases. Therefore, we employed sliding window analysis to determine the optimal bacterial conversion factor for each specific ecosystem under each climate, ranging from 28.8 (tropical wetland) to 34.2 (cold glacier) (*SI Appendix*, Table S2). As such, for

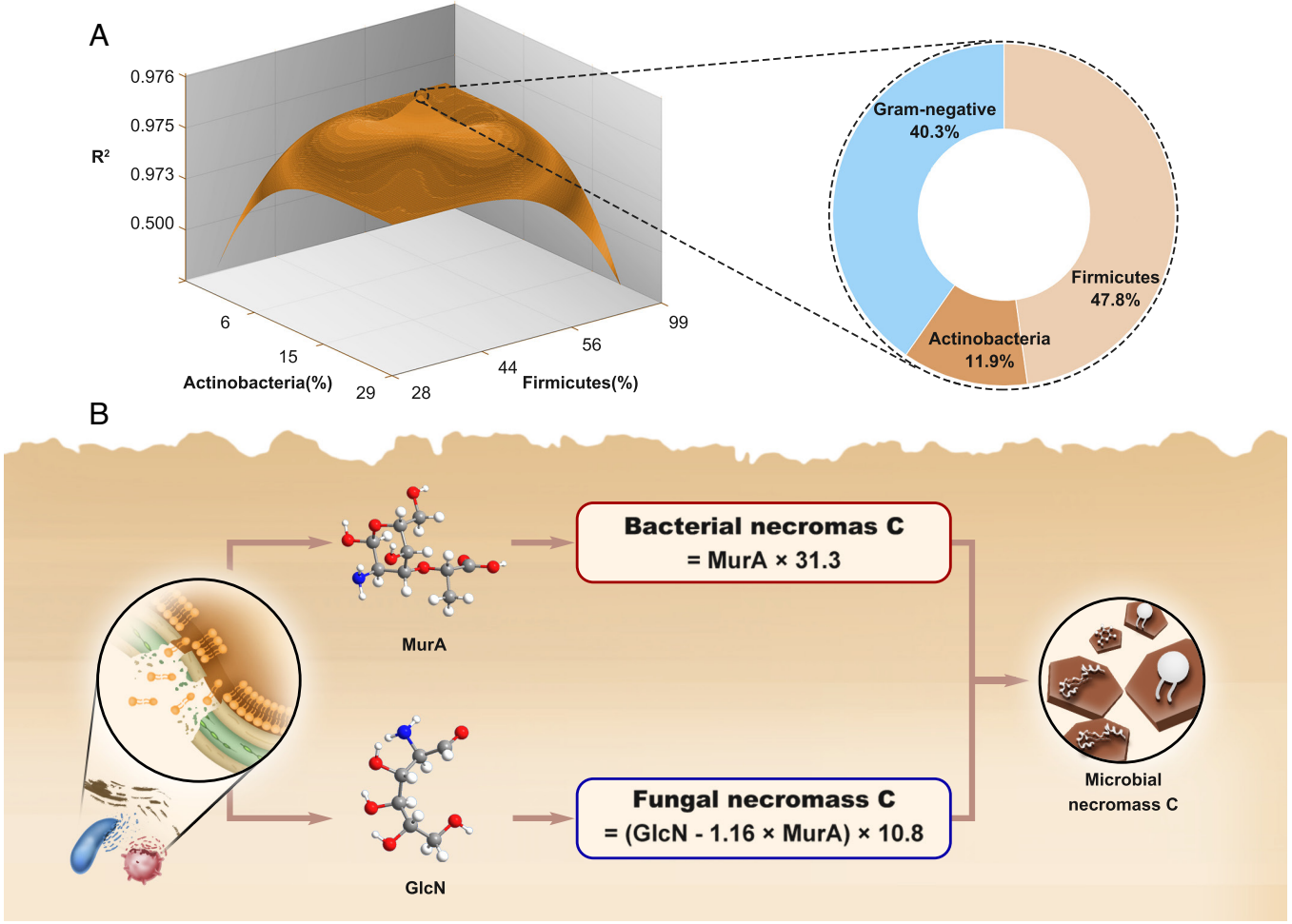

Fig. 2. Determination of formulas for bacterial- and fungal-derived carbon. (*A*) Determination of the optimal Firmicutes:Actinobacteria:GN ratio for minimizing the impact of global variations in bacterial group composition on estimating bacterial-derived carbon. The data on the x-, y-, and z-axes are nonlinearly distributed. Please refer to the *Materials and Methods* section for more detailed information. (*B*) The formulas for estimating bacterial- and fungal-derived carbon. MurA and GlcN represent the concentrations of MurA and GlcN in the soil, respectively.

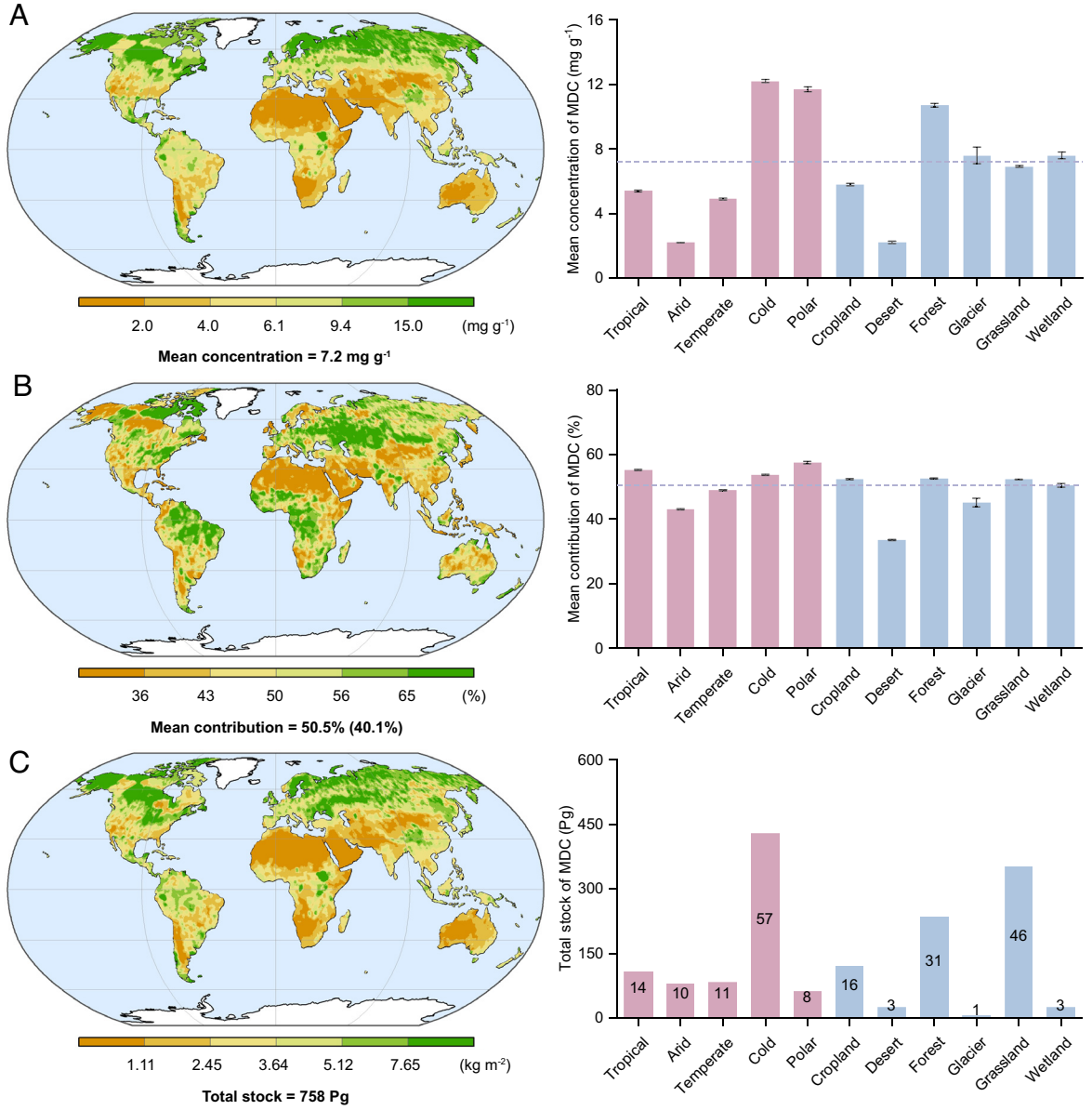

Fig. 3. Global distributions of concentrations (*A*), contributions (*B*), and stocks (*C*) of microbial-derived carbon (MDC). The data in the histograms are presented as the mean ± the SE across climates (in red) and ecosystems (in blue). The purple dotted lines indicate the global mean. In panel *B*, the number in brackets indicates the weighted mean contribution, while the number outside brackets indicates the arithmetic mean contribution. In panel *C*, numbers in columns indicate the percentage of total global MDC stock (unit in percentage). All maps have a spatial resolution of 5 arcminutes (~10 km).

some studies, particularly those involving cross-climatic research, we recommend using a variable bacterial conversion factor based on the specific ecosystem and climate rather than relying solely on a constant bacterial conversion factor (31.3).

It is important to note that there are still limitations in converting soil amino sugar concentrations into MDC concentrations. Certain peptidoglycans resist degradation, resulting in the preservation of amino sugars within their intact cell wall (10, 30, 31). This preservation is supported by the consistent 1:1 ratio of MurA to d-alanine found in bacterial peptidoglycan, although at different positions than MurA (10, 31). Moreover, even within the same bacterial strain, bacteria in substrate-limited soils may exhibit higher MurA concentrations than those in nutrient-rich soils due to the smaller cell sizes of starving soil bacteria (21). This could lead to an overestimation of the bacterial conversion factor in nutrient-poor soils, resulting in the contribution of MDC to SOC exceeding 100% in these areas. Our analysis showed that the area with MDC contributions exceeding 100% accounted for 1.0% of the global area according to the previous formulas but only 0.5% according to the revised formulas. This disparity indicates that our study partially mitigates this constraint, yet additional investigation is warranted. Addressing these limitations remains challenging.

Estimation of the Global MDC Stock. We estimated the global distributions of MDC concentrations, contributions, and stocks in topsoils (0 to 30 cm) and subsoils (30 to 100 cm) (*SI Appendix*, Figs. S6–S8). Our findings indicated that the global mean concentration of total MDC was 11.3 g kg^{-1} in topsoils, and its contribution to SOC was 43.3% in topsoils, with bacteria contributing 12.5% and fungi contributing 30.9%. In subsoils, the mean concentration of total MDC was 5.5 g kg^{-1} on a global scale, contributing 57.6% of the total SOC, with bacteria contributing 24.5% and fungi contributing 33.1%. We estimated the global

total MDC stock to be approximately 758 Pg, with 311 Pg stored in topsoils (67 Pg for bacteria and 244 Pg for fungi) and 447 Pg stored in subsoils (130 Pg for bacteria and 317 Pg for fungi).

The contribution of MDC to SOC in tropical climates (55.2%) was comparable to that in cold climates (54.2%). However, the bulk of the MDC stocks remains predominantly in high-latitude boreal regions due to the large SOC pool (Fig. 3). Our estimates showed that the average MDC concentration was 12.2 mg g^{-1} in cold and polar climates, representing 490 Pg MDC or 65% of the total global MDC stock (Fig. 3). This may be attributed to the slow turnover of microbial biomass at low temperatures (32). It is important to note that these regions store the most substantial amount of MDC but are experiencing warming at rates faster than at lower latitudes (33). As a consequence, the stored MDC is subject to decomposition, potentially exerting considerable impacts on the atmospheric carbon pool. Furthermore, the replenishment of MDC stocks is projected to require a prolonged period, potentially spanning several decades or even longer (10, 15). The MDC loss in these regions may have lasting effects on local ecosystems. Therefore, accurately assessing the stocks of MDC is crucial for developing effective policies and implementing measures to counteract the challenges posed by climate change.

Taken together, we have refined the formula for quantifying MDC concentrations. We estimated the MDC contributions to SOC at an average of 50.5% globally and the global total MDC stock at approximately 758 Pg. Our results not only enhance the precision of MDC estimation but also preserve the simplicity and practicality of the estimation formulas. This advancement is crucial for several reasons. First, it enables more accurate parameterization of next-generation microbial models, which is essential for predicting SOC dynamics. Second, it contributes to a deeper comprehension of the soil-atmosphere carbon cycle, offering insights that could help mitigate climate change by promoting atmospheric CO_2 sequestration on land. Moreover, the integration of AI techniques in our study demonstrates their significant potential in the field of soil science. These techniques have been instrumental in quantifying global patterns of amino sugars and microbial group composition, showcasing the vast applicability of AI in advancing our understanding of soil ecosystems.

Materials and Methods

Compilation of the Global Database. To compile a comprehensive database, we utilized the Web of Science (http://apps.webofknowledge.com), Google Scholar (https://scholar.google.com), and the China National Knowledge Infrastructure Database (http://www.cnki.net) to search for peer-reviewed articles published before April 1, 2023. This process resulted in the collection of six datasets.

The first dataset included the MurA concentrations of various bacterial strains. The search was conducted using the keywords "bacteria" and "muramic acid". A total of 358 bacterial strains, including 120 Firmicutes, 88 Actinobacteria, and 150 GN bacteria, were extracted from 57 articles. The MurA concentrations of these strains are expressed in terms of cell dry weight. For bacterial strains for which only MurA concentrations in cell walls were reported, we converted these values to cell dry weight using a conversion factor of 0.465 (23). After calculating the mean value, outliers were identified as data points falling outside 1.5 times the interquartile range. Accordingly, four outliers for Actinobacteria, six outliers for Firmicutes, and eight outliers for GN bacteria were removed from the box plots. Please refer to *SI Appendix*, Appendix 1 for further details on this dataset.

The second dataset included data on GlcN concentrations in fungal strains. The search was conducted using the keywords "fungi" and "glucosamine or hexosamine". A total of 700 fungal strains were collected from 123 articles, providing their GlcN concentrations in terms of cell dry weight. For fungal strains in which only GlcN concentrations in cell walls were reported, we converted these values to cell dry weight using a conversion factor of 0.2 (23). After calculating the mean value, outliers were identified as data points falling outside 1.5 times the interquartile range. Accordingly, ten outliers were removed from the box plot. Please refer to *SI Appendix*, Appendix 1 for further details on this dataset.

The third and fourth datasets include data on MurA and GlcN concentrations in soils on a global scale. These datasets were compiled from six previous meta-analyses focusing on soil MurA and GlcN concentrations (please refer to *SI Appendix*, Appendix 2 for detailed information). To avoid duplication, the same articles from different meta-analyses were included only once in our dataset. Additionally, data on environmental variables were collected. In total, we gathered 1,604 and 1,636 observations of MurA and GlcN concentrations in soils, respectively. Data on latitude, longitude, elevation, mean annual air temperature (MAT), mean annual precipitation (MAP), mean annual potential evaporation (PET), ecosystem type, SOC, soil total nitrogen (TN), soil total phosphorus (TP), microbial biomass C and N (MBC, MBN), pH, and soil texture (clay, silt, and sand) were also collected. Please refer to *SI Appendix*, Appendix 2 for detailed data.

The fifth dataset includes data on the Firmicutes:GN and Actinobacteria:GN ratios in soil bacterial communities from various locations worldwide, including our own experiments. The search was conducted using the keywords "bacteria," "ratio," and "Gram." To minimize publication bias, the data were screened based on specific criteria. First, bacterial biomass was measured using phospholipid fatty acids (PLFAs), and bacterial group identification followed the methods of Joergensen (34). Second, the latitude and longitude of the soil collection sites were reported. Third, laboratory-incubated soils were excluded unless the blank treatment soil was used in the incubation experiments. Fourth, pot experiments were not considered. Fifth, only undisturbed soils were included, and sieved soil was excluded. Sixth, plant litter layers were not taken into account. The data were extracted from the graphs using GetData software (v.2.22). Additionally, our own data from cropland soils, comprising 414 observations, were incorporated into this dataset. In total, the dataset comprises 3,063 and 2,066 observations of Firmicutes:GN and Actinobacteria:GN ratios, respectively, from different global locations. Please refer to *SI Appendix*, Appendix 3 for detailed information on the dataset. The PLFA approach enables the differentiation of bacteria into Gram-positive (GP) and Gram-negative (GN) categories, with further capability to distinguish GP bacteria into specific groups such as Firmicutes and Actinobacteria (34). Importantly, these bacterial groups can be identified using the conventional PLFA method without necessitating any extra experiments. This means that our refined estimation formulas preserve the simplicity of the original models, ensuring their continued applicability and ease of use in future studies.

The sixth dataset includes data on the molar ratio of GlcN to MurA in bacterial strains. The search was conducted using the keywords "bacteria" and "muramic acid." The molar ratios of a total of 353 bacterial strains, comprising 216 GP, 124 GN, and 12 unidentified strains, were collected from 115 articles. For those bacterial strains for which only the MurA and GlcN concentrations in the cell dry weight or cell walls were reported, we calculated their molar ratios based on their relative molecular masses (179.17 for MurA and 251.23 for GlcN). After calculating the mean value, outliers were identified as data points falling outside 1.5 times the interquartile range. Accordingly, thirty outliers were removed from the box plot. Please refer to *SI Appendix*, Appendix 4 for further details on this dataset.

Furthermore, for some observations in the third, fourth, and fifth datasets, the environmental variables were not reported in the text. To address this issue, missing data were filled in using the following global databases:

- The SOC, TN, TP, pH, and soil texture were obtained from the gridded Global Soil Dataset at a 0.083° spatial resolution. (http://globalchange.bnu.edu.cn/research/soilw).
- MBC and MBN were obtained from the Oak Ridge National Laboratory Distributed Active Archive Center (https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1264).
- PET data were obtained from the CGIAR Data Science Academy (https://cgiarcsi.community/data/global-aridity-and-pet-database).

Prediction of the Global distributions. To minimize the effect of global differences in bacterial group composition on the estimation of BDC, we used machine learning to map the global distributions of Firmicutes:GN, Actinobacteria:GN, MurA, and GlcN concentrations in soils. The proposed machine learning approach

consists of four steps, namely, data preprocessing, model pool construction, hyperparameter optimization, and model training and selection, which are introduced as follows.

Data preprocessing. The tasks of predicting the global distributions are tabular regression tasks, where the four labels to be predicted (i.e., the output of a machine learning model) are ln(Firmicutes:GN) (logarithmic form for better prediction), ln(Actinobacteria:GN), MurA, and GlcN. These tasks have 19 variables (also known as features, i.e., the input of a machine learning model), including 13 soil variables (SOC, TN, TP, SOC/TN, SOC/TP, TN/TP, MBC, MBN, MBC/MBN, pH, clay, silt, sand), 3 climatic variables (MAT, MAP, PET), elevation, latitude, and ecosystem type. Among these variables, only "ecosystem type" is categorical. We handled it by using a widely adopted technique called one-hot encoding, which transforms each category value into a 0–1 vector, thereby enhancing the model's predictive performance. For each task, we randomly split the data into a training set and a test set (70%:30%).

Model pool construction. We built a pool of machine learning models that included as many representative models as possible. Currently, regression methods are primarily divided into two categories: classical models and deep neural network models (35). Within classical models, tree ensemble methods have achieved impressive results in a series of practical applications and competitions (e.g., KDD Cup and Kaggle competitions) (36). Hence, we selected six representative tree ensemble methods, including Random Forest (37), extreme gradient boosting model-XGBoost (38), light gradient boosting machine model-LightGBM (39), category gradient boosting decision trees model-CatBoost (40), Deep Forest (41), and Auto-Sklearn (42). We also selected the classical model, multilayer perceptron-MLP, which is the basis of neural networks. For deep neural networks, we used three models, convolutional neural networks-CNN (43), ResNet (44), and feature tokenizer Transformer-FT-Transformer (45), which have different network structures.

Hyperparameter optimization. The hyperparameters of machine learning models significantly impact their performance; thus, their optimization is necessary. We used an advanced hyperparameter optimization algorithm, Bayesian optimization (46), in this work. Bayesian optimization is a sample-efficient approach for expensive black-box optimization (with applications including hyperparameter tuning, experimental design, and protein design) that approximates the objective function by a Gaussian process surrogate model and then selects the most valuable point for evaluation by optimizing an acquisition function based on the posterior of the surrogate model. It can utilize all historical evaluation information about the objective function and has the ability to automatically decide where to explore the search space and where to prune. Recent works have shown that Bayesian optimization is efficient for hyperparameter optimization [e.g., tuning the hyperparameters of AlphaGo (47) and tuning the Swiss Free Electron Laser (48)] and has won many machine learning hyperparameter tuning competitions (49). Thus, in this work, the hyperparameters of a machine learning model are tuned by the popular Bayesian optimization module Optuna, where the search space is the space of all hyperparameter configurations of the machine learning model, and the objective function is defined as the mean squared error (MSE) of a fivefold cross-validation to be minimized. The experimental results in *SI Appendix*, Table S3 also show that Bayesian optimization is more suitable than grid search for our task.

Model training and selection. After the model hyperparameters are determined, we select a good model for each task. To mitigate the impact of training randomness, we repeated the training processes of each model multiple times (e.g., five times in our experiments) and conducted a statistical significance test to compare these trained models, as shown in *SI Appendix*, Table S4. Specifically, on each task, a t-test with a confidence level of 0.05 is performed between the best-performing model (i.e., having the highest R^2 and the lowest MSE) and each other model, and those models that are almost equivalent (i.e., have no significant difference) to the best-performing model are selected as good candidate models. If a good candidate model is unique, we select it directly as the final model. If there are multiple good models whose performance has no significant difference, we further perform explainable model analysis by Kernel SHAP (which can estimate the importance of each feature for the prediction) (50) and finally make an integrated decision based on model performance and feature importance with knowledge and experience in soil science.

The results in *SI Appendix*, Table S4 show that ResNet is significantly better than all the other models on the ln(Firm:GN) task, and the FT-Transformer is significantly better than all the other models on the GlcN task. Thus, ResNet and FT-Transformer are selected as the final models for these two tasks. However, there are models that are almost equivalent to the best-performing model on the tasks ln(Act:GN) and MurA, and model selection relies on explainable model analysis. For the ln(Act:GN) task, ResNet and FT-Transformer are two good candidate models (*SI Appendix*, Table S4) that recognize feature elevation and soil pH as the most important, respectively, as shown in *SI Appendix*, Fig. S9*A*. Because the bacterial group composition is more likely to be influenced by soil pH than by elevation based on the knowledge and experience in soil science, the FT-Transformer was chosen for the task ln(Act:GN). For the MurA task, CatBoost, Deep Forest, and FT-Transformer are three good candidate models (*SI Appendix*, Table S4). Because soil nutrients (for example, TN, TP and their stoichiometry) are generally considered to be the environmental factors closely linked to amino sugars, the Deep Forest model was chosen for the task MurA, which recognizes the N:P (ratio of TN to TP), TN, and TP as the three most important factors according to Kernel SHAP, as shown in *SI Appendix*, Fig. S9*C*.

Detailed information on the above four steps is provided in *SI Appendix*, Supporting Information Text 3. The results showed that the most suitable models for predicting the global distributions of ln(Firmicutes:GN), ln(Actinobacteria:GN), MurA, and GlcN were ResNet, FT-Transformer, Deep Forest, and FT-Transformer, respectively. All mapping was performed using ArcGIS. We also provided the code, Python libraries, and installation guidelines in *SI Appendix, Supporting File* to make our results easier to reproduce and to make our approach easier to apply to other prediction tasks in scientific research. Please see *SI Appendix*, Supporting Information Texts 4 and 5 for the systematic difference verification and reproducibility, respectively.

Determination of the Optimal Ratio. First, we transformed the global distributions of Firmicutes:GN and Actinobacteria:GN into those of the bacterial conversion factor (based on Eq. **1**). We multiplied the global distribution of the bacterial conversion factor with that of the soil MurA concentration in the form of a raster to estimate the global distribution of BDC (hereafter referred to as Map A). Second, we assumed a Firmicutes:Actinobacteria:GN ratio and input this assumed ratio into Eq. 1, resulting in a constant value for the bacterial conversion factor at a global scale. We multiplied the global distribution of the soil MurA concentration by the constant bacterial conversion factor in the form of a raster to obtain the global distribution of the assumed BDC (hereafter referred to as Map B). Third, we divided the global Firmicutes:GN map by the global Actinobacteria:GN map in the form of a raster to obtain a global Firmicutes:Actinobacteria map. Based on the mean MurA concentrations of Firmicutes (20.9 mg g^{-1}) and Actinobacteria (28.4 mg g^{-1}), we transformed the global map of Actinobacteria:GN into that of the weighted average MurA concentrations of GP bacteria. Then, we assumed a GP:GN ratio and entered this assumed ratio into the equation in *SI Appendix*, Supporting Information Text 1 to obtain a global distribution of the bacterial conversion factor. We multiplied the global distribution of the soil MurA concentration by that of the bacterial conversion factor in the form of a raster to obtain the global distribution of the assumed BDC (hereafter referred to as Map C).

In brief, the data on Map A were calculated based on the results of machine learning (Real data), while the data on Maps B and C were calculated based on the assumed Firmicutes:Actinobacteria:GN ratio (Assumed data). As the assumed Firmicutes:Actinobacteria:GN ratio changed, the global distributions of Maps B and C changed, resulting in varying R^2 values between Maps A and B and between Maps A and C. When the product of these two R^2 ratios is maximized, the assumed Firmicutes:Actinobacteria:GN ratio represents the optimal ratio that can minimize the effect of global differences in bacterial group composition on the estimation of the BDC. Sliding window analysis was employed to determine the optimal Firmicutes:Actinobacteria:GN ratio (step length = 0.01). The results showed that when the assumed Firmicutes:Actinobacteria:GN ratio was 0.441:0.152:0.407, the R^2 (that is, the R^2 in the z-axis in Fig. 2*A*) was maximized.

Estimation of Carbon Stocks. The carbon stock can be estimated as follows (51):

$$C\ stock = \frac{C\ concentration}{1{,}000} \times SLT \times BD \times \frac{100 - sand}{100},$$

where C stock is the carbon stock (unit in kg m^{-2}), C concentration is the carbon concentration (unit in g kg^{-1}), SLT is the soil layer thickness (unit in m), BD is the bulk density (unit in kg m^{-3}), and sand is the volumetric fraction of sand in soils (unit in percent).

Data, Materials, and Software Availability. All study data and code are included in the article and/or supporting information.

ACKNOWLEDGMENTS. We received funding from the National Natural Scientific Foundation of China (42377121), the National Key Research and Development Program of China (2021YFD1900400), Jiangsu Natural Science Foundation (BK20240015), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA28030102), the National Natural Scientific Foundation of China (32241037), the Innovation Program of Institute of Soil Science (ISSASIP2201), and the Youth Innovation Promotion Association of Chinese Academy of Sciences.

Author affiliations: [a]State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China; [b]University of the Chinese Academy of Sciences, Beijing 100049, China; [c]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China; [d]School of Artificial Intelligence, Nanjing University, Nanjing 210023, China; [e]Department of Soil Biology and Plant Nutrition, University of Kassel, Kassel 34117, Germany; [f]Institute of Earth Surface Dynamics, University of Lausanne, Lausanne CH-1015, Switzerland; [g]Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang 110016, China; [h]Key Lab of Conservation Tillage and Ecological Agriculture, Liaoning Province, Shenyang 110016, China; [i]Department of Agroecology, Aarhus University, Tjele 8830, Denmark; [j]Aarhus University Centre for Circular Bioeconomy, Aarhus University, Tjele 8830, Denmark; [k]Interdisciplinary Centre for Climate Change, Aarhus University, Roskilde 4000, Denmark; [l]Department of Chemistry and Biochemistry, Old Dominion University, Norfolk, VA 23529; [m]School of Biological Sciences, University of Oklahoma, Norman, OK 73069; and [n]Department of Environmental Systems Science, Institute of Integrative Biology, ETH Zurich 8092, Switzerland

Author contributions: Y.L., J. Zhou, T.W.C., Z.-H.Z., and J. Zhang designed research; H.H., R.G.J., X.Z., J.C., Y.S., H.N., J.D., J.M., and Y.L. performed research; C.Q., K.X., R.G.J., M.K., C.L., X.Z., W.H., R.-X.T., J. Zhou, T.W.C., Z.-H.Z., J. Zhang, and Y.L. contributed new reagents/analytic tools; H.H., C.Q., R.G.J., X.Z., J.C., Y.S., J.D., W.H., R.-X.T., and Y.L. analyzed data; and H.H., K.X., R.G.J., M.K., C.L., J.C., J.M., J. Zhou, T.W.C., Z.-H.Z., J. Zhang, and Y.L. wrote the paper.

1. H. Eswaran, E. Vandenberg, P. Reich, Organic carbon in soils of the world. *Soil Sci. Soc. Am. J.* **57**, 192–194 (1993).
2. E. A. Davidson, S. E. Trumbore, R. Amundson, Biogeochemistry – Soil warming and organic carbon content. *Nature* **408**, 789–790 (2000).
3. C. Liang, J. P. Schimel, J. D. Jastrow, The importance of anabolism in microbial control over soil carbon storage. *Nat. Microbiol.* **2**, 17105 (2017).
4. E. A. Davidson, I. A. Janssens, Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* **440**, 165–173 (2006).
5. J. Lehmann, M. Kleber, The contentious nature of soil organic matter. *Nature* **528**, 60–68 (2015).
6. P. C. de Ruiter, E. Morrien, Global soil map pinpoints key sites for conservation. *Nature* **610**, 634–635 (2022).
7. C. Liang, M. Kästner, R. G. Joergensen, Microbial necromass on the rise: The growing focus on its role in soil organic matter development. *Soil Biol. Biochem.* **150**, 108000 (2020).
8. A. Miltner, P. Bombach, B. Schmidt-Brücken, M. Kästner, SOM genesis: Microbial biomass as a significant source. *Biogeochemistry* **111**, 41–55 (2012).
9. M. W. I. Schmidt *et al.*, Persistence of soil organic matter as an ecosystem property. *Nature* **478**, 49–56 (2011).
10. C. Liang, W. Amelung, J. Lehmann, M. Kaestner, Quantitative assessment of microbial necromass contribution to soil organic matter. *Glob. Change Biol.* **25**, 3578–3590 (2019).
11. C. M. Kallenbach, S. D. Frey, A. S. Grandy, Direct evidence for microbial-derived soil organic matter formation and its ecophysiological controls. *Nat. Commun.* **7**, 13630 (2016).
12. W. Amelung, S. Brodowski, A. Sandhage-Hofmann, R. Bol, Combining biomarker with stable isotope analyses for assessing the transformation and turnover of soil organic matter. *Adv. Agron.* **100**, 155–250 (2008).
13. M. Kleber *et al.*, Mineral–organic associations: Formation, properties, and relevance in soil environments. *Adv. Agron.* **130**, 1–140 (2015).
14. J. K. Jansson, K. S. Hofmockel, Soil microbiomes and climate change. *Nat. Rev. Microbiol.* **18**, 35–46 (2020).
15. E. D. Whalen *et al.*, Clarifying the evidence for microbial- and plant-derived soil organic matter, and the path toward a more quantitative understanding. *Glob. Change Biol.* **28**, 7167–7185 (2022).
16. R. Benner, Biosequestration of carbon by heterotrophic microorganisms. *Nat. Rev. Microbiol.* **9**, 75 (2011).
17. T. Ma *et al.*, Divergent accumulation of microbial necromass and plant lignin components in grassland soils. *Nat. Commun.* **9**, 3480 (2018).
18. T. Camenzind, K. Mason-Jones, I. Mansour, M. C. Rillig, J. Lehmann, Formation of necromass-derived soil organic carbon determined by microbial death pathways. *Nat. Geosci.* **16**, 115–122 (2023).
19. G. Angst, K. E. Mueller, K. G. J. Nierop, M. J. Simpson, Plant- or microbial-derived? A review on the molecular composition of stabilized soil organic matter *Soil Biol. Biochem.* **156**, 108–189 (2021).
20. B. Wang, S. An, C. Liang, Y. Liu, Y. Kuzyakov, Microbial necromass as the source of soil organic carbon in global ecosystems. *Soil Biol. Biochem.* **162**, 108422 (2021).
21. R. G. Joergensen, Amino sugars as specific indices for fungal and bacterial residues in soil. *Biol. Fertil. Soils* **54**, 559–568 (2018).
22. G. Guggenberger, S. D. Frey, J. Six, K. Paustian, E. T. Elliott, Bacterial and fungal cell-wall residues in conventional and no-tillage agroecosystems. *Soil Sci. Soc. Am. J.* **63**, 1188–1198 (1999).
23. A. Appuhn, R. G. Joergensen, Microbial colonisation of roots as a function of plant species. *Soil Biol. Biochem.* **38**, 1040–1051 (2006).
24. K. S. Khan, R. Mack, X. Castillo, M. Kaiser, R. G. Joergensen, Microbial biomass, fungal and bacterial residues, and their relationships to the soil organic matter C/N/P/S ratios. *Geoderma* **271**, 115–123 (2016).
25. A. J. Simpson, M. J. Simpson, E. Smith, B. P. Kelleher, Microbially derived inputs to soil organic matter: Are current estimates too low? *Environ. Sci. Technol.* **41**, 8070–8076 (2007).
26. D. S. Jenkinson, "The determination of microbial biomass carbon and nitrogen in soil" in *Advances in nitrogen cycling in agricultural ecosystems*, J. R. Wilson, Ed. (CAB International, 1988).
27. B. Engelking, H. Flessa, R. G. Joergensen, Shifts in amino sugar and ergosterol contents after addition of sucrose and cellulose to soil. *Soil Biol. Biochem.* **39**, 2111–2118 (2007).
28. X. Zhang *et al.*, Local community assembly mechanisms shape soil bacterial beta diversity patterns along a latitudinal gradient. *Nat. Commun.* **11**, 1–10 (2020).
29. M. Delgado-Baquerizo *et al.*, A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018).
30. W. Amelung, Nitrogen biomarkers and their fate in soil. *J. Plant Nutr. Soil Sci.* **166**, 677–686 (2003).
31. W. Amelung, S. Brodowski, A. Sandhage-Hofmann, R. Bol, Combining biomarker with stable isotope analyses for assessing the transformation and turnover of soil organic matter. *Adv. Agron.* **100**, 155–250 (2008).
32. X. Wang *et al.*, Elevated temperature increases the accumulation of microbial necromass nitrogen in soil via increasing microbial turnover. *Glob. Change Biol.* **26**, 5277–5289 (2020).
33. F. J. W. Parmentier *et al.*, The impact of lower sea-ice extent on Arctic greenhouse-gas exchange. *Nat. Clim. Change* **3**, 195–202 (2013).
34. R. G. Joergensen, Phospholipid fatty acids in soil-drawbacks and future prospects. *Biol. Fertil. Soils* **58**, 1–6 (2022).
35. V. Borisov *et al.*, Deep neural networks and tabular data: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **35**, 7499 (2022).
36. Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms (Chapman and Hall/CRC, 2012).
37. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
38. T. Chen, C. Guestrin, "Xgboost: A scalable tree boosting system" in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, R. Rastogi, Eds. (Association for Computing Machinery, New York, NY, 2016), pp. 785–794.
39. G. Ke *et al.*, Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inform. Process. Syst.* **30**, 3149–3157 (2017).
40. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inform. Process. Syst.* **31**, 6639–6649 (2018).
41. Z.-H. Zhou, J. Feng, Deep forest. *Natl. Sci. Rev.* **6**, 74–86 (2019).
42. M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, F. Hutter, Auto-sklearn 2.0: Hands-free automl via meta-learning. *J. Mach. Learn. Res.* **23**, 1–61 (2022).
43. L. Du *et al.*, "TabularNet: A neural Network architecture for understanding semantic structures of tabular data" in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, F. Zhu, B. C. Ooi, C. Miao, Eds. (Association for Computing Machinery, New York, NY, 2021), pp. 322–331.
44. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, L. Agapito, T. Berg, J. Kosecka, L. Zelnik-Mano, Eds. (IEEE Computer Society, Washington, DC, 2016), pp. 770–778.
45. Y. Gorishniy, I. Rubachev, V. Khrulkov, A. Babenko, Revisiting deep learning models for tabular data. *Adv. Neural Inform. Process. Syst.* **34**, 18932–18943 (2021).
46. R. Garnett, Bayesian Optimization (Cambridge University Press, 2023).
47. D. Silver *et al.*, Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
48. J. Kirschner, M. Mutny, N. Hiller, R. Ischebeck, A. Krause, "Adaptive and safe bayesian optimization in high dimensions via one-dimensional subspaces" in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri, R. Salakhutdinov, Eds. (Proceedings of Machine Learning Research, Cambridge, MA, 2019), pp. 3429–3438.
49. R. Turner *et al.*, "Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020" in *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, H. J. Escalante, K. Hofmann, Eds. (Proceedings of Machine Learning Research, Cambridge, MA, 2021), pp. 3–26.
50. S. M. Lundberg, S. I. Lee, "A Unified Approach to Interpreting Model Predictions" in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett, Eds. (Curran Associates, Red Hook, NY, 2017), pp. 4765–4774.
51. G. Patoine *et al.*, Drivers and trends of global soil microbial carbon over two decades. *Nat. Commun.* **13**, 4195 (2022).