UC Riverside

UC Riverside Previously Published Works

Title

ENVnet provides a global molecular resource of dissolved organic matter

Permalink

https://escholarship.org/uc/item/71k8c9hg

Journal

Nature Biotechnology

ISSN

1087-0156

Authors

Bowen, Benjamin P
Harwood, Thomas V
de Raad, Markus
et al.

Publication Date

2026-07-21

DOI

10.1038/s41587-026-03230-0

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Accepted Manuscript

ENVnet provides a global molecular resource of dissolved organic matter

Accepted for publication in Nature Biotechnology.

This manuscript is the author-approved accepted version of the article, after peer review and before publisher copyediting, typesetting, and proof correction.

The DOI number is: 10.1038/s41587-026-03230-0

Submitted: November 5th, 2024
Accepted: June 19th, 2026
Published online: July 21st, 2026, https://www.nature.com/articles/s41587-026-03230-0

Authors

Benjamin P. Bowen[1,2,*], Thomas V. Harwood[2], Markus de Raad[1], Katherine B. Louie[2], Suzanne M. Kosina[1], Katherine D. McMahon[3,4,5], Neslihan Taş[1,6], Nicholas J. Bouskill[6,7], Terry C. Hazen[8,9], Shellie R. Bench[10], Rachel Mackelprang[11], Daniel Petras[12], Mingxun Wang[13], Fernando T. Maestre[14], Stephen J. Giovannoni[2,15], Trent R. Northen[1,2,*]

* address correspondence to Ben Bowen (bpbowen@lbl.gov) and Trent Northen (trnorthen@lbl.gov)

1. Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Lab, Berkeley, CA, 94720
2. The Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
3. Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, Wisconsin, USA
4. Wisconsin Energy Institute, University of Wisconsin-Madison, Madison, Wisconsin, USA
5. Department of Bacteriology, University of Wisconsin-Madison, Madison, Wisconsin, USA
6. Climate and Ecosystems Division, Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, California, USA, 94720
7. Department of Biological and Ecological Engineering, Oregon State University, Corvallis, Oregon, USA
8. University of Tennessee, Knoxville, TN, 37916
9. Oak Ridge National Laboratory, Oak Ridge, TN 37831
10. Global Viral - Now LRC systems, 1336 N Carolan Ave Suite 206, Burlingame, CA 94010
11. Biology Department, California State University Northridge, 18111 Nordhoff St., Northridge, CA 91330, USA
12. Department of Biochemistry, University of California, Riverside, 69 Aberdeen Dr, Riverside, CA 92507
13. Department of Computer Science and Engineering, University of California Riverside, 900 University Ave. Riverside CA 92521
14. Environmental Science and Engineering, Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Kingdom of Saudi Arabia
15. Department of Microbiology, Oregon State University, Corvallis, Oregon, USA

Abstract

Dissolved organic matter (DOM) is a central component of Earth's carbon cycle and one of the planet's most chemically diverse pools, yet the molecular structures of its constituents remain largely unresolved. This limitation has hindered our ability to link DOM composition to microbial processes and ecosystem function. Here, we present ENVnet, a global molecular repository built from tandem mass spectrometry data collected across 13 terrestrial and aquatic environment types, including 419 newly generated samples that expand publicly available DOM metabolomics data and cover previously underrepresented environments. By computationally deconvolving chimeric mass spectra, a longstanding challenge in environmental metabolomics, we recover high-quality fragmentation data for >22,000 distinct molecular features (defined by a specific precursor mass and fragmentation pattern). Using ENVnet, we uncover conserved and environment-specific molecular patterns in DOM composition and underlying biogeochemical processes. We also use molecular features encoded in ENVnet to train predictive models of DOM persistence, allowing molecular-level assessment of microbial turnover in independent systems.

Main

There is intense interest in improving our understanding of the global carbon cycle[1,2]. While metagenomic sequencing technologies have produced high resolution genomic maps of the microbial communities that drive the carbon cycle[3], the chemical composition of terrestrial and marine dissolved organic matter (DOM) remains largely inaccessible at the molecular level[4]. While techniques such as FT-ICR-MS and 2D NMR have illuminated DOM's bulk formula composition and specific atomic bonding environments such as quaternary carbons and functional group distributions[5,6], compound-level identification and cross-environment structural comparisons at scale remain largely unexplored[7–9]. This barrier has impeded comparison of molecular features across ecosystems and understanding of Earth's microbial communities.

Metabolomic technologies, which are well suited for global profiling of complex organic mixtures, have potential to address this gap[10]. However, their use requires high-quality tandem mass spectrometry (MS/MS) data for the identification of chemical structures and formulas[10–12]. This is why spectral libraries of authentic standards have played such a critical role in biomedical applications of metabolomics[13]. However molecular resources analogous to spectral libraries are currently unavailable for a wide range of environments. This is mainly due to the large number of molecules with similar chromatographic properties and masses, such that individual MS/MS spectra often contain fragments from multiple structurally distinct molecules—so-called *chimeric spectra*[14] —because several compounds are frequently isolated and fragmented at the same time. Mass difference matching (MDM) was introduced to extract structural information from chimeric spectra by analyzing characteristic fragment mass differences associated with known precursor ions[9,15]. However, this approach does not facilitate the reconstruction of individual, molecule-specific MS/MS spectra from mixed fragmentation events. Consequently, it is unable to generate a molecular resource or a curated spectral reference library of DOM molecules that would enable structural identification and cross-environment comparison like those employed in biomedical metabolomics. While techniques such as FT-ICR-MS and 2D NMR provide complementary structural information, a curated tandem mass spectrometry reference library would uniquely enable large-scale molecular networking and structural comparisons across thousands of environmental samples needed to elucidate and potentially manage biological and chemical mechanisms of DOM cycling[16,17].

Here, we introduce ENVnet, a global molecular resource designed to resolve and compare the molecular composition of dissolved organic matter across diverse environments using tandem mass spectrometry. We built ENVnet using an extensive dataset of tandem mass spectra generated from 1,389 organic-matter metabolomics samples: 419 newly produced and 970 publicly available. Using a custom deconvolution algorithm and a specialized molecular-networking workflow, ENVnet is purposefully designed to overcome the challenges inherent to DOM tandem mass spectra. These datasets represent a wide array of environments, including global croplands, drylands, late Pleistocene permafrost, temperate and tropical forest soils, rivers,

lakes, subsurface groundwater, and coastal seawater DOM. Because solid samples (for example, soils) were analyzed as solvent extracts, they may include compounds that are not freely dissolved in situ; for clarity, we nonetheless refer to all samples collectively as DOM throughout the manuscript. ENVnet was constructed using custom, high-performance computing implementations of chimeric spectra deconvolution[9] and molecular networking[18]. Using ENVnet workflows in the GNPS2 Analysis Hub, a web-based mass spectrometry analysis platform[12,19], we explore molecularly resolved data on organic matter composition from these environments and investigate the effects of various treatments. To show how ENVnet enables predictive insight, we used soil DOM incubation experiments to train a machine-learning model that assesses the microbial lability of individual DOM molecules with high accuracy. These findings illustrate how ENVnet can be used by researchers from a wide range of disciplines to explore hypotheses regarding DOM dynamics, enhance existing models of carbon cycling and storage through integrated, molecularly resolved analyses.

Results

Selection of diverse global environmental sample datasets for network construction

To develop ENVnet as a durable and static molecular resource that spans a variety of biomes, we integrated appropriate and publicly available datasets with additional new datasets (SI Table 1A). This collection of spectra and the resulting network as a version-controlled molecular resource can be updated as new datasets become available, informing additional details about ENVnet nodes and attributes (particularly for unknown metabolites). By doing so, the nodes and regions of ENVnet become identifiable, and can be used to relate changes between experiments using a common identifier. By disseminating ENVnet as a freely available, open-source molecular resource, the complex calculations described below can be leveraged and applied by researchers to explore a wide range of DOM-focused research questions.

ENVnet was constructed from a diverse collection of 1,389 samples spanning four major categories: freshwater (n=221), soil (n=443), seawater (n=698), and International Humic Substances Society (IHSS) reference materials (n=27), as summarized in Figure 1a and SI Table 1a. An ontology for environments (ENVO) was used for specifying the environment classification for each sample[20]. We focus exclusively on negative ionization mode data, as it is the one in which the MDM neutral losses were described[9]. While adding positive mode data would likely increase the number of compounds detected, limiting this analysis to negative ionization mode data allowed us to focus on the highly oxidized molecules most relevant to the study of DOM with extended residence times and resistance to rapid microbial turnover (referred to as “persistent” throughout for simplicity).

This compilation of negative mode tandem liquid chromatography mass spectrometry (LC-MS/MS) data from public repositories (MassIVE)[12] was performed using a combination of programmatic data cleaning and manual data curation[21]. Selection of MassIVE datasets (SI Table 1C) was accomplished by manual curation of project metadata based on both the description of the dataset and data quality. Datasets were manually screened for relevance based on sample metadata (terrestrial or aquatic DOM samples) and analysis type (metabolomics, excluding proteomics or lipidomics). We further removed files containing added xenobiotics, tissues, or unnatural treatments, leaving only metabolomics data from environmental samples. A February 2024 search of MassIVE (SI Table 1C) using keywords “DOM”, “dissolved organic matter”, “soil”, and “water” yielded numerous datasets, but the overwhelming majority lacked negative ionization mode data and were unsuitable for our adaptation of the MDM approach. Duplicate files were removed, and we ensured that the majority of signals from these files were calibrated within 0.001 Da by matching the measured precursor *m/z* to the theoretical *m/z* from chemical formulae derived *in silico* (MS-BUDDY)[11]. The resulting files can be seen in SI Table 1A.

To complement publicly available data (where soil and freshwater representation was limited and sometimes inconsistent in processing and analysis) we generated new DOM LC-MS/MS datasets from underrepresented

environments. These new datasets encompass a diverse range of environments, including dryland soils, agricultural lands, river systems, peatlands, permafrost, groundwater, and tropical forests, thereby enabling a more comprehensive analysis of DOM composition and dynamics across global ecosystems. These newly generated samples were consistently processed and analyzed within a single laboratory to ensure uniform data quality and comparability. Although interlaboratory and sample preparation variability remains inherent to the broader ENVnet dataset, integrating these consistently processed samples improves overall cross-dataset comparability and broadens the environmental representation of public DOM data. Further details on sample collection and processing are available in Methods and SI Table 1A. The new data we generated substantially enlarged the ENVnet database and provides a more comprehensive analysis of DOM across global environments.

Refining fragmentation data to maximize molecular classification

The resulting database included a total of 5.6 million negative mode MS/MS spectra (Fig. 1a & SI tables 1a and 1b). A primary obstacle hindering the effectiveness of conventional metabolomics analysis for complex environmental samples lies in the low abundance, sheer magnitude, and diversity of the molecules present in organic matter. As a result of this diversity, and the current technical limitations of mass filters used during precursor ion selection, much of the fragmentation data collected from these samples are chimeric (i.e., contain fragments from multiple parent ions).

