Molecular tradeoffs in soil organic carbon composition at continental scale

Steven J. Hall[1*†], Chenglong Ye[1,2*], Samantha R. Weintraub[3], William C. Hockaday[4†]

[1]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011, USA

[2]Ecosystem Ecology Lab, College of Resources and Environmental Sciences, Nanjing Agricultural University, Nanjing 210095, China

[3]National Ecological Observatory Network, Battelle, Boulder, Colorado, 80301

[4]Department of Geosciences, Baylor University, Waco, TX 76798, USA

*Contributed equally to this manuscript

[†]Co-corresponding authors: stevenjh@iastate.edu; Phone: 515-294-7650; Fax: 515-294-1337

william_hockaday@baylor.edu; Phone: 254-710-2639; Fax: 254-710-2673

The molecular composition of soil organic carbon remains contentious. Microbial-, plant-, and fire-derived compounds may each contribute, but do they vary predictably among ecosystems? Here we present carbon functional groups and molecules from a diverse spectrum of North American surface mineral soils, primarily collected from the National Ecological Observatory Network, quantified by nuclear magnetic resonance spectroscopy and a molecular mixing model. Soils varied widely in relative contributions of carbohydrate, lipid, protein, lignin, and char-like carbon, but each compound class had similar overall abundance. Three principal component axes explained 90% of the variance in carbon composition: the first showed a tradeoff between lignin and protein, the second showed a tradeoff between carbohydrate and char, and the third was explained by lipids. Reactive aluminum, crystalline iron oxides, and pH plus overlying organic horizon thickness best explained variation along each respective axis; these predictors were ultimately related to climate. Together, our data point to continental-scale tradeoffs in soil carbon molecular composition which are linked to environmental and geochemical variables known to predict carbon mass concentrations. Controversies regarding the genesis of soil carbon and its potential responses to global change can be partially reconciled by considering diverse ecosystem properties that drive complementary persistence mechanisms.

Soil organic carbon (SOC) is generally understood to comprise a diverse suite of biomolecules representing the decomposition products of plant and microbial biomass and the imprint of abiotic processes such as fire[1,2]. However, the fundamental mechanisms controlling the molecular composition of SOC within and among mineral soils remain contentious[3]. Do disparate soils converge along a predictable molecular continuum of SOC composition driven by the inexorable transformation of plant detritus to a consistent suite of low-molecular-weight decomposition products[1,4–6]? Or conversely, do diverse biogeochemical factors such as climate, vegetation, or mineralogy lead to distinct molecular differences in SOC among ecosystems[7,8]? We can increasingly predict the spatial distribution of SOC as a function of biogeochemical properties[9], as well as the partitioning of SOC between particulate and mineral-associated pools[10]. An equivalent framework for predicting SOC molecular composition among ecosystems remains elusive but could inform our understanding of the functional properties of SOC and its dynamics under global change[3,11].

Debates on the importance of different mechanisms of SOC persistence rest in part on our contested understanding of its molecular composition. Plant-derived aromatic compounds like lignin were historically thought to dominate SOC due to their macromolecular structure[12], which requires strong oxidants for depolymerization[13]. Subsequent work challenged this view by demonstrating that aromatic and lignin-like moieties may be minor constituents[5] that decompose faster than bulk SOC[14,15]. Microbial necromass and low-molecular-weight decomposition products (carbohydrates, proteins, and lipids) have assumed key roles in current SOC paradigms given the potential for efficient microbial metabolism and recycling of these molecules. In this view, SOC persistence does not derive from chemical complexity[16] or stability but rather from protective physico-chemical interactions with minerals and aggregates with microbial detritus playing a dominant role[6,17–21].

However, the importance of microbial vs. direct plant contributions to SOC could vary among ecosystems[22]. Microbial growth and necromass production may be decoupled from SOC accumulation in stressful environments where decomposition is inefficient[23]. Significant contributions of lignin and other plant-derived compounds to mineral-associated SOC were recently observed[24–26]. In fact, lignin-derived C may have been systematically underestimated due to methodological biases[25,27,28]. Finally, char-like molecules presumably derived from pyrolytic decomposition are prevalent in many ecosystems[29,30] despite evidence that the increased stability of these molecules does not guarantee long-term persistence[31,32]. In spite of significant theoretical and empirical progress[3,10,33], we still lack a consistent framework for reconciling potential controls on SOC molecular composition across diverse ecosystem types. Such a framework would enhance our conceptual understanding of the origins and persistence of soil carbon to inform modeling and management of this critical resource.

Here, we leveraged a unique sample archive and datasets provided by the National Ecological Observatory Network (NEON), along with additional samples, to characterize SOC molecular composition and its relationships with biogeochemical factors across 42 North American surface mineral soils (Extended Data Fig. 1). Samples spanned 11 of the 12 US Department of Agriculture soil orders (all except Histosols, which were explicitly excluded) and the major ecosystem gradients of North America (tropics to tundra; Supplementary Table 1, Extended Data Fig. 2). We quantified the molecular functional groups of bulk SOC of demineralized samples using solid-state ^{13}C cross-polarization magic-angle-spinning (CPMAS) nuclear magnetic resonance (NMR) spectroscopy. We confirmed the robustness of results by assessing sample pretreatment and analytical biases in companion measurements employing cross- and direct-polarization NMR (Supplementary Discussion; Supplementary Tables 1–6, Supplementary Figures 1–5).

Molecular variation in SOC at the continental scale

The ^{13}C CPMAS NMR spectra (Extended Data Fig. 3) illustrated substantial variability in SOC molecular composition (Fig. 1a). Across all samples, O-alkyl, alkyl, and aromatic C were the largest constituents, with mean values of 23%, 22%, and 21% of total SOC, respectively; each varied as much as 2.5-, 3.4-, and 2.8-fold among samples (Fig. 1a). Amide/carboxyl C was less abundant (13%; $P < 0.05$) but was greater than phenolic, N-alkyl/methoxyl, and di-O-alkyl C, which each comprised 7% of SOC, on average (Fig. 1a). These results challenge a previous literature synthesis of similar measurements, where disparate soils tended towards a consistent ranking of C functional groups (O-alkyl > alkyl > aromatic > carbonyl)[5]. Across our diverse soils, no constituent was dominant overall, and either O-alkyl, alkyl, or aromatic C could predominate within an individual sample.

The molecular mixing model implied that five constituent molecules had similar relative abundance across the dataset as a whole: carbohydrate, lignin, lipid, protein, and char comprised mean values of 21%, 21%, 18%, 18%, and 17% of SOC, respectively (Fig. 1b). Despite their similar means, these molecules varied greatly among soils, by as much as 4-, 33-, 46-, 10-, and 6-fold, respectively. Carbonyl C which represented the oxidized products of various molecules was consistently less abundant (5% of SOC). To illustrate molecular changes during decomposition, we compared these SOC data with previous litter measurements conducted using comparable NMR methods (Supplementary Table 7). Although highly variable among ecosystems, litter is typically dominated by carbohydrates (36–80%; mean 56%), with lesser contributions from lignin (12–43%; mean 24%), lipids (0–21%; mean 8%) and proteins (0.3–28%; mean 7%). Litter data available from a subset of the forested NEON sites also had a mean lignin content of 24% (Extended Data Fig. 2). Therefore, on average, SOC from our surface mineral soils (Fig. 1b) tended to have less carbohydrate, similar lignin, and more lipid and protein as compared with typical organic matter inputs, even after accounting for carbohydrate losses during HF treatment (Supplementary Information). Our data are consistent with the emerging consensus that readily decomposable biomolecules, especially carbohydrates and proteins, are often important SOC components[33], and they reinforce the importance of lipids[1,4]. However, our data challenge the view that lignin is disproportionately lost relative to other molecules as microbes degrade litter to form SOC[5,14,15]. Although variable, mean lignin abundance in SOC (21%) was similar to the other dominant molecules and similar to mean lignin abundance in litter.

Some SOC molecules covaried with vegetation and management characteristics (Extended Data Fig. 4). Lignin was significantly greater (23% vs. 16%) and protein was a smaller component (15% vs. 23%) of SOC in forests than grasslands/shrublands ($P < 0.05$). Char was greater in ecosystems experiencing periodic prescribed fire (22% vs. 16%, $P < 0.05$), but intriguingly, char was not limited to fire-prone ecosystems. All soils contained measurable char (> 6%). Interpretation of char-like C remains contentious, as it might be produced by non-fire-related processes[34]. However, the fact that char significantly increased in soils with a known history of fire (Extended Data Fig. 4) indicates the importance of pyrogenesis. Five soils were from perhumid climates where mean annual precipitation (MAP) exceeded potential evapotranspiration (PET) by > 1 m (Supplementary Table 1). This implicates a plausible role for ancient or anthropogenic fire in producing extant char. For example, anthropogenic charcoal production, but not natural fire, was documented in the rainforests of the Luquillo Mountains, Puerto Rico[35], where two samples were collected.

Consistent molecular tradeoffs linked to ecosystem factors

Molecules covaried in predictable ways within samples despite high variability in composition among samples. Lignin and carbonyl C were positively correlated while lignin and protein, and carbohydrate and char, were each negatively correlated (corrected $P < 0.01$, $P < 0.0001$, and $P < 0.0001$, respectively; Extended Data Fig. 5). Principal components analysis showed that the correlation matrix of SOC molecules was dominantly explained ($R^2 = 0.90$) by three axes, which we rotated orthogonally to maximize interpretability and thus refer to as rotated components (Fig. 2, Supplementary Table 3). The first rotated component (RC1) scores were positively correlated with lignin (r = 0.9) and carbonyl (r = 0.8) and negatively correlated with protein (r = -0.8; Fig. 2a, Extended Data Fig. 6). The second rotated component (RC2) scores were positively correlated with carbohydrate (r = 0.88) and negatively correlated with char (r = -0.92). The third rotated component (RC3) scores were strongly correlated with lipid (r = -0.98) and weakly correlated with other molecules (r < 0.53). The molecular tradeoffs implied by the RC axes indicated the importance of multiple SOC persistence mechanisms enabling differential accrual of molecules among ecosystems.

To assess potential mechanisms underlying observed variation in SOC composition, we analyzed correlations between molecule relative abundance and biogeochemical predictors and performed multiple regression analyses and structural equation models (SEMs) for each RC axis. Several SOC molecules showed significant correlations with geochemical, biological, and climate variables (Fig. 3, Extended Data Fig. 7, Extended Data Fig. 8). Lignin and carbonyl C correlated positively with concentrations of oxalate-extractable aluminum (Al_o), which represents Al in short-range-ordered (SRO) mineral phases and/or organo-metal complexes that can protect SOC from microbial decomposition[36]. In contrast, protein had a negative correlation with Al_o and with copy numbers of the fungal internal transcribed spacer (ITS) region, and a positive correlation with pH. Lipid C correlated negatively with mean annual temperature (MAT), pH, and sulfate-extractable calcium and magnesium ($Ca_s + Mg_s$), which may participate in cation bridging with clay minerals. Lipid C correlated positively with the thickness of the overlying organic (O) horizon, which in turn had a strong negative relationship with MAT (Fig. 4).

Consistent with these pairwise correlations, different sets of variables best predicted variation along each RC axis (Extended Data Fig. 9). We present multiple models fit by backwards selection using more conservative ($P < 0.01$) and liberal ($P < 0.05$) variable selection criteria, respectively (Methods). We also compared models fit to the NEON samples only vs. the complete dataset, given that not all potential predictors were available for all samples (e.g., root and microbial data). Across all models, Al_o concentration was the best predictor of RC1 ($r = 0.63–0.69$, $P < 0.001$), with increasing values reflecting greater lignin vs. protein. The more liberal models indicated that $Ca_s + Mg_s$, forest vegetation, and prescribed fire were also positively correlated with RC1, as was ITS copy number. For RC2, crystalline iron mineral concentration (Fe_{d-o}) was a consistently important predictor across models ($r = 0.28–0.48$, $P < 0.01$), which was associated with increased carbohydrate vs. char. The liberal models also indicated a negative correlation of RC2 with mineral horizon thickness and fine root C:N, and a positive correlation with fine root biomass. For RC3, soil pH and O horizon thickness were the strongest predictors ($r = 0.40–0.60$, $P < 0.001$); MAP-PET and fine root C:N also correlated positively with RC3. The more acidic soils with thicker O horizons were associated with greater lipid relative abundance.

The SEMs showed that the strongest biogeochemical predictors of SOC composition were ultimately related to climate, either directly, or via proxies for soil development which were also related to climate (thickness of the O horizon and surface mineral genetic horizon; Fig. 4). Concentrations of Al_o increased with MAP-PET (excess moisture drives dissolution of Al-bearing minerals[37]) and decreased with MAT. Similarly, Fe_{d-o}, which accumulates as soils progressively weather, also increased with MAP-PET. Temperature impacted SOC composition both directly and indirectly. Increasing MAT decreased O horizon thickness, consistent with increased decomposition of unprotected organic matter with warmer temperature[38]. Organic horizon thickness was directly linked to RC3, and also indirectly linked via effects on pH (O horizons may promote acidification by leaching organic acids[39]). Thinner O horizons were also associated with thinner mineral surface genetic horizons in our dataset, possibly reflecting differences in soil profile development related to litter decomposition rates. Surface mineral horizon thickness, in turn, was proximately linked to SOC composition (RC2). Including vegetation type or fire did not improve any of the SEMs, possibly because these factors were adequately reflected by climate or soil-horizon-related variables.

The relationships between SOC composition and biogeochemical predictors observed here provide a molecular-level explanation for trends in SOC content among ecosystems noted elsewhere. The concentration of Al_o is among the best predictors of SOC content at local to global scales[9,40], reflecting its formation of protective complexes with SOC[36]. Our data imply that specific geochemical associations between lignin- and carbonyl-derived SOC and Al_o could explain increases in SOC content with Al_o among soils. The finding that lignin was the only molecule whose relative abundance significantly increased with SOC content (Fig. 3) also accords with this interpretation. A strong Al_o-lignin relationship is consistent with previous evidence of ligand exchange by carboxylated aromatics on SRO ordered Al phases[41] and high concentrations of lignin-derived C observed in humid tropical soils rich in SRO phases[25,40,42].

While lignin was greater in forested than non-forested soils (Extended Data Fig. 4), the relationship between lignin and Al_o—and ultimately, climate—was much stronger than the relationship with vegetation (Figs. 3,4, Extended Data Table 3). Relationships between Al_o and lignin partially reconcile aspects of old and new SOM paradigms: lignin-derived C may contribute significantly to SOC in some soils[12] (Fig. 1), but not because of inherent recalcitrance[31,32]. Rather, lignin (and carbonyl C, which was strongly correlated with lignin) may vary among ecosystems as a function of geochemical context. Our statistical models also supported a role for Ca and Mg in protecting lignin; these cations can provide protective bridging between anionic SOC functional groups and negatively charged mineral surfaces[9]. Minerals and metals are effective predictors of SOC content[31,36], and interactions with specific SOC molecules may underlie these patterns.

Complementary mechanisms of SOC persistence

The first observed tradeoff, between lignin and protein (Figs. 2,5), may reflect multiple underlying mechanisms. First, where SRO mineral phases (i.e., Al_o) and physicochemical protection are scarce, protein relative abundance may increase because it is a major microbial biomass component that can be efficiently recycled between living and dead microbes, whereas most lignin C is decomposed to carbon dioxide[18,21]. Second, in acidic soils, low abundance of protein (Fig. 3) vs. lignin may be driven by inefficient litter decomposition and low microbial necromass production[23]. Third, the negative correlation between fungal ITS copies and protein (Fig. 3) suggests that fungi may play a role in the lignin-protein tradeoff. Fungi are dominant decomposers of lignin[13] but have a higher biomass C:N (lower protein content) than bacteria[43]. Finally, the observed tradeoff between lignin and protein in SOC could reflect the fundamental role of the lignin:N ratio in controlling litter decay rates. Protein is a dominant soil N pool, and limited N availability to produce lignin-degrading enzymes could constrain lignin mass loss[44].

The second observed tradeoff, between carbohydrate and char (Figs. 2,4), was most closely linked to Fe_{d-o}. Crystalline Fe phases are protective sorbents that may promote soil aggregation[36,41,42], and these physicochemical protection mechanisms may explain increased relative abundance of carbohydrate, an easily decomposed molecule. Increasing quality (lower C:N ratios) and quantity of fine root biomass were also associated with greater carbohydrate. In contrast, char became more abundant as Fe_{d-o} and fine root quantity and quality decreased. We interpret this second tradeoff as follows: where physicochemical protection is lacking, a complex molecular structure involving a greater diversity of bond types and more stable bonds (such as those contained in polyaromatic char-like SOC) becomes increasingly important. Accepting that molecular structure alone cannot guarantee long-term persistence[32], the logistical challenges of char decomposition[16,29] may increase its relative contribution to SOC where other protection mechanisms are unavailable and root C inputs are small.

The third SOC axis was related most strongly to lipid content (Figs. 2,4). Temperature and pH are known to impact SOC content of mineral soils[9,38], and our data indicate that this may be influenced by lipid accrual (Figs. 3,4). Lipids were largely independent of other molecules and increased in cold, acidic soils with thick overlying organic horizons, comprising up to 59% of SOC. Constraints on microbial physiology may promote lipid persistence. Lipids are the most chemically reduced constituents of SOC, requiring greater activation energy for oxidation than other compounds[45]. Because decomposition reactions are temperature dependent, the Arrhenius equation predicts that molecules with the highest activation energies (i.e. lipids) exhibit the greatest increase in decomposition rate with increasing temperature if other protection mechanisms are unavailable[38]. As such, accrual of lipids in cold mineral soils is consistent with thermodynamic expectations.

Collectively, our continental-scale dataset supports a concise new framework for understanding multiple complementary mechanisms of SOC persistence among ecosystems (Fig. 5). Debates as to the relative importance of microbial necromass vs. lignin in SOC[20,21,25,28] can be reconciled in part by considering the biogeochemical heterogeneity of ecosystems: necromass may be more important than lignin where reactive Al phases are scarce, and vice-versa. Similarly, the contested role of molecular stability in SOC persistence[16,31,32] is also illuminated by examination of broad ecosystem gradients: where physicochemical protection mechanisms mediated by Fe are scarce and high-quality root C inputs are small, char assumes a more important role. Finally, differences in temperature and pH among ecosystems were ultimately linked to lipid abundance, informing debates as to which SOC forms may be most impacted by near-term warming and acidification[38,46]. Over longer timescales, temperature and moisture influence all three axes of this conceptual framework via soil development (Fig. 4). Collectively, our data point to the power of a macrosystems approach in reconciling paradigmatic controversies in SOC research.

References

1. Baldock, J. A., Masiello, C. A., Gélinas, Y. & Hedges, J. I. Cycling and composition of organic matter in terrestrial and marine ecosystems. *Mar. Chem.* **92**, 39–64 (2004).
2. Sutton, R. & Sposito, G. Molecular structure in soil humic substances: the new view. *Environ. Sci. Technol.* **39**, 9009–9015 (2005).

3. Lehmann, J. & Kleber, M. The contentious nature of soil organic matter. *Nature* **528**, 60–68 (2015).

4. Baldock, J. A. *et al.* Assessing the extent of decomposition of natural organic materials using solid-state ^{13}C NMR spectroscopy. *Aust. J. Soil Res.* **35**, 1061–1084 (1997).

5. Mahieu, N., Randall, E. W. & Powlson, D. S. Statistical analysis of published carbon-13 CPMAS NMR spectra of soil organic matter. *Soil Sci. Soc. Am. J.* **63**, 307 (1999).

6. Grandy, A. S. & Neff, J. C. Molecular C dynamics downstream: The biochemical decomposition sequence and its impact on soil organic matter structure and function. *Sci. Total Environ.* **404**, 297–307 (2008).

7. Baldock, J. A. *et al.* Aspects of the chemical structure of soil organic materials as revealed by solid-state ^{13}C NMR spectroscopy. *Biogeochemistry* **16**, 1–42 (1992).

8. Ahmad, R., Nelson, P. N. & Kookana, R. S. The molecular composition of soil organic matter as determined by ^{13}C NMR and elemental analyses and correlation with pesticide sorption. *Eur. J. Soil Sci.* **57**, 883–893 (2006).

9. Rasmussen, C. *et al.* Beyond clay: towards an improved set of variables for predicting soil organic matter content. *Biogeochemistry* **137**, 297–306 (2018).

10. Cotrufo, M. F., Ranalli, M. G., Haddix, M. L., Six, J. & Lugato, E. Soil carbon storage informed by particulate and mineral-associated organic matter. *Nat. Geosci.* **12**, 989–994 (2019).

11. Wagai, R. *et al.* Linking temperature sensitivity of soil organic matter decomposition to its molecular structure, accessibility, and microbial physiology. *Glob. Change Biol.* **19**, 1114–1125 (2013).

12. Waksman, S. A. & Iyer, K. R. N. Contribution to our knowledge of the chemical nature and origin of humus: I. On the synthesis of the “humus nucleus”. *Soil Sci.* **34**, 43–69 (1932).

13. Kirk, T. K. & Farrell, R. L. Enzymatic" combustion": the microbial degradation of lignin. *Annu. Rev. Microbiol.* **41**, 465–501 (1987).

14. Amelung, W., Brodowski, S., Sandhage-Hofmann, A. & Bol, R. Combining biomarker with stable isotope analyses for assessing the transformation and turnover of soil organic matter. *Adv. Agron.* **100**, 155–250 (2008).

15. Thevenot, M., Dignac, M.-F. & Rumpel, C. Fate of lignins in soils: A review. *Soil Biol. Biochem.* **42**, 1200–1211 (2010).

16. Bosatta, E. & Ågren, G. I. Soil organic matter quality interpreted thermodynamically. *Soil Biol. Biochem.* **31**, 1889–1891 (1999).

17. Miltner, A., Bombach, P., Schmidt-Brücken, B. & Kästner, M. SOM genesis: microbial biomass as a significant source. *Biogeochemistry* **111**, 41–55 (2011).

18. Cotrufo, M. F., Wallenstein, M. D., Boot, C. M., Denef, K. & Paul, E. The Microbial Efficiency-Matrix Stabilization (MEMS) framework integrates plant litter decomposition with soil organic matter stabilization: do labile plant inputs form stable soil organic matter? *Glob. Change Biol.* **19**, 988–995 (2013).

19. Kallenbach, C. M., Frey, S. D. & Grandy, A. S. Direct evidence for microbial-derived soil organic matter formation and its ecophysiological controls. *Nat. Commun.* **7**, 13630 (2016).

20. Ma, T. *et al.* Divergent accumulation of microbial necromass and plant lignin components in grassland soils. *Nat. Commun.* **9**, 1–9 (2018).

21. Liang, C., Amelung, W., Lehmann, J. & Kästner, M. Quantitative assessment of microbial necromass contribution to soil organic matter. *Glob. Change Biol.* **25**, 3578–3590 (2019).

22. Khan, K. S., Mack, R., Castillo, X., Kaiser, M. & Joergensen, R. G. Microbial biomass, fungal and bacterial residues, and their relationships to the soil organic matter C/N/P/S ratios. *Geoderma* **271**, 115–123 (2016).

23. Malik, A. A. *et al.* Land use driven change in soil pH affects microbial carbon cycling processes. *Nat. Commun.* **9**, 3591 (2018).

24. Córdova, S. C. *et al.* Plant litter quality affects the accumulation rate, composition, and stability of mineral-associated soil organic matter. *Soil Biol. Biochem.* **125**, 115–124 (2018).

25. Huang, W. *et al.* Enrichment of lignin-derived carbon in mineral-associated soil organic matter. *Environ. Sci. Technol.* **53**, 7522–7531 (2019).

26. Wan, D. *et al.* Iron oxides selectively stabilize plant-derived polysaccharides and aliphatic compounds in agricultural soils. *Eur. J. Soil Sci.* **70**, 1153–1163 (2019).

27. Hernes, P. J., Kaiser, K., Dyda, R. Y. & Cerli, C. Molecular trickery in soil organic matter: hidden lignin. *Environ. Sci. Technol.* **47**, 9077–9085 (2013).

28. Klotzbücher, T., Kalbitz, K., Cerli, C., Hernes, P. J. & Kaiser, K. Gone or just out of sight? The apparent disappearance of aromatic litter components in soils. *SOIL* **2**, 325–335 (2016).

29. Preston, C. M. & Schmidt, M. W. I. Black (pyrogenic) carbon: a synthesis of current knowledge and uncertainties with special consideration of boreal regions. *Biogeosciences* **3**, 397–420 (2006).

30. Lehmann, J. *et al.* Australian climate–carbon cycle feedback reduced by soil black carbon. *Nat. Geosci.* **1**, 832–835 (2008).

31. Mikutta, R., Kleber, M., Torn, M. S. & Jahn, R. Stabilization of soil organic matter: association with minerals or chemical recalcitrance? *Biogeochemistry* **77**, 25–56 (2006).

32. Kleber, M. What is recalcitrant soil organic matter? *Environ. Chem.* **7**, 320–332 (2010).

33. Schmidt, M. W. I. *et al.* Persistence of soil organic matter as an ecosystem property. *Nature* **478**, 49–56 (2011).

34. DiDonato, N., Chen, H., Waggoner, D. & Hatcher, P. G. Potential origin and formation for molecular components of humic acids in soils. *Geochim. Cosmochim. Acta* **178**, 210–222 (2016).

35. Scatena, F. An introduction to the physiography and history of the Bisley experimental watersheds in the Luquillo Mountains of Puerto Rico. *USDA For. Serv. Gen. Tech. Rep.* **SO-72**, 1–22 (1989).

36. Kleber, M. *et al.* Mineral–organic associations: formation, properties, and relevance in soil environments. *Adv. Agron.* **130**, 1–140 (2015).

37. Slessarev, E. W. *et al.* Water balance creates a threshold in soil pH at the global scale. *Nature* **540**, 567–569 (2016).

38. Davidson, E. A. & Janssens, I. A. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* **440**, 165–173 (2006).

39. Lundström, U. S., van Breemen, N. & Bain, D. The podzolization process. A review. *Geoderma* **94**, 91–107 (2000).

40. Kramer, M. G., Sanderman, J., Chadwick, O. A., Chorover, J. & Vitousek, P. M. Long-term carbon storage through retention of dissolved aromatic acids by reactive particles in soil. *Glob. Change Biol.* **18**, 2594–2605 (2012).

41. Kaiser, K. & Guggenberger, G. The role of DOM sorption to mineral surfaces in the preservation of organic matter in soils. *Org. Geochem.* **31**, 711–725 (2000).

42. Coward, E. K., Ohno, T. & Plante, A. F. Adsorption and molecular fractionation of dissolved organic matter on iron-bearing mineral matrices of varying crystallinity. *Environ. Sci. Technol.* **52**, 1036–1044 (2018).

43. Throckmorton, H. M., Bird, J. A., Dane, L., Firestone, M. K. & Horwath, W. R. The source of microbial C has little impact on soil organic matter stabilisation in forest ecosystems. *Ecol. Lett.* **15**, 1257–1265 (2012).

44. Moorhead, D. L. & Sinsabaugh, R. L. A theoretical model of litter decay and microbial interaction. *Ecol. Monogr.* **76**, 151–174 (2006).

45. LaRowe, D. E. & Van Cappellen, P. Degradation of natural organic matter: A thermodynamic analysis. *Geochim. Cosmochim. Acta* **75**, 2030–2042 (2011).

46. Ye, C. *et al.* Reconciling multiple impacts of nitrogen enrichment on soil carbon: plant, microbial and geochemical controls. *Ecol. Lett.* **21**, 1162–1173 (2018).

Correspondence and requests for materials should be addressed to S.J.H (stevenjh@iastate.edu) and W.C.H. (william_hockaday@baylor.edu)

Acknowledgements

We gratefully acknowledge the efforts of NEON and NRCS staff in conducting soil sampling and analyses, and E. Ayres for providing access to ‘Megapit’ samples. This research was supported in part by NSF projects DEB 1802745 and EAR 1132124. The National Ecological Observatory Network is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle Memorial Institute.

Author contributions

S.J.H., S.R.W., and W.C.H. developed the research concepts, C.Y. and W.C.H. conducted the NMR analyses, S.J.H., C.Y., S.R.W, and W.C.H. analyzed data, and S.J.H. and C.Y. wrote the paper with contributions from all authors.

Competing interests

The authors declare no competing interests.

Additional Information

Supplementary Information is available for this paper.

Figure Captions

Fig. 1: **Boxplots of carbon abundance as the fraction of total SOC in each sample.** Values were determined directly from ^{13}C CPMAS NMR peak areas (**a)** and by applying a molecular mixing model (**b)**. Grey dots represent observations (n = 42). Center lines are medians; box limits are upper and lower quartiles; whiskers are 1.5x the interquartile ranges; points are outliers.

Fig. 2: Rotated principal components analysis of SOC molecules. RC1, RC2, and RC3 represent rotated components 1–3, which respectively explained 35%, 29%, and 26% of the total variation (90% overall) in the correlation matrix of SOC molecule relative abundance. Grey dots represent soil samples, and labeled green arrows indicate correlations between SOC molecules and RCs, with the correlation coefficient indicated on the top and right axes (carbohyd denotes

carbohydrate). Several samples with RC values > 3 (Supplementary Table 3) are not shown for clarity.

Fig. 3: **Heatmap of correlations (r) between SOC molecules and biogeochemical predictors**. The symbols *, **, and **** denote corrected significance at $P < 0.05$, $P < 0.01$, and $P < 0.0001$, respectively. MAT, mean annual temperature; MAP-PET, mean annual precipitation minus potential evapotranspiration. Additional descriptions of biogeochemical predictors and any data transformations are provided in the Methods and Supplementary Table 5.

Figure 4: Parsimonious structural equation models of SOC molecular composition. The response variables (RC1, RC2, RC3) are the rotated principal components of SOC molecule relative abundance. Solid yellow and blue lines indicate significant positive or negative piecewise relationships between variables ($P < 0.05$). Dashed lines indicate non-significant piecewise relationships that improved overall model fit as indicated by comparing AIC of nested models. Numbers in boxes are scaled correlation coefficients. Fisher's C statistic refers to the test of the overall model fit, where high P values indicate plausibility of the overall model.

Figure 5: Conceptual model of three-dimensional tradeoffs in SOC composition linked to complementary persistence mechanisms as supported by our data. Soil samples fall within a spherical space indicating the relative predominance of different SOC molecules, which are constrained according to three major axes of variation. The location of a sample along each axis indicates the relative importance of different SOC persistence mechanisms as described in the text.

Methods

Soil sampling and analysis. We analyzed the molecular SOC composition of surface mineral soil samples spanning 32 sites in the NEON Megapit archive, along with 10 additional soils which were selected to encompass additional diversity in biogeochemical characteristics (Supplementary Table 1). Vegetation included forests (n = 29) and grasslands or open canopy shrublands (n = 13), including both managed (burned or grazed) and wildland sites. Soils were sampled from the upper-most mineral soil horizon at a given site (organic horizons were excluded); complete soil profile descriptions for the NEON sites are provided in Supplementary Table 2. Briefly, in the dominant soil and vegetation type at each NEON terrestrial site, a soil profile was characterized and sampled by horizon with the help of US Department of Agriculture Natural Resource Conservation (NRCS) staff and archived by NEON[47–49]. We requested subsamples of A horizon material from each site in the Megapit archive that was available in September 2019. The Gellisols had extensive organic (O) horizons, such that we requested material from the mineral horizon closest to the surface (described as Bg/Oajj, A/Cjj, and Bg at BONA, HEAL, and TOOL, respectively; Supplementary Table 1). The 10 non-NEON samples analyzed here were each collected from 0–10 cm depth with a clean shovel after removing any litter or O horizon material. All soils were air dried to constant mass and sieved to 2 mm. Visible root fragments were removed with tweezers and soils were finely ground with a mortar and pestle prior to subsequent analyses.

^{13}C CPMAS NMR analyses and sample preparation All 42 samples were prepared for NMR analyses, allowing a comparative characterization of organic C molecular composition. In order

to increase the NMR sensitivity and remove paramagnetic materials, soils were pre-treated with hydrochloric acid (HCl, 10% wt.) and hydrofluoric acid (HF, 10%, wt.) to remove any calcium carbonate and mineral phases, respectively[50]. Briefly, 2–3 g of finely ground soil was weighed into a 50 mL sealed polyethylene centrifugation tube, saturated with 30 mL HCl, and allowed to settle for 30 min. After centrifugation and discarding HCl, the remaining slurry was then shaken with 40 ml of mixed HF (10% wt.) and HCl (10% wt.) for 8 h, and subsequently centrifuged. The supernatant was removed and discarded appropriately. After repeating the procedure four times, each sample was washed with distilled water three times and dried at 50 ^{o}C under a stream of dinitrogen gas.

Solid-state ^{13}C CP-MAS and ^{13}C DP-MAS NMR spectra were recorded at room temperature (23 °C) using a 300 MHz Bruker AVANCE III NMR spectrometer equipped with a 4 mm magic angle spinning (MAS) probe (Bruker BioSpin, Billerica, MA) at Baylor University (Waco, TX). The 60–130 mg HF-treated sample was placed in a zirconium rotor with a diameter of 4 mm and Kel-F caps to maximize the C mass and signal intensity. A MAS rate of 12 kHz was used for all NMR measurements. Cross polarization (CP) experiments used a ramped-amplitude (50% to 100%) contact pulse and rotor synchronized Hahn echo[51]. The contact time and recycle delay were set to 2 ms and 1.2 s, respectively, and composite pulse proton decoupling was applied during signal acquisition. Direct polarization (DP) ^{13}C spectra were acquired with a 90-degree excitation pulse and rotor-synchronized Hahn echo[52], with a recycle delay of 180 s. Glycine was used as an external standard for setting pulse angles, chemical shift and Hartman-Hahn matching conditions. DPMAS spectra were obtained for 11 HF-treated soil samples as a means against which to assess relative quantitation bias in CPMAS NMR data[53]. These samples were selected to span a broad range of biogeochemical diversity (nine soil orders;

Supplementary Table 1) and contained sufficient SOC (> 2.9% C in the original samples) to enable timely analysis by DPMAS.

CPMAS spectra for HF-treated samples were acquired with more than 6000 scans. To assess potential impacts of HF treatment on SOC composition, 11 untreated samples with relatively high SOC concentration (> 6%) were also selected for NMR analysis (this set differed slightly from the CPMAS/DPMAS comparison given the differing selection criteria). These samples included six soil orders and spanned a broad range of paramagnetic element content (14–58 mg Fe g^{-1}). Spectra for these untreated samples were recorded using the same operation conditions of HF-treated samples and were acquired with more than 44000 scans. After baseline correction, quantification was performed by dividing the spectra into seven chemical shift regions: 0–45 ppm, 45–60 ppm, 60–95 ppm, 95–110 ppm, 110–145 ppm, 145–165 ppm and 165–215 ppm, assigned to alkyl C, N-alkyl + methoxyl C, O-alkyl C, Di-O-alkyl C, aromatic C, phenolic C, amide + carbonyl C, respectively. Subsequently, a molecular mixing model was applied to the seven integrated spectra regions, to estimate the relative abundances of six molecular SOC constituents (carbohydrate, protein, lignin, lipid, carbonyl and char)[1]. The elemental concentrations of C and N were measured on the HF-treated samples by combustion/elemental analysis at Baylor University (Costech 4010, Valencia, CA) and were used as additional constraints on the molecular mixing model solutions[1].

Biogeochemical analyses Megapit soil samples and vegetation in proximity to the soil pit were subjected to numerous physical and chemical analyses[54]. Here, we utilized measurements of total elemental content and particle size from the Megapit samples. We also used measurements of the copy number of bacteria/archaea (16S) and fungi (ITS) coding regions calculated by quantitative polymerase chain reaction (qPCR), which were conducted on separate fresh soil samples collected in the vicinity of each sampled Megapit profile[54]. Briefly, these soils were flash frozen in the field on dry ice and shipped to an analytical facility for DNA extraction and amplification. Soil samples for qPCR analysis were collected periodically (approximately three times per year) from each site from 0–30 cm depth, and cores were visually separated according to organic and mineral horizons; only samples from mineral soil were used here. We selected samples from plots in proximity to each Megapit (i.e., within several hundred m; denoted as Tower plots in NEON terminology) and averaged the mean 16S and ITS abundance for each site based on the 2016–2018 data. Fine root biomass was measured by depth in three pit profiles within the Megapit and sorted into live/dead classes for fine (< 2 mm or < 4 mm, depending on the site) and coarse diameter classes. Here, we denoted the combined < 2 mm and < 4 mm fractions as fine roots for subsequent analyses. Roots were dried, weighed, and combusted for analysis of carbon (C) and nitrogen (N) content. We averaged root data from 0–30 cm depth for use in subsequent analyses. No NEON root data were available from TOOL, so we used previous published data from the same site[55]. Samples for foliar and/or litter chemistry were available from a subset of the NEON sites (15 and 16 sites, respectively), as these are collected from each site on a five-year rolling schedule. Foliar samples represented clips of bulk herbaceous samples from the plant community. Litter samples included debris from trees and shrubs. We used measurements of foliar and litter C:N and a proxy for lignin content (acid-unhydrolyzable residue)[54]. No root or microbial or litter chemistry data were available from the non-NEON samples from which we collected ^{13}C NMR spectra.

We conducted several additional soil extractions of all samples to quantify reactive metals. Subsamples were extracted in parallel with sodium dithionite (1:150 ratio of soil:solution) to quantify pedogenic iron (denoted Fe_d) and ammonium oxalate (1:60 ratio of soil:solution) to quantify Fe and aluminum in short-range-ordered phases and organo-metal complexes (termed Fe_o and Al_o). The concentration of crystalline Fe minerals was then calculated as the difference between Fe_d and Fe_o (Fe_{d-o}). Subsamples were also sequentially extracted with deionized water and sodium sulfate (1:150 ratio of soil:solution). The calcium and magnesium concentration of the sodium sulfate extraction (termed $Ca_s + Mg_s$), which followed the water extraction, was interpreted as a proxy for Ca and Mg that may have participated in divalent cation bridging between clays and organic matter[46]. All metals were analyzed by inductively coupled plasma optical emission spectroscopy at Iowa State University (ICP-OES; Perkin Elmer Optima 5300 DV, Waltham Massachusetts). Mean annual precipitation and temperature data were estimated for each NEON site using previously synthesized data[56]. Potential evapotranspiration (PET) data were extracted from a global 1-km resolution mean annual evapotranspiration dataset from 2000-2014[57].

Statistical analyses Correlation heatmaps were calculated between SOC molecules and biogeochemical predictors, and some variables were log10-transformed because of skewness (Fe_o, Al_o, Fe_{d-o}, $Ca_s + Mg_s$, ITS, 16S). Significance of correlations was calculated by multiplying *P* values according to a Bonferroni correction to correct for multiple comparisons and an α of 0.10. We used a rotated principal components analysis to assess relationships among the relative abundances of the six SOC molecules calculated from the molecular mixing model and soil biogeochemical variables. Principal components were calculated from the correlation matrix of the C molecule data and rotated orthogonally (varimax rotation) using the "Psych" package[58] in R version 3.6.0. Rotation is commonly used in PCA to simplify interpretation of principal components by maximizing/minimizing the correlations between factors and component axes. These rotated components (RC) of the correlation matrix were used to facilitate interpretation of each component in terms of dominant C molecule(s). To investigate relationships among RCs and biogeochemical variables, we fit multiple linear regression models for each RC using the lm function in R. The global model contained the following potential explanatory variables: mean annual temperature, mean annual precipitation, mean annual precipitation minus potential evapotranspiration, forest vs. non-forest vegetation, presence/absence of recurring fire, Al_o, Fe_o, $Fe_{c\text{-}o}$, $Ca_s + Mg_s$, fine root biomass, fine root C:N ratio, the ratio of total base cations to zirconium (a weathering ratio sensu[59]), and copy numbers of 16S and ITS genes quantified by qPCR. Most of the non-forested ecosystems were grazed, such that a separate variable for grazing was not included. Certain predictor variables were only available for the NEON samples (n = 32; Supplementary Table 3), such that model selection was conducted independently for both datasets. Candidate models for each RC were carefully investigated for multicollinearity of predictors and assumptions of normality and heteroscedasticity by calculating variance inflation factors (VIF) and graphically examining plots of residuals. Prior to model selection, individual predictors with VIF > 3 were sequentially deleted[60], the reduced global model was refit, and VIF values were calculated again. After removing collinear predictors, we performed model selection by backwards elimination; more conservative and more liberal models yielded by $\alpha = 0.01$ and $\alpha = 0.05$, respectively, were presented for completeness. Following multiple linear regression, to better understand interrelationships among proximate predictors of RC axes and soil forming factors, we fit SEMs using the piecewiseSEM package v. 2.1.0 in R[61]. Candidate SEMs included the direct biogeochemical predictors identified by multiple linear regressions, along with climate, soil, and vegetation variables that might influence those biogeochemical predictors. The optimum models for each RC were selected by comparing AIC values among nested models.

Data availability

Summarized NMR data are available in the Supplementary Information, and raw NMR spectra data and sample biogeochemical characteristics are available in the Environmental Data Initiative digital repository: https://doi.org/10.6073/pasta/2284825ecb8460f056ae5b0e7d355cc8

Code availability

R scripts used for post-processing data are available in the Environmental Data Initiative digital repository: https://doi.org/10.6073/pasta/2284825ecb8460f056ae5b0e7d355cc8

References

47. Ayres, E., *et al.* NEON Field and Lab Procedure and Protocol: TIS Soil Pit Sampling Protocol. NEON.DOC.001307. https://data.neonscience.org/data-products/DP1.00097.001 (2017).

48. Ayres E., & Durden, D. NEON Field and Lab Procedure and Protocol: TIS Soil Archiving. NEON.DOC.000325. https://data.neonscience.org/data-products/DP1.00097.001 (2017).

49. Ayres, E. NEON procedure and protocol: producing TIS soil archive subsamples for users. NEON.DOC.001306. https://data.neonscience.org/data-products/DP1.00097.001 (2017).

50. Gélinas, Y., Baldock, J. A. & Hedges, J. I. Demineralization of marine and freshwater sediments for CP/MAS ^{13}C NMR analysis. *Org. Geochem.* **32**, 677–693 (2001).

51. Harbison, G. S. *et al.* High-resolution carbon-13 NMR of retinal derivatives in the solid state. *J. Am. Chem. Soc.* **107**, 4809–4816 (1985).

52. Mao, J.-D. *et al.* Quantitative characterization of humic substances by solid-state carbon-13 nuclear magnetic resonance. *Soil Sci. Soc. Am. J.* **64**, 873–884 (2000).

53. Longbottom, T. L. & Hockaday, W. C. Molecular and isotopic composition of modern soils derived from kerogen-rich bedrock and implications for the global C cycle. *Biogeochemistry* **143**, 239–255 (2019).

54. NEON (National Ecological Observatory Network). DP1.00096.001, DP1.10066.001, DP1.10102.001, DP1.10109.001 (accessed September 1, 2019), DP1.10026.001, DP1.10033.001, DP1.10031.001 (accessed May 15, 2020). http://data.neonscience.org.

55. Sullivan, P. F. *et al.* Climate and species affect fine root production with long-term fertilization in acidic tussock tundra near Toolik Lake, Alaska. *Oecologia* **153**, 643–652 (2007).

56. SanClements, M. *et al.* Collaborating with NEON. *BioScience* **70**, 107–107 (2020).

57. Mu, Q., Zhao, M. & Running, S. W. Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sens. Environ.* **115**, 1781–1800 (2011).

58. Revelle, W. psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.12 (2018).

59. Chittleborough, D. J. Indices of weathering for soils and palaeosols formed on silicate rocks. *Aust. J. Earth Sci.* **38**, 115–120 (1991).

60. Hair, J. F., Risher, J. J., Sarstedt, M. & Ringle, C. M. When to use and how to report the results of PLS-SEM. *Eur. Bus. Rev.* **31**, 2–24 (2019).

61. Lefcheck, J. S. piecewiseSEM: Piecewise structural equation modelling in R for ecology, evolution, and systematics. *Methods Ecol. Evol.* **7**, 573–579 (2016).