While several approaches exist for spectral deconvolution, many require specialized data acquisition strategies, limiting their use on diverse public datasets[22–26]. To address the challenge of pervasive spectral chimerism in our dataset, we developed and applied a mass-based separation algorithm to computationally separate these complex chimeric data into a set of distinct, pure MS/MS spectra. Conceptually, our *ab initio* approach asks which fragment ions are most likely to originate from the same parent molecule, and then computationally separates each chimeric MS/MS spectrum into a set of purified, molecule-specific spectra without prior knowledge of the true precursor identities. Details of the algorithm are provided in the Supplementary Information (Appendix S1); briefly, the method generates a cloud of hypothetical precursor masses by adding known neutral losses (from Simon et al[9]) to each fragment ion and then applies a statistical clustering procedure to jointly identify co-isolated precursors and partition their associated fragment ions. The output is a set of distinct, computationally purified MS/MS spectra. Applying this algorithm to the 5.6 million spectra served as a primary quality filter; only MS/MS events with sufficient fragment evidence to form a robust cluster (≥3 ions) yielded a high-confidence output, resulting in approximately 4.7 million pure spectra. To generate a non-redundant molecular resource, we performed a final consolidation step in which spectra were clustered based on inferred precursor m/z (within 0.002 Da) and spectral similarity (cosine similarity > 0.9), retaining only the highest-quality spectrum from each cluster. This process yielded the final set of 22,128 unique spectra that represent the molecular nodes of ENVnet.

To systematically validate the performance of the deconvolution algorithm, we constructed 50 realistic, synthetic chimeric spectra by merging MS/MS spectra from high-quality reference library standards around a nominal precursor *m/z* and then applied the algorithm to these mixtures (Fig. 2c-j). At the fragment level, the algorithm achieves high precision and recall (Fig. 2c,d), indicating that most fragment ions are correctly assigned to their true parent molecules, with few incorrect or missing assignments. The precursor-level precision achieved is near-perfect (score consistently above 0.95, Fig. 2e), confirming that the algorithm's statistical clustering approach does not invent spurious precursors. The broader distribution observed in precursor recall (Fig. 2f) reflects a deliberately conservative strategy that minimizes false discovery, occasionally merging two very closely related precursors rather than incorrectly splitting a single one. This quantitative performance is visualized in representative deconvolution examples (Fig. 2g-j), where the algorithm successfully resolves complex mixtures of multiple precursors. The algorithm-inferred precursor masses (red dashed lines) align accurately with the ground-truth precursor-fragment relationships (blue circles), correctly partitioning the fragment ions (black dots) to their true parents. This successful transformation of a single, mixed spectrum into multiple, clean spectra provides the foundational data quality required for accurate downstream metabolite identification and is the basis for the ENVnet molecular resource.

To further ensure these algorithmically-inferred precursors represented real molecules and not computational artifacts, we performed a critical validation step. We cross-referenced each newly inferred precursor *m/z* against MS1 scans acquired at nearby retention time to confirm the presence of the inferred parent ions (within +/- 1 minute and 10ppm). This validation was successful for 93% of the spectra; the remaining 7% were discarded as likely artifacts, ensuring the high quality of the final ENVnet molecular resource.

The original spectra, deconvoluted spectra, and deconvoluted precursor *m/z* values were searched against a custom library of reference tandem mass spectra. This library comprised a deduplicated union of all publicly available spectra from the GNPS2 library, along with private spectra from the NIST20 and NIST14 databases. Library searches were performed using cosine scoring (original vs original and deconvoluted vs deconvoluted)[12,27], retaining only high-quality matches (cosine score > 0.7, matching ions > 3, and precursor m/z difference < 0.002 Da). For each spectrum, the higher-scoring match—whether from the original or deconvoluted version—was selected. We assigned annotations using a tiered confidence hierarchy based on Metabolomics Standards Initiative (MSI) levels. This framework serves to explicitly communicate the varying degrees of structural ambiguity across the dataset.[28] Notably, reflecting the limited availability of authentic standards for environmental molecules, only 782 spectra (4.5%) produced library matches, yielding putative identifications (MSI Level 2). To provide deeper structural information for the vast number of remaining unannotated features, we implemented a comprehensive *in silico* workflow. First, we used MS-BUDDY[11], an *in silico* tool to predict the chemical formula of each node (MSI Level 4) (see SI Appendix S1)[11]. Second, the SIRIUS platform[29] was used to propose a tentative structure (MSI Level 3) via CSI:FingerID[30] and to predict a chemical class via CANOPUS[31] for every spectrum (SI Tables 2a & 2b). While the confidence in any single annotation is necessarily limited, this multi-level annotation of the entire 22,000-node network provides the robust chemical context needed to uncover large-scale ecological patterns. Notably, the distribution of compound classes, as shown in Figure 1d, reveals a prevalence of structures containing rings (aliphatic and aromatic) and oxygenated functional groups, with many featuring multiple stereocenters. This is consistent with the top compound classes predicted *in silico* (Fig. 1e), which include a range of oxygenated molecules from simple phenolics to more complex terpenoids[5]. Molecules with these characteristics have been previously identified as key constituents of persistent DOM[5,6,15]. Carbon count, H/C ratio, and O/C ratio of annotated formula match our expectation of what has been described previously from FT-ICR analysis of soil organic matter (Extended Data Fig. 1). The resulting formula assignments confirmed the highly oxygenated nature of the dataset (96.7% of nodes), and revealed substantial representation of nitrogen- (31.6%), sulfur- (19.0%), and phosphorus-containing (6.7%) molecules known to be critical for nutrient cycling (SI Fig. 1).

We next leveraged molecular networking to infer chemical relationships between tandem mass spectra[12,18]. Standard similarity algorithms, like modified cosine, struggle with the specialized nature of DOM spectra because it is composed of many molecules with similar MS/MS values (SI Fig. 2). We therefore developed Reaction Modification - Blur and Link (REM-BLINK), a spectral similarity algorithm designed to recognize chemically meaningful relationships among closely related DOM molecules. Unlike modified cosine, which considers unshifted and a single mass shift, REM-BLINK calculates a vector of over 100 similarity scores for each spectral pair, corresponding to a curated list of common DOM-related chemical differences from Simon et al[9]. This high-dimensional vector is then input into a machine learning model to produce a single similarity score that can capture more complex structural relationships (see SI Appendix S2 for additional details).

To validate this approach, we benchmarked REM-BLINK against modified cosine. Specifically, we calculated precision (the fraction of predicted similar pairs that are truly similar) and recall (the fraction of all actual similar pairs correctly identified) across a range of score thresholds using a held-out test set of tandem mass spectral pairs. Using the resulting area under the precision-recall curve (AUC-PR) as the primary performance metric, these analyses show that REM-BLINK substantially outperforms modified cosine for DOM-derived deconvoluted spectra (Fig. 3a). Furthermore, when used to build networks, REM-BLINK connects nodes with much narrower distributions of chemical properties (e.g., carbon count, H/C ratio) than other methods (SI Fig. 2), indicating it produces a more chemically coherent network (Fig. 3b-d). The resulting molecular network (ENVnet) consisted of 14,870 spectra/nodes and 389,656 edges in one interconnected subgraph; the

remaining nodes consisted of 5,752 singletons and several small subgraphs. As expected, 86% of all ENVnet nodes were originally from spectra that contained two or more co-isolated precursors and 71% of those had mass differences exceeding 10 ppm from the intended isolation target (Fig. 2ab)[14]. These results suggest that for complex DOM samples analyzed by traditional metabolomic-workflows, co-isolation may be widespread and warrants careful consideration.

High degree of DOM conservation across environments

ENVnet contains 8,874 spectra common to multiple environments and 2,088 that are present in all environments (Fig. 1a & SI table 3). To evaluate the molecular conservation of DOM on a global scale, we compared the ENVnet node peak area profiles of all 1,389 samples. Principal component analysis (PCA) revealed a distinct clustering of samples based on their environment of origin, illustrating both shared and unique molecular signatures across different biomes (Fig. 4a). While this separation is driven in part by the presence or absence of unique molecular signatures, these quantitative results should be interpreted with caution, as signal intensities vary across instruments and laboratories and may reflect technical as well as biological differences. Nevertheless, the PCA provides a valuable overview of the dataset, highlighting the dominant patterns of molecular similarity and divergence across global ecosystems. These observations are consistent with incomplete degradation of common metabolites through microbial activity such that each environment retains some unique molecular structures[5,6,15,32].

To resolve the specific chemical drivers of these patterns, we examined the distribution of chemical classes across environments (Fig. 4b). This analysis revealed a high degree of conservation; a core set of ubiquitous compound classes, including simple phenolic acids, amino acids, isocoumarins, and dicarboxylic acids, was found to be abundant across disparate ecosystems from tropical soils to permafrost. In contrast, other chemical classes exhibited strong environment-specificity. Notably, various terpenoid classes showed distinct distribution patterns, suggesting that while the core oxidative products of DOM are conserved, specific biogenic precursors and their derivatives drive the chemical divergence observed between environments[3]. However, considering common challenges in the analysis of complex mixtures, some ENVnet nodes are presumably unresolved chimeric MS/MS, in-source degradation products of larger molecules, or partially chromatographically resolved complex mixtures of isomers [8,9,33]. Addressing these analytical challenges will require a multi-pronged approach, combining advancements in instrumentation, data processing algorithms, and the development of comprehensive spectral libraries, ultimately leading to a more complete understanding of DOM composition and dynamics.

Using ENVnet for comparing, visualizing, and integrating DOM analysis

We incorporated ENVnet and the associated spectral library into an open source GNPS2 workflow (Fig. 5a, SI Fig. 3, 4, 5), a widely used, free online tool for metabolomic analysis, to make ENVnet accessible to diverse scientific communities[12]. This workflow allows scientists to map their MS/MS as well as MS1 (e.g. from Fourier transform ion cyclotron resonance mass spectrometry) onto ENVnet to provide structural information, visualization, and intersample comparisons. To further enhance accessibility, we have provided an online video tutorial guiding users through workflow execution and output interpretation (see Data Availability). We used this workflow to assess the identification of significant differences for two common DOM experiment designs: comparing samples from different environments and comparing treated to untreated samples. In the first case, we performed a soil incubation experiment and used ENVnet to train a model of DOM persistence to microbial transformations. In the second, we used ENVnet to compare DOM composition across a permafrost chronosequence. Together, these examples illustrate how ENVnet enables molecular-level insights that support predictive models, while also providing a detailed view of DOM transformations across environments[34,35].

Our soil incubation experiment was performed using DOM extracted from organic agricultural soil, soil microbes extracted from the same soil, and essential inorganic nutrients (Wolf's vitamins and minerals)[36]. The

incubations were supplemented with different amounts of labile carbon to simulate the 'priming effect'[37]. After an incubation of 7 days, the DOM was extracted and analyzed by LC-MS/MS. Nodes in ENVnet were classified as either depleted, enriched, or unchanged. Depleted nodes had a statistically significant ($p<0.05$) negative log2 fold-change (less than -0.5) between the incubation time points, whereas enriched nodes had a significant ($p<0.05$) positive log2 fold-change (greater than 0.5). Nodes exhibiting a change in relative abundance between timepoints, as determined by a two one-sided t-tests (TOST) procedure, between -0.5 and 0.5, were classified as unchanged.

A total of 2,101 DOM compounds were mapped to ENVnet by tandem mass spectra (>3 matching ions and >0.6 cosine score) and 18,064 by parent *m/z* with a mean peak area greater than 1e7 in one or more conditions. Microbial growth was detected in all microbial incubations along with noticeable decreases in depleted compounds, especially with labile carbon additions (Fig. 6 and SI Fig. 6). Terpenoids were more likely to remain unchanged (median 92.4% across conditions) compared 86.0% across all compound classes, whereas low molecular weight aromatic classes such as phenolic acids, lignans, phenylpropanoids, and coumarins were predominantly depleted or enriched (median 32.1% combined) compared to 14.8% across all displayed compound classes, with the direction of change dependent on soil type and labile carbon addition.

To predict which DOM molecules persist (e.g. 'unchanged' across all incubation studies) during microbial processing, we trained a gradient-boosted decision tree classifier using molecular features derived from both tandem mass spectra and formula characteristics associated with each node (SI Appendix S4)[38]. The resulting model demonstrated strong performance across five-fold cross-validation, yielding a mean accuracy of 0.91 and a Matthews Correlation Coefficient of 0.80 (SI Fig. 7). This high predictive capability suggests that specific chemical properties encoded within ENVnet are key drivers of microbial and abiotic lability[39].

To gain additional chemical insights into specific molecular drivers of persistence we applied SHAP (SHapley Additive exPlanations)[40]. Oxygen-to-carbon (O/C) ratio was the most dominant predictor, exhibiting a largely linear, inverse relationship with stability. As expected, the Nominal Oxidation State of Carbon (NOSC) was found to decrease the likelihood of a compound remaining unchanged (Extended Data Fig. 2) [41]. Together, these indicate that preferential microbial turnover of oxidized substrates, whereas reduced, aliphatic structures (high H/C) were more likely to persist, consistent with literature[41,42]. A high aromaticity index (AI) negatively impacted stability, indicating active turnover of aromatic moieties in these soils rather than preservation. This aligns with previous reports that aromatic compounds and lignin-derived carbon exhibit lower persistence in soil and are more susceptible to microbial decomposition compared to more stable, non-protonated aromatic carbons that are not bonded to oxygen [43,44]. SHAP analysis reveals consistent clustering of features associated with direct impact on stability (e.g., O/C) (Extended Data Fig. 2b) [45]. Conversely, features where high values are distributed across both positive and negative SHAP values imply a context-dependent relationship, where the feature's influence on stability is determined by interactions with other attributes in the feature vector. While elemental ratios like O/C, H/C, and AI showed generally linear impacts, other features such as the carbon atom count, sulfur content, and specific neutral losses (e.g., CO_2, $2xCO_2$) displayed inconsistent or context-dependent behaviors, implying that for complex heteroatomic structures, influences of specific structural configurations can outweigh influences of stoichiometry. Together, these results indicate that oxidation state strongly governs DOM persistence, while finer structural features modulate stability in a context-dependent manner. Notably, the agreement between these ENVnet-derived predictors and drivers of lability identified in prior FT-ICR-MS incubation studies[7,46,47] independently validates the chemical accuracy of ENVnet's molecular assignments, and demonstrates that this type of analysis is now accessible from widely available LC-MS/MS data.

To demonstrate how ENVnet enables models trained in one system to be applied more broadly, we extended our DOM stability model to diverse ecosystems, revealing new chemical insights into DOM transformations. This revealed distinct relationships between total intensity and predicted persistence. For instance, aquatic environments such as humic lakes and coastal seawater displayed some of the highest median node peak areas, yet possessed the lowest proportion of total intensity derived from predicted stable nodes (Extended Data Fig. 2c, d). This suggests that while specific compounds may be highly abundant in these aquatic

systems, the model identifies them as chemically susceptible to biological turnover rather than intrinsically stable. Conversely, tropical forest soils exhibited a higher fraction of total intensity (50.6%) derived from predicted stable nodes compared to the mean of all samples (30.6%) (Extended Data Fig. 2c), as would be expected from the high levels of activity in these soils[48,49]. In a related application, we used ENVnet to examine the effects of 27 different *P. trichocarpa* (poplar) genotypes on rhizosphere DOM LC-MS/MS analysis from a published study[50]. From this we see variation in persistent classes of DOM metabolites associated with known Poplar lines (See SI Appendix S5 and SI Fig. 8). While DOM turnover can strongly depend on local microbial and physicochemical factors, these predictions reveal variability in soil-stable signatures across environmental pools and demonstrate ENVnet's ability to compare emergent molecular properties at scale. Despite these promising results, future work is needed to greatly expand the model's training data which is currently limited to a single field site.

We next evaluated how permafrost age influences DOM molecular composition[12]. Samples used for this assessment were obtained from late Pleistocene permafrost that formed approximately 19,000 and 27,000 years before present (19 kya and 27 kya, respectively) in Alaska[51]. The environmental conditions in interior Alaska during deposition of the 27 kya permafrost samples were likely cooler and drier than deposition conditions of the 19 kya samples[52,53]. These climatic differences, which correspond to differences in paleovegetation and time to permafrost formation, shape the OM "starting material" and the amount of degradation that can occur prior to entrainment in permafrost[51,54,55]. Slow rates of carbon turnover by microbial processes through geologic time subsequently alter the characteristics of permafrost carbon[54,56].

We found 2,268 nodes in ENVnet that matched permafrost tandem mass spectra (>3 matching ions and >0.5 cosine score) and 13,470 nodes with MS1 peak-areas greater than 1e7. A variety of in silico compound classes increased with permafrost age (SI Fig. 9), such as cyclic compounds with multiple carboxyl groups. Surprisingly, the relative abundance of matches to authentic standards was higher in 19 kya permafrost (Fig. 3d) compared with 27 kya permafrost. Conversely many of the abundant features of 27 kya permafrost were both more oxidized and had higher molecular weights, and they were structural neighbors to authentic standards rather than exact matches (Extended Data Fig. 1 and Fig. 5c).

Discussion

Here we introduce ENVnet, a molecular resource for DOM analysis that integrates existing and newly developed algorithms to process over five million tandem mass spectra from samples spanning diverse environments. ENVnet is implemented within the freely accessible, web-based GNPS2 analysis hub, enabling users to compare DOM composition across environments and conditions, and easily visualize and compare metabolite structures, substructures, and compound classes. Designed to complement established DOM characterization approaches such as FT-ICR-MS and 2D NMR spectroscopy, ENVnet provides a fragmentation-based structural reference that uniquely enables compound-level identification and the contextualization of new data within a molecular network spanning thousands of samples. Furthermore, by reducing errors caused by chimeric and redundant spectra, ENVnet improves DOM compositional comparisons, facilitating hypothesis generation and supporting the development of mechanism-targeted approaches for DOM management, including detailed examination of conserved DOM to link recalcitrant organics to their sources and identify approaches to alter their abundances.

Our findings provide chemical insights into a key question in DOM cycling: is persistent DOM simply leftover material that microbes could not process, or is it composed of new compounds actively formed over time? Our observations support the latter, suggesting that recalcitrant DOM compounds are derivatives of necromass, actively produced through *in situ* transformations (biotic and abiotic) including microbial metabolism[54,56,57]. This mechanism is plausible in permafrost, where studies show some microbes maintain low activity and slowly transform organic matter, even at subzero temperatures, over geologic time[58,59]. ENVnet results align with the continuous transformation model: not just exact chemical matches, but also numerous 'structural neighbors',

chemically related molecules that differ by small modifications such as additional oxidation steps, were observed (see Fig. 5d, which shows a proliferation of partially oxidized analogs in aging permafrost). We acknowledge that the use of a curated list of DOM-related chemical differences in REM-BLINK may preferentially detect transformations well-represented in this list, and uncommon chemical modifications may be underrepresented. The unique value of ENVnet is that, by making processes visible in networks as diverse clusters of related compounds, it can link nuanced compositional changes in complex DOM to transformation mechanisms[7,60,61]. By providing a generalized approach for reducing high complexity DOM to patterns based on chemically informative relationships, ENVnet establishes a standardized framework for researchers to systematically analyze DOM transformations, facilitating the future identification of the diverse biochemical mechanisms that drive carbon cycling. Predictive models of DOM molecular stability are relevant in carbon valorization, agricultural optimization, land management and other applications. ENVnet highlights underrepresented environments and identifies where additional sampling could reveal novel DOM signatures. Because many DOM molecules remain undetected with current technologies, ENVnet also provides a benchmark dataset for evaluating emerging technologies that increase metabolite detection[28,62]. All current annotations have been rigorously classified according to Metabolomics Standards Initiative confidence levels, primarily molecular formulas (Level 4), tentative in silico structures (Level 3), and putative spectral-library identifications (Level 2)[28,62]. The scientific strength of ENVnet lies not in the certainty of any single molecular annotation but in the robust, reproducible patterns that emerge across thousands of consistently classified features and can be interpreted by ecologists, modelers, and biotechnologists.

As a living resource, we anticipate that continued community contributions of new data and standards through our GNPS2 and GitHub workflows will steadily elevate these annotations toward higher-confidence putative (Level 2) and ultimately validated (Level 1) identifications, further expanding ENVnet's utility as foundational molecular resource for understanding DOM dynamics and drivers and for developing technologies for carbon management across terrestrial and aquatic ecosystems.

Acknowledgments

We thank Thomas A. Douglas for permafrost sampling and expertise and Michael Snyder for help working with permafrost samples. F. T. M. acknowledges support from King Abdullah University of Science and Technology. The U.S. Government retains, and the publisher, by accepting the article for publication, acknowledges, that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.

Funding Statement

The construction of ENVnet was supported by ENIGMA- Ecosystems and Networks Integrated with Genes and Molecular Assemblies (http://enigma.lbl.gov) and m-CAFEs - Microbial Community Analysis and Functional Evaluation in Soils a Science Focus Area Programs at Lawrence Berkeley National Laboratory under contract DE-AC02-05CH11231 with the U.S. Department of Energy; the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility (https://ror.org/04xm1d337) including Community Science Projects to McMahon (10.46936/10.25585/60000947), Bench (10.46936/10.25585/60000046), Tas (10.46936/10.25585/60001146) and Bouskill (10.46936/10.25585/60001233); and the National Energy Research Scientific Computing Center (NERSC) a DOE Office of Science User Facility all operated under Contract No. DE-AC02-05CH11231 with the U.S. Department of Energy. The soil incubation experiments were supported by RESTOR-C - Center for Restoration of Soil Carbon by Precision Biological Strategies under Contract No. DE-AC02-05CH11231 with the U.S. Department of Energy. The BIODESERT global dryland survey was supported by the European Research Council (ERC grant agreement 647038). D.P. was supported by the Simons Foundation International through a Simons Early Career Investigator in Aquatic Microbial Ecology and Evolution Award (SFI-LS-ECIAMEE-00013858). F.T.M. was supported by King Abdullah University of Science and Technology. Sample collection from Wisconsin lakes was supported by the National

Science Foundation (North Temperate Lakes Long Term Ecological Research site DEB-2025982, DEB-1440297, and DEB-0822700). R.M. was supported by the National Science Foundation Division of Environmental Biology and Office of Polar Programs, grant number DEB 2029573.

Author Contributions

B.P.B. and T.V.H conception, design, analysis, interpretation, new software used in this work, draft and revision of the manuscript; M.d.R., K.B.L., and S.M.K, data and sample acquisition, K.D.M., N.T., N.T., N.J.B., T.C.H., S.R.B., R.M., and D.P. data contributions, M.W. new software used in this work, F.T.M. and S.J.G. data contribution and substantial revisions; T.R.N. conception, design, analysis, interpretation, draft and revision of the manuscript.

Competing Interest Statement

T.R.N. is a founder of two non-profits, Prosper Soils and Bioaligned Labs with prior approval from LBNL. M.W. is a co-founder of Ometa Labs LLC. The remaining authors declare no competing interests.

Figures & Captions

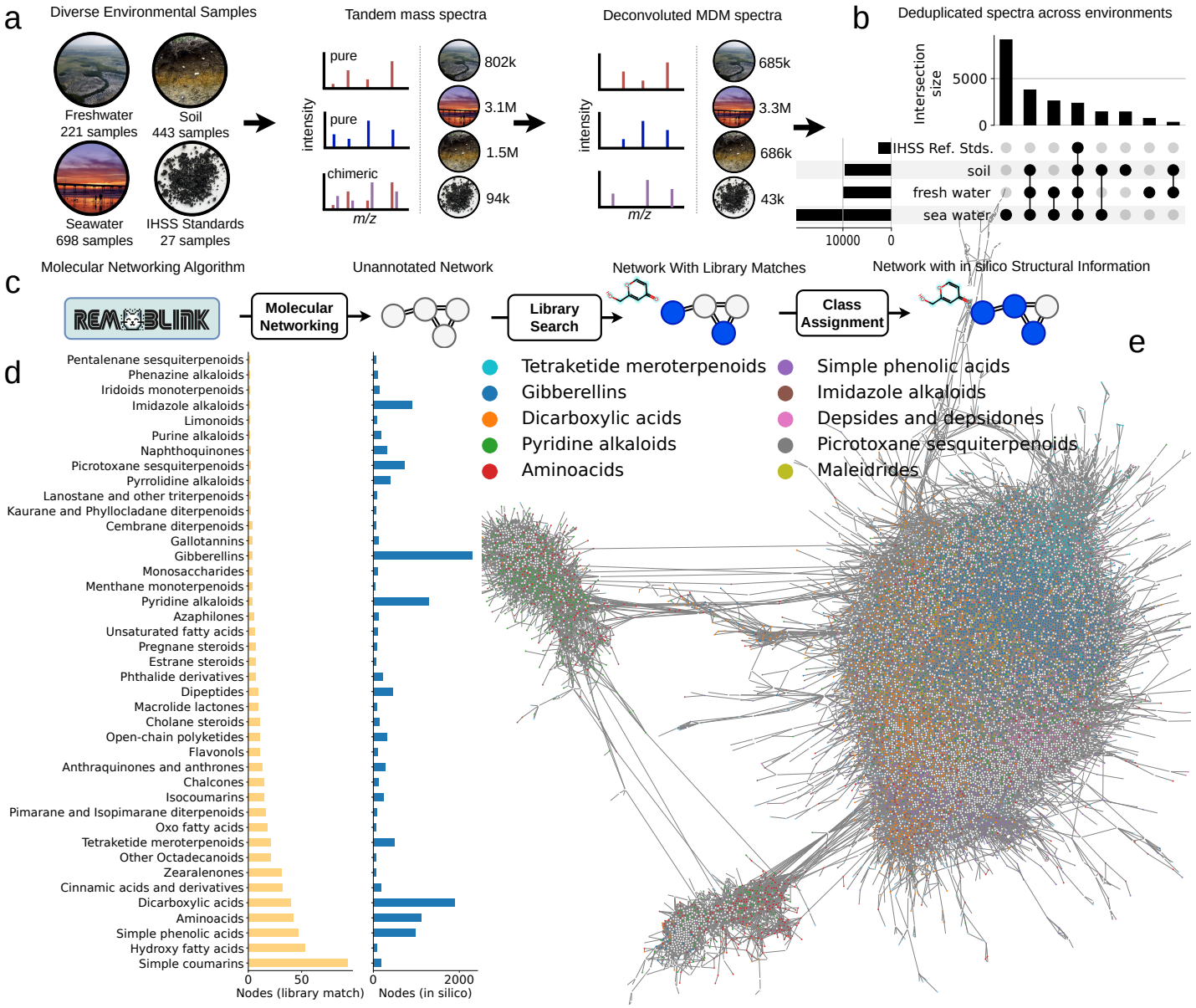


Figure 1. Overview of the methodology and results from the analysis of metabolomics data collected from diverse environmental samples. a) Tandem mass spectra from geographically diverse soil and water samples that were deconvoluted and clustered. b) These unique molecular signals reveal the distinct and shared molecules across many ecosystems with cosine scores greater than 0.5 and 3 or more matching ions to ENVnet. c) Using Reaction Modification - Blur and Link (REM-BLINK), spectra were linked based on spectral similarity, generating a molecular network which could be used to group related molecules and to relate identity matches to unknown molecules. d) Node frequency distribution of compound classes from identity matches as well as *in silico* predictions (left) provides a comprehensive structural reference for investigation. e) The completed ENVnet (right) shows the top 10 most common *in silico* predicted compound classes.

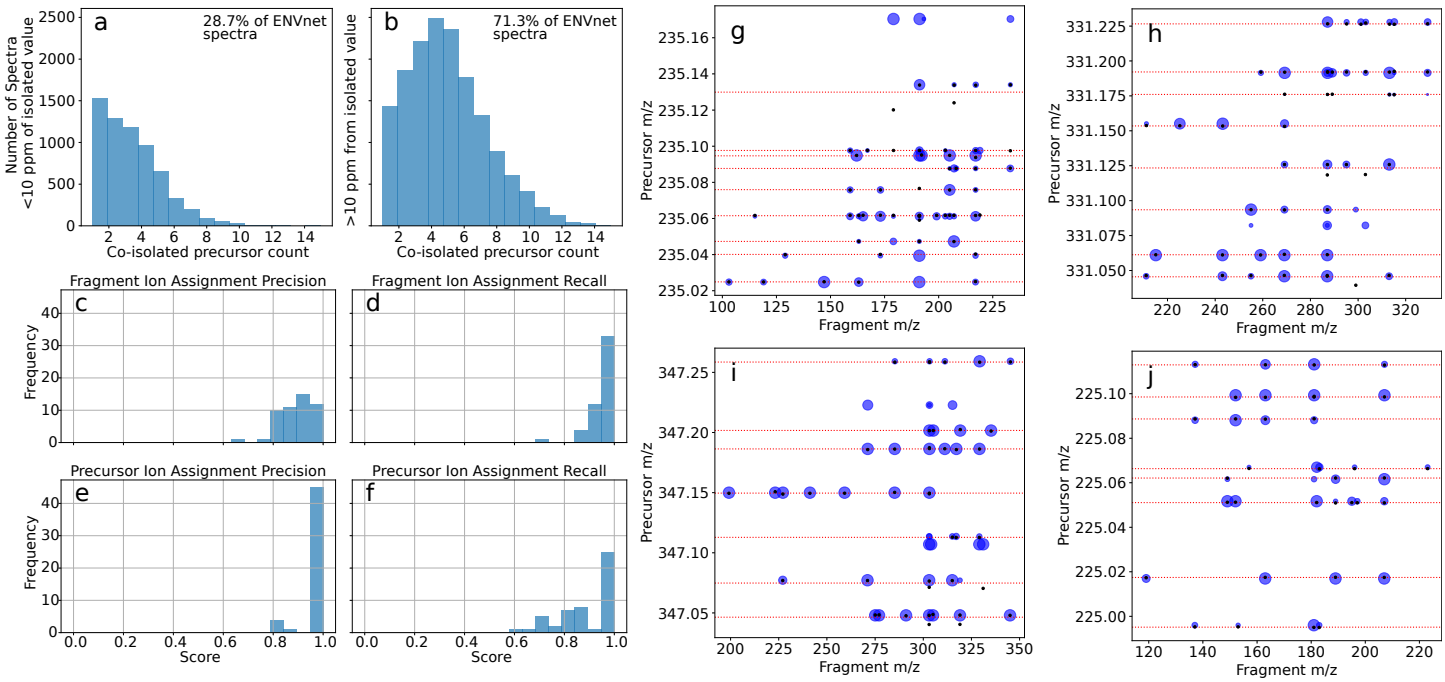


Figure 2. ENVnet computationally separates chimeric spectra, revealing extensive co-isolation and enabling robust precursor-fragment assignment. (a,b) Analysis of a large-scale environmental dataset shows as expected that spectral chimerism is widespread (86%): only 28.7% of MS/MS spectra contain precursors within 10 ppm of the isolation target and would work well for traditional workflows (a), whereas 71.3% include multiple co-isolated precursors with substantial mass differences (b), limiting the reliability of traditional analysis. (c-f) Systematic validation of the deconvolution algorithm on 50 synthetic chimeric spectra demonstrates high performance. The algorithm achieves excellent precision and recall for fragment ion assignment (c,d), correctly linking most fragments to their true parent. Precursor identification attains near-perfect precision (e), indicating no false-positive precursors, and a conservative recall (f) that prioritizes accuracy. (g-j) Representative examples show successful separation of complex mixtures containing up to six co-isolated precursors. Ground-truth fragments (blue circles, plotted at their true precursor m/z) are correctly assigned to the algorithm-inferred parent precursor m/z (red dashed lines), as reflected by the matched ions (black dots). These results illustrate the algorithm's ability to transform a single, highly chimeric spectrum into multiple, clean, and interpretable MS/MS spectra.

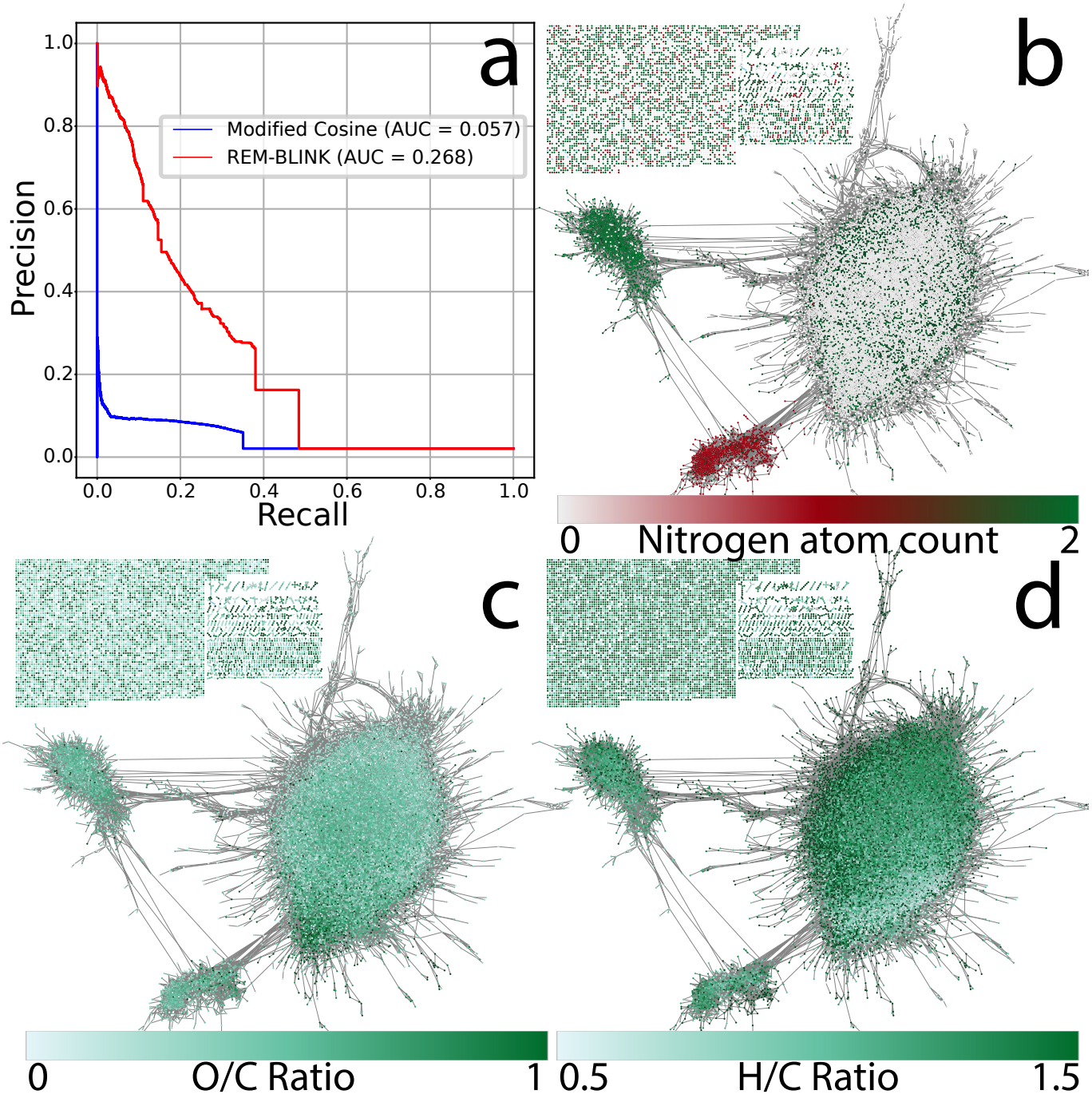


Figure 3. REM-BLINK enhances chemical interpretability in molecular networks. REM-BLINK provides a clearer view of molecular relationships within a network and benchmarks well against Modified Cosine, a comparable spectral similarity algorithm. This results in better node separation and improved visual clarity in 2d visualizations of molecular network layouts. (a) Precision-recall analysis comparing the performance of REM-BLINK (red) and modified cosine (blue) on a library of known structures using maximum common substructure (MCS) greater than 0.7 as ground truth. The area under the curve (AUC) values are shown for each algorithm. (b-d), The molecular network (ENVnet) constructed using REM-BLINK, colored by nitrogen atom count (b), O/C ratio (c), and H/C ratio (d). These visualizations illustrate the chemically coherent network produced by REM-BLINK. By coloring the network according to stoichiometric attributes, clear structural clusters and compositional gradients emerge. This organized topology indicates that the calculated edges capture underlying chemical similarities.

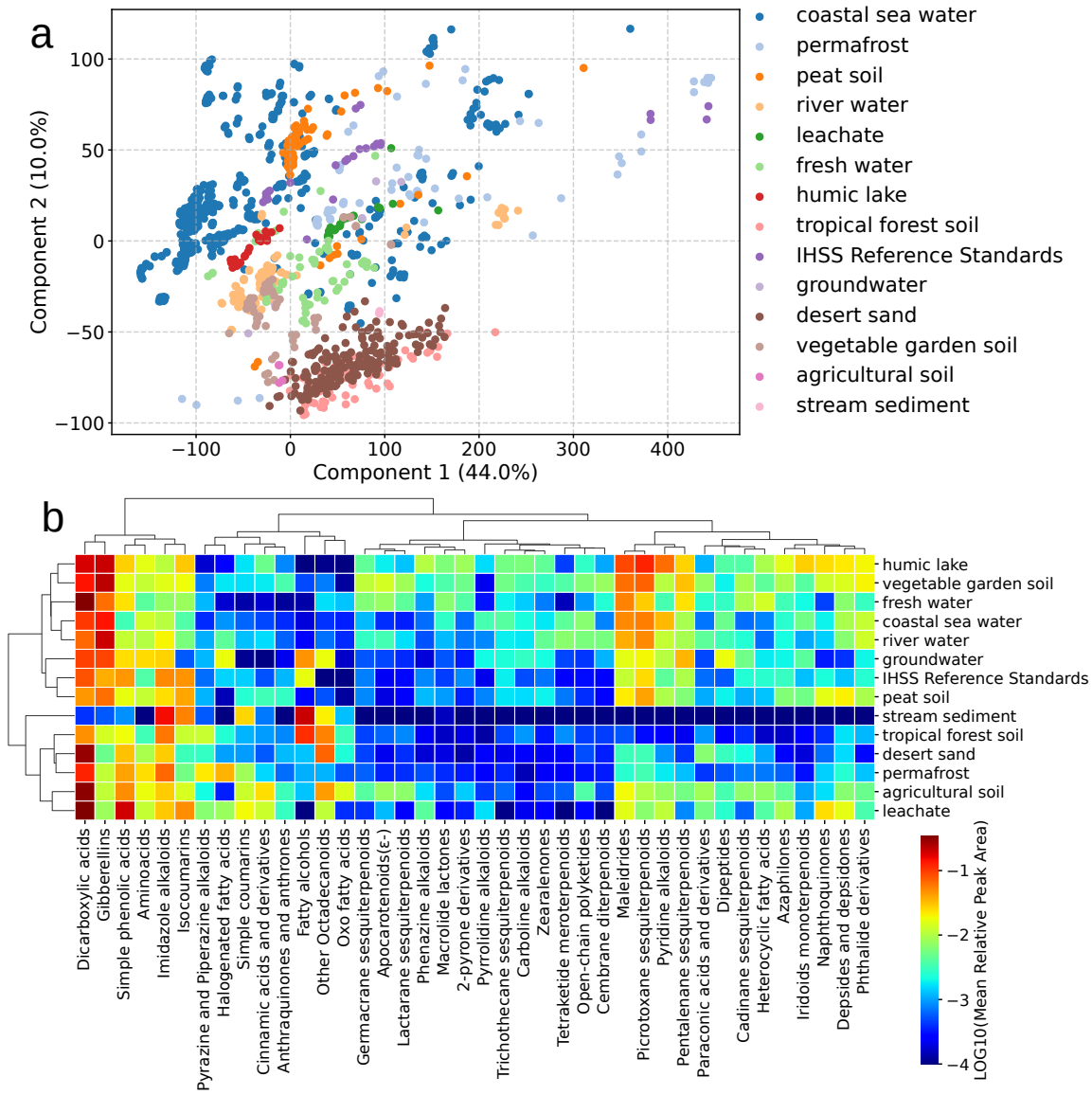


Figure 4. Global comparison of DOM composition reveals conserved and environment-specific molecular signatures. (a) Principal Component Analysis (PCA) of ENVnet node peak areas for all samples, colored by environment type. The plot illustrates the broad separation of samples, indicating unique molecular fingerprint of MS1 peak area across ENVnet for distinct biomes (e.g., coastal seawater vs. permafrost) alongside shared variance. Note that due to the multi-platform nature of the dataset, some observed variance likely reflects technical batch effects in addition to biological differences; quantitative separation should therefore be interpreted with caution. (b) Class averaged hierarchical clustering (Ward linkage, Euclidean distance) of the top 40 most abundant chemical classes (CANOPUS predictions) across environments. The color scale represents the mean log10-transformed sum of relative peak area of all nodes within a given class for each environment matching ENVnet with cosine score greater than 0.5 and 3 or more matching ions. The heatmap highlights a "core" DOM composition of ubiquitous classes (e.g., dicarboxylic acids, simple phenolic acids) present in all environments, contrasted with clusters of environment-specific classes, particularly diverse terpenoids, which are differentially abundant across ecosystem types.

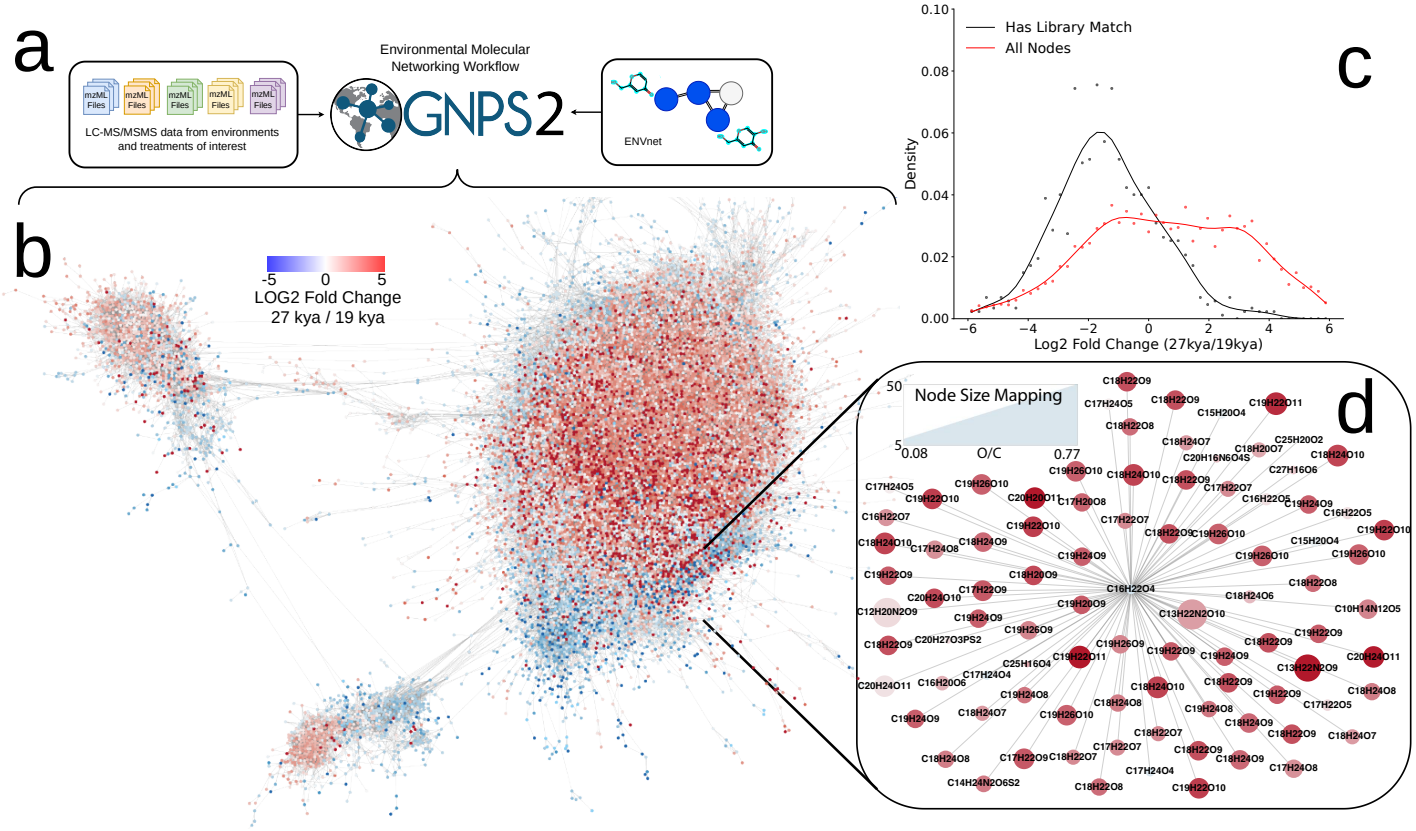


Figure 5. Using ENVnet to analyze and compare molecular compositions in permafrost of different ages. a) Comparisons between environments and treatments can be performed in GNPS2. b) Relative intensity comparisons between 19 kya and 27 kya permafrost reveal broad compositional trends as well as individual molecules whose abundances differ with p permafrost age. c) In line with the view that younger permafrost contains more recognizable necromass-derived compounds, nodes with *MS/MS* library matches (black) were more abundant in the 19 kya samples. d) Molecular networking analysis focusing on the library-matched node with the greatest number of connections to nodes significantly elevated in older (27 kya) permafrost shows that most of these neighboring nodes have MS-BUDDY–assigned formulas consistent with oxidized structural analogs.

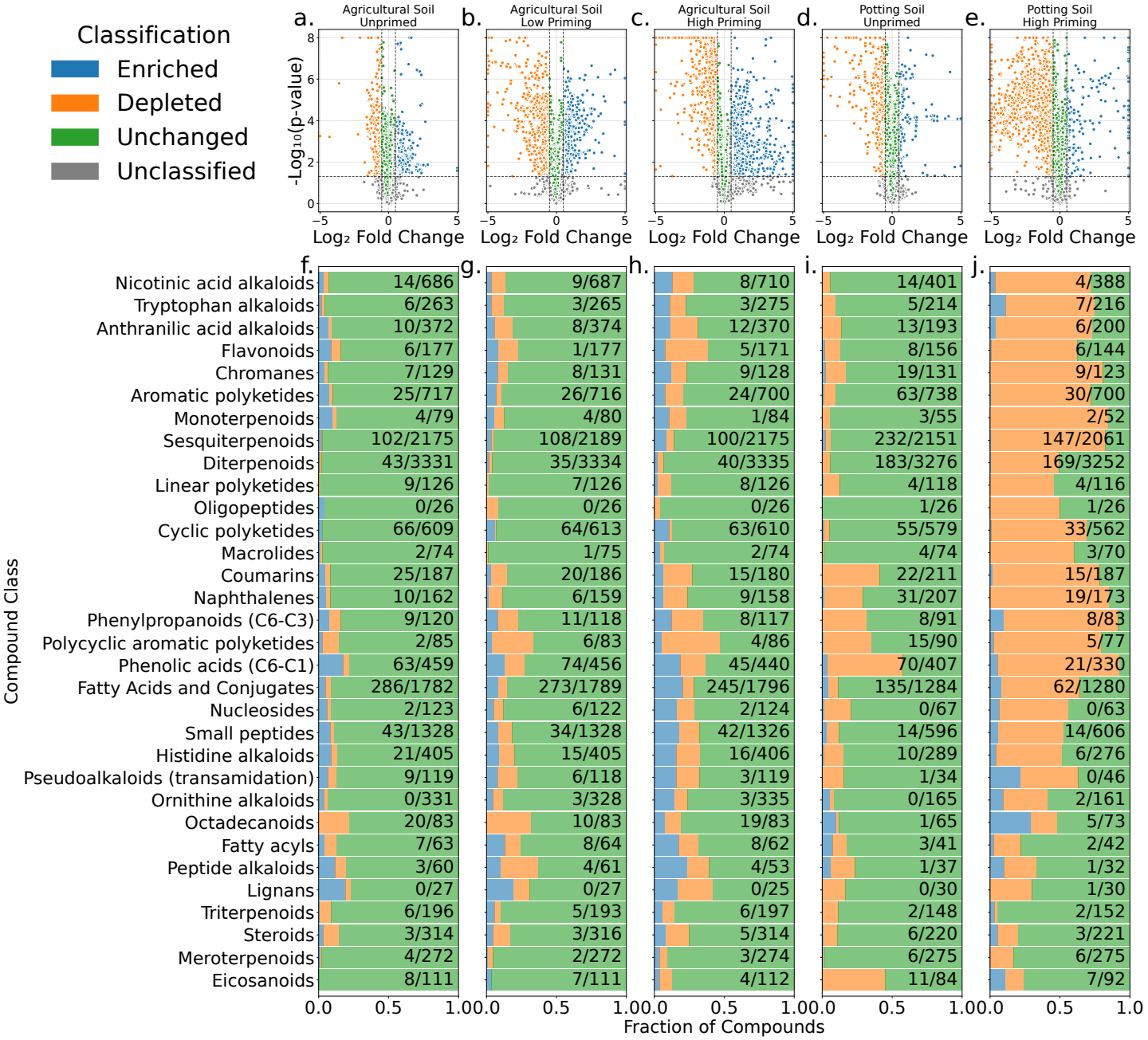


Figure 6. Assessing the inherent persistence of DOM using ENVnet. Organic matter extracted from either potting soil or agricultural soil was incubated with its native microbiome and a range of labile growth media for seven days at room temperature in the dark. a–e, Volcano plots show ENVnet node-level changes in abundance from day 0 to day 7. For each node, day-7 and day-0 peak areas were compared using a two-sided independent-samples Student's t-test with equal variances assumed. Nodes were classified as enriched or depleted if the magnitude of the $\log_2$ fold change was greater than 0.5 and $p < 0.05$. Nodes were classified as unchanged using two one-sided equivalence tests, TOST, implemented as two independent-samples z-tests comparing day-0 and day-7 peak areas with equivalence bounds of $\log_2$ fold change between 0.5 and -0.5; the TOST p-value was defined as the larger of the two one-sided p-values, and unchanged nodes required TOST $p<0.05$. Nodes not meeting the enriched, depleted or unchanged criteria were labeled as unclassified. f–j, Stacked bars show the fraction of ENVnet nodes within each compound class assigned to each treatment category. Compound classes are shown if they contain at least 20 MS1 nodes across the analyzed data. Numbers beside bars indicate the number of nodes with MS/MS evidence and the total number of MS1 nodes in that superclass.

References

1. Lawrence, M. G. *et al.* Evaluating climate geoengineering proposals in the context of the Paris Agreement temperature goals. *Nat. Commun.* **9**, 3734 (2018).
2. Friedlingstein, P. Carbon cycle feedbacks and future climate change. *Philos. Trans. A Math. Phys. Eng. Sci.* **373**, 20140421 (2015).
3. Shaffer, J. P. *et al.* Standardized multi-omics of Earth's microbiomes reveals microbial and metabolite diversity. *Nature Microbiology* **7**, 2128–2150 (2022).
4. He, C. *et al.* Molecular composition of dissolved organic matter across diverse ecosystems: Preliminary implications for biogeochemical cycling. *J. Environ. Manage.* **344**, 118559 (2023).
5. Li, S. *et al.* Dearomatization drives complexity generation in freshwater organic matter. *Nature* **628**, 776–781 (2024).
6. Hertkorn, N. *et al.* Characterization of a major refractory component of marine dissolved organic matter. *Geochimica et Cosmochimica Acta* **70**, 2990–3010 (2006).
7. Dittmar, T. *et al.* Enigmatic persistence of dissolved organic matter in the ocean. *Nature Reviews Earth & Environment* **2**, 570–583 (2021).
8. Hawkes, J. A., Sjöberg, P. J. R., Bergquist, J. & Tranvik, L. J. Complexity of dissolved organic matter in the molecular size dimension: insights from coupled size exclusion chromatography electrospray ionisation mass spectrometry. *Faraday Discuss.* **218**, 52–71 (2019).
9. Simon, C. *et al.* Mass difference matching unfolds hidden molecular structures of dissolved organic matter. *Environmental Science & Technology* **56**, 11027–11040 (2022).
10. Hawkes, J. A., Dittmar, T., Patriarca, C., Tranvik, L. & Bergquist, J. Evaluation of the orbitrap mass spectrometer for the molecular fingerprinting analysis of natural dissolved organic matter. *Analytical Chemistry* **88**, 7698–7704 (2016).
11. Xing, S., Shen, S., Xu, B., Li, X. & Huan, T. BUDDY: molecular formula discovery via bottom-up MS/MS interrogation. *Nature Methods* **20**, 881–890 (2023).
12. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **34**, 828–837 (2016).
13. de Jonge, N. F. *et al.* Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. *Metabolomics* **18**, 103 (2022).
14. Papadopoulos Lambidis, S. *et al.* Two-dimensional liquid chromatography tandem mass spectrometry untangles the deep metabolome of marine dissolved organic matter. *Environ. Sci. Technol.* **58**, 19289–19304 (2024).
15. Zark, M. & Dittmar, T. Universal molecular structures in natural dissolved organic matter. *Nature Communications* **9**, 3178 (2018).
16. Georgiou, K. *et al.* Global stocks and capacity of mineral-associated soil organic carbon. *Nature Communications* **13**, 3797 (2022).
17. Griscom, B. W. *et al.* Natural climate solutions. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 11645–11650 (2017).
18. Watrous, J. *et al.* Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E1743–52 (2012).
19. Yu, J. S. *et al.* A versatile toolkit for drug metabolism studies with GNPS2: from drug development to clinical monitoring. *Nat. Protoc.* 1–35 (2025) doi:10.1038/s41596-025-01237-6.
20. Buttigieg, P. L. *et al.* The environment ontology: contextualising biological and biomedical entities. *Journal of biomedical semantics* **4**, 43 (2013).
21. Jarmusch, A. K. *et al.* ReDU: a framework to find and reanalyze public mass spectrometry data. *Nature Methods* **17**, 901–904 (2020).
22. Xing, S. *et al.* Recognizing Contamination Fragment Ions in Liquid Chromatography-Tandem Mass Spectrometry Data. *Journal of the American Society for Mass Spectrometry* **32**, 2296–2305 (2021).
23. Zhao, T., Xing, S., Yu, H. & Huan, T. De Novo Cleaning of Chimeric MS/MS Spectra for LC-MS/MS-Based Metabolomics. *Analytical Chemistry* **95**, 13018–13028 (2023).
24. Stancliffe, E., Schwaiger-Haber, M., Sindelar, M. & Patti, G. J. DecoID improves identification rates in

metabolomics through database-assisted MS/MS deconvolution. *Nature Methods* **18**, 779–787 (2021).
25. Kushnir, M. M., Rockwood, A. L. & Nelson, G. J. Simultaneous quantitative analysis of isobars by tandem mass spectrometry from unresolved chromatographic peaks. *Journal of Mass Spectrometry* **39**, 532–540 (2004).
26. Nikolskiy, I., Mahieu, N. G., Chen, Y.-J., Tautenhahn, R. & Patti, G. J. An untargeted metabolomic workflow to improve structural characterization of metabolites. *Analytical Chemistry* **85**, 7713–7719 (2013).
27. Harwood, T. V. *et al.* BLINK enables ultrafast tandem mass spectrometry cosine similarity scoring. *Scientific Reports* **13**, 13462 (2023).
28. Schymanski, E. L. *et al.* Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environmental Science & Technology* **48**, 2097–2098 (2014).
29. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
30. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 12580–12585 (2015).
31. Dührkop, K. *et al.* Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* **39**, 462–471 (2021).
32. Davenport, R. *et al.* Decomposition decreases molecular diversity and ecosystem similarity of soil organic matter. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2303335120 (2023).
33. Giera, M., Aisporna, A., Uritboonthai, W. & Siuzdak, G. The hidden impact of in-source fragmentation in metabolic and chemical mass spectrometry data interpretation. *Nature Metabolism* **6**, 1647–1648 (2024).
34. Chandel, A. K., Jiang, L. & Luo, Y. Microbial models for simulating soil carbon dynamics: A review. *Journal of Geophysical Research: Biogeosciences* **128**, (2023).
35. Schwarz, E., Ghersheen, S., Belyazid, S. & Manzoni, S. When and why microbial-explicit soil organic carbon models can be unstable. *Biogeosciences* **21**, 3441–3461 (2024).
36. Wolf, K. M. *et al.* The century experiment: the first twenty years of UC Davis' Mediterranean agroecological experiment. *Ecology* **99**, 503 (2018).
37. Kuzyakov, Y., Friedel, J. K. & Stahr, K. Review of mechanisms and quantification of priming effects. *Soil Biology and Biochemistry* **32**, 1485–1498 (2000).
38. Harwood, T. V., Wang, M., Northen, T. R. & Bowen, B. P. Increasing the scale of the mass spectrometry query language compendium with explainable AI. *Anal. Chem.* **97**, 18860–18866 (2025).
39. Mendez, K. M., Reinke, S. N. & Broadhurst, D. I. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics : Official journal of the Metabolomic Society* **15**, 150 (2019).
40. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4768–4777 (Curran Associates Inc., Red Hook, NY, USA, 2017).
41. Gunina, A. & Kuzyakov, Y. From energy to (soil organic) matter. *Glob. Chang. Biol.* **28**, 2169–2182 (2022).
42. Grant, K. E., Galy, V. V., Haghipour, N., Eglinton, T. I. & Derry, L. A. Persistence of old soil carbon under changing climate: The role of mineral-organic matter interactions. *Chem. Geol.* **587**, 120629 (2022).
43. Mesgar, M., Kia, S., Voroney, P. R., Lo, A. & Gillespie, A. W. Role of aromatic and non-protonated aromatic carbon in the stability of soil organic matter. *Org. Geochem.* **201**, 104897 (2025).
44. Huang, W. *et al.* Contrasting geochemical and fungal controls on decomposition of lignin and soil carbon at continental scale. *Nat. Commun.* **14**, 2227 (2023).
45. Rodríguez-Pérez, R. & Bajorath, J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *J. Comput. Aided Mol. Des.* **34**, 1013–1026 (2020).
46. Boiteau, R. M. *et al.* Relating molecular properties to the persistence of marine dissolved organic matter with liquid chromatography-ultrahigh-resolution mass spectrometry. *Environ. Sci. Technol.* (2024) doi:10.1021/acs.est.3c08245.
47. Koch, B. P., Kattner, G., Witt, M. & Passow, U. Molecular insights into the microbial formation of marine dissolved organic matter: recalcitrant or labile? *Biogeosciences* **11**, 4173–4190 (2014).
48. Malhi, Y. & Grace, J. Tropical forests and atmospheric carbon dioxide. *Trends Ecol. Evol.* **15**, 332–337 (2000).

49. Chacon, S. S. *et al.* Divergent responses of soil microorganisms to throughfall exclusion across tropical forest soils driven by soil fertility and climate history. *Soil Biology and Biochemistry* **177**, 108924 (2023).
50. Schadt, C. *et al.* An integrated metagenomic, metabolomic and transcriptomic survey of Populus across genotypes and environments. *Sci. Data* **11**, 339 (2024).
51. Mackelprang, R. *et al.* Microbial survival strategies in ancient permafrost: insights from metagenomics. *The ISME Journal* **11**, 2305–2318 (2017).
52. Muhs, D. R., Ager, T. A. & Begét, J. E. Vegetation and paleoclimate of the last interglacial period, central Alaska. *Quaternary science reviews* **20**, 41–61 (2001).
53. Hamilton, T. D., Craig, J. L. & Sellmann, P. V. The Fox permafrost tunnel: A late Quaternary geologic record in central Alaska. *GSA Bulletin* (1988).
54. Leewis, M.-C. *et al.* Life at the frozen limit: microbial carbon metabolism across a late pleistocene permafrost chronosequence. *Frontiers in Microbiology* **11**, 1753 (2020).
55. Schirrmeister, L. *et al.* Late Quaternary paleoenvironmental records from the Chatanika River valley near Fairbanks (Alaska). *Quaternary science reviews* **147**, 259–278 (2016).
56. Waldrop, M. P. *et al.* Permafrost microbial communities and functional genes are structured by latitudinal and soil geochemical gradients. *The ISME Journal* (2023) doi:10.1038/s41396-023-01429-6.
57. Liang, C., Schimel, J. P. & Jastrow, J. D. The importance of anabolism in microbial control over soil carbon storage. *Nature Microbiology* **2**, 17105 (2017).
58. Nikrad, M. P., Kerkhof, L. J. & Häggblom, M. M. The subzero microbiome: microbial activity in frozen and thawing soils. *FEMS Microbiology Ecology* **92**, fiw081 (2016).
59. Rivkina, E. M., Friedmann, E. I., McKay, C. P. & Gilichinsky, D. A. Metabolic activity of permafrost bacteria below the freezing point. *Applied and Environmental Microbiology* **66**, 3230–3233 (2000).
60. Liang, C., Amelung, W., Lehmann, J. & Kästner, M. Quantitative assessment of microbial necromass contribution to soil organic matter. *Global Change Biology* **25**, 3578–3590 (2019).
61. Lehmann, J. *et al.* Persistence of soil organic carbon caused by functional complexity. *Nature geoscience* (2020) doi:10.1038/s41561-020-0612-3.
62. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**, 211–221 (2007).

Methods

The following new data sets of samples were generated for this study:

1. The dryland soil dataset consists of 652 topsoil (0-7.5 cm) samples coming from the BIODESERT survey, a standardized global field survey of dryland ecosystems from 25 countries and six continents (Maestre et al. 2022a). See Maestre et al. (2022b) for a complete description of the study sites surveyed and of the methodology used to collect and process these samples[1,2]. A subset of 192 samples from this collection were used here.
2. Organic tomato-corn rotation (OMT) soil samples were collected in autumn 2023 at the UC Davis Russell Ranch Sustainable Agricultural Facility site. Samples were taken from the top soil (0-20 cm) of the profile from plots with an OMT treatment from the long term century experiment[3].
3. Sacramento river delta samples were collected from surface soil from an organic farm field located inside the levee (this field floods periodically because it is near the Sacramento River and not protected by the levee). The exact location was 39°14'22.5"N 122°00'36.2"W, north of the town of Colusa. The field is tilled and has been used to grow a variety of different crop types.
4. Peat soil sample collected from The U.S. Department of Energy's (DOE) Next Generation Ecosystem Experiment (NGEE) Arctic Council (64.51°N, 163.39°W) site in the northeast of Nome on the Seward Peninsula in Alaska. The location is within the sub-Arctic transitional region between boreal forest and tundra. The sample location represents the moist acidic tussock tundra, which is dominated by Sphagnum spp. (moss), Eriophorum vaginatum (cotton grass) and Vaccinium uliginosum (bog blueberry). The soils at this location are classified as Typic Histoturbels in the USDA system. The soil sample was collected in August 2018 from 5cm–20cm in depth after removing non-decomposed dead

plant materials (Oi-horizon). The soil samples were stored on blue ice in the field and at −20 °C during shipment and until further analysis.

5. River sediment sample was collected in July 2017 from The U.S. Department of Energy's (DOE) East River Community Observatory (ER) in the Upper Colorado River Basin (managed by the Lawrence Berkeley National Laboratory Watershed Function Science Focus Area). ER is a 300 square kilometer headwater catchment representative of watersheds in the Rocky Mountains of the Western United States[4 8]. The sediment sample was collected from the Pump House location (8.922°N 106.951°W) which is representative of a lower montane subregion with characteristic meanders and hillslopes. The sediment sample was drained shortly in the field, stored on dry ice during transportation and at −20°C until further analysis.
6. Oak ridge groundwater from well MLSB5 at the Oak Ridge Field Research Centre. MLSB5 was selected for additional sampling due to its central location and constrained sampling depth (screened from 4.31-4.47 m below ground surface) within Area 2, an uncontaminated area of the Oak Ridge Field Research site. The site is an ongoing campaign to understand the relationships between planktonic and sediment microbial communities to groundwater and porewater geochemistry.
7. Freshwater Lakes. Extracted water from oligotrophic, dystrophic, and eutrophic lakes to dissect carbon cycling metabolisms. The study sites include three lakes that are the subject of intense study through the North Temperate Lakes Long Term Ecological Research site: Sparkling Lake (oligotrophic), Lake Mendota (eutrophic), and Trout Bog Lake (dystrophic).
8. Tropical forest soils were collected from four distinct lowland forests across the Isthmus of Panama that range in mean annual precipitation from 2,335 to 3,421 mm. This region experiences a monsoonal climate with a short dry season (December to April). The soils (which include oxisols, alfisols, and ultisols) were sampled to a depth of 10 cm. See Chacon et al., (2023) for further details on site physicochemical characteristics and sampling details[5].
9. Permafrost samples were collected from the US Army Cold Regions Research and Engineering (CRREL) Permafrost Tunnel near Fairbanks, Alaska (64.951°N, 147.621°W) in November of 2017. The tunnel is located in a region of discontinuous permafrost that formed syngenetically through sediment deposition and upward growth of the active layer[6 10]. The permafrost tunnel provides access to late Pleistocene-aged permafrost deposits that are ice- and organic- rich silt (termed Yedoma), at temperatures near -3°C[7]. The age of permafrost located at 20 m and 55 m from the tunnel portal was estimated to be approximately 19,000 calendar years before present (19 kya) and 27,000 calendar years before present (27 kya), respectively[8].

Sample extraction and reconstitution of new samples

The water (J.T.Baker, Allentown, PA, United States), formic acid (Honeywell International Inc, Charlotte, NC, United States) acetonitrile (OmniSolv, Sigma Aldrich, St. Louis, MO, United States), ethyl acetate (Sigma Aldrich), and methanol (Sigma Aldrich) used for sample extraction were of LC-MS grade.

Dissolved organic matter isolation from global BIODESERT soils

In 50mL conical tubes, 1g of each soil was weighed out and then suspended in 5mL of -20C chilled LCMS grade water. The aqueous soil slurries were vortexed for 10 seconds and then bath sonicated in ice water for 1 hour followed by a 1 hour incubation in the ice water bath. Aqueous soil slurries were then shaken at 4C overnight at 200 RPM on an orbital platform shaker. Aqueous soil slurries were then centrifuged at 3,200 rcf for 20 minutes and aqueous supernatants were collected and frozen at -20C in a new tube, after 5 hours samples were transferred to -80C and then lyophilized and stored dry at -80C. Pellets were resuspended in 5mL of 2:1 chloroform:methanol for sequential extraction, vortexed for 10 seconds and then ice bath sonicated for 30 minutes. Chloroform:methanol soil slurries were then shaken at 4C for 1 hour at 200 RPM on an orbital platform shaker. Chloroform:methanol soil slurries were centrifuged and Chloroform:methanol supernatants were collected and dried under nitrogen gas, then stored dry at -80C. Dried aqueous and chloroform:methanol extracts were each resuspended in 0.75mL 100% LCMS grade methanol, vortexed for 10 seconds, ice bath sonicated for 10 minutes, then combined into one tube, vortexed again for 10 seconds and ice bath sonicated for an additional 10 minutes. The combined extracts were then centrifuged at 10,000 rcf, for 5 minutes at 10C;

supernatants were collected in new tubes and dried by vacuum concentration. The dried extracts were resuspended in 150uL of ice cold methanol containing internal standard mix[9], vortexed twice for 10 seconds, ice bath sonicated for 15 minutes and then centrifuged at 10,000 rcf for 5 minutes at 10C; supernatants were then filtered through 0.2 µm modified nylon membrane centrifugal filters (Pall Corporation, Port Washington, NY, United States) using the same centrifugation parameters. Filtrates were transferred to amber glass LCMS vials for LC-MS/MS analysis.

DAC-FA2_20230206 & PyrToDAC_20221219 & SynDAC_20230223 & DOM-TargetedMS2_pilot1

Suwannee River Fulvic Acid standard II (2S101F, IHSS, St. Paul, MN, United States), Suwannee River Fulvic Acid standard III (3S101F, IHSS), Pahokee Peat Fulvic Acid standard III (2S103F, IHSS) and Suwannee River NOM (2R101N, IHSS) were dissolved in either water and/or in methanol at different concentrations ranging from 0-25 mg/mL. The pH of the standards in water was adjusted with 1M KOH to around 5, except for the Pahokee Peat Fulvic Acid standard III and a subset of the Suwannee River Fulvic Acid standard III. For the standards in water, 200 uL was transferred to 1.5 mL Eppendorf tubes and dried by vacuum concentration (ThermoScientific Speed-Vac, Waltham, MA, United States). Dried material was resuspended in 200 uL methanol containing internal standards[9], vortexed twice for 10 seconds, bath sonicated in an ice water bath for 30 minutes and centrifuged at 10,000 xg for 5 minutes at 4°C. The standards in methanol were diluted in 200 uL methanol containing internal standards. The supernatants and the standards diluted in methanol were filtered through 0.2 µm modified nylon membrane centrifugal filters (Pall Corporation, Port Washington, NY, United States) and the filtrates were collected for LC-MS/MS analysis. A subset of the Suwannee River Fulvic Acid standard II and Suwannee River Fulvic Acid standard III diluted in 200 uL methanol containing internal standards were directly collected for LC-MS/MS analysis without filtration.

HumicAcid_20221110

Humic acid (technical, Sigma Aldrich) was dissolved at 10 mg/mL in water with 0.1M KOH. Humic acid was centrifuged for at 10,000 xg for 5 minutes at 4°C. Supernatant was both collected for LC-MS/MS analysis and filtered through a 3kDa filter (Amicon Ultra, Sigma Aldrich), and flow-through was collected for LC-MS/MS analysis.

Sample preparation for FRCgw2_20230830 and FRCgwater_20230105

Groundwater was collected at a constant pump rate of 1 L/min until geochemical parameters (dissolved oxygen, pH, oxidation reduction potential, and conductivity) had stabilized. 10 L was filtered sequentially through 10 µm, 0.2 µm and 0.1 µm polyethersulfone filters in the field. The filtrate was collected in sterile bottles and stored at -80°C until shipment and extraction. Metabolites from groundwater from well-MLSB5 were collected using Solid Phase Extraction (SPE). First, the groundwater (500 mL) was acidified to pH 5 using formic acid (mass spec grade). Next, OASIS HLB SPE cartridges (6 cc, 200 mg sorbent per cartridge, Waters, Milford, MA, United States) were conditioned using 5 mL methanol and equilibrated with 5 mL formic acid (2% v/v) in water. Acidified groundwater was loaded onto the SPE cartridge, and the cartridge was washed with 5 mL water. Metabolites were eluted using 5 mL methanol into glass vials. Either 250 uL, 500 uL or 2 mL of the elution fraction was transferred to 2 mL Eppendorf tubes and dried using vacuum concentration. Samples were resuspended in 150-250 uL methanol containing internal standards[9], vortexed twice for 10 seconds, bath sonicated in an ice water bath for 10 minutes and centrifuged at 10,000 x g for 5 minutes at 4°C. The supernatants were filtered through 0.2 µm modified nylon membrane centrifugal filters and the filtrates were collected for LC-MS/MS analysis.

Sample preparation for OMTSoil50g_Pilot, SacRDelta_20240131& CenturyExpPilot_20230414

OMT soil (0.5 gr/mL, 50 gr/sample), Sacramento River Delta Soil (0.5 gr/mL, 10 gr/sample), soil from Peat plot 14 (N6EE-A Council) (0.25 gr/mL, 5 gr/sample) and East River Sediment (Pump House) (0.25 gr/mL, 5gr /sample) were resuspended in 75%/25% methanol/water. Potting soil (Sunshine MIx, the Home Depot) (0.2/gr/mL, 2 or 5 gr/sample) was resuspended in either water, methanol, acetonitrile or ethyl acetate (water saturated). Soil slurries were bath sonicated in an ice water bath for 15 minutes, and shaken for 1.5-16 hours at 4°C at 180 rpm on an orbital shaker. The methanol/water soil slurries were centrifuged at 3,000 × g for 15 minutes at 4°C and the methanol was removed from the supernatants by vacuum concentration and/or N2

purging. The remaining water fractions were centrifuged at 3,000 × g for 10 minutes at 4°C, supernatants were filtered through a 0.22 um filter (PES membrane filter, Corning, Corning, NY, United States) and dried by lyophilization. The potting soil slurries were centrifuged at 10,000 × g for 10 minutes at 10°C and the supernatants were centrifuged again at 14,000 - 16,000 × g for 10 minutes at 10°C. Supernatants from the methanol, acetonitrile and ethyl acetate slurries were dried by vacuum concentration and/or N2 purging. The aqueous supernatants were dried by lyophilization. Dried material was resuspended in 1.5 mL of methanol, vortexed twice for 10 seconds, bath sonicated in an ice water bath for 15 minutes and centrifuged at 10,000 x g for 5 minutes at 4°C. Supernatants were transferred to 2 mL eppendorf tubes and dried by vacuum concentration. Dried soil samples were resuspended in 100 uL methanol containing internal standards[9], vortexed twice for 10 seconds, bath sonicated in an ice water bath for 15 minutes and centrifuged at 10,000 x g for 5 minutes at 4°C. The supernatants were filtered through 0.2 µm modified nylon membrane centrifugal filters (Pall Corporation, Port Washington, NY, United States) and the filtrates were collected for LC-MS/MS analysis.

GEODES_All3

To extract metabolites, collected water samples between 120-500 mL were frozen and lyophilized dry. Next, 1 mL of LC-MS 100% methanol was added to each dried sample, followed by a brief 10 second vortex, water bath sonication for 10 minutes, then centrifugation at 7500 rcf for 5 minutes to pellet debris. Next, 875 uL of supernatant was transferred to a 2 mL Eppendorf and dried in a SpeedVac. Extracts were stored frozen at -80 degrees C. In preparation for LC-MS, dried extracts were resuspended by adding 500 uL 100% methanol containing internal standards[9], vortexed then sonicated in a water bath prior to centrifugation for 2.5 minutes at 7500 rcf, then supernatant centrifuge-filtered (0.22 µM hydrophilic PVDF membrane, #UFC30GV00, Millipore) for 5 minutes at 7500 rcf and transferred to LC-MS vials.

Pleistocene Permafrost

Cores were collected at 20 and 55 m into the tunnel, as measured from the tunnel portal, using a Snow, Ice and Permafrost Research Establishment (SIPRE) coring auger (7.62-cm inner barrel diameter, 100 cm in length) as described previously [8]. Cores were subsectioned in a room dedicated to processing permafrost samples using measures to prevent contamination. Briefly, the outer ~3 cm was removed and discarded using wet tile saw (run without water) and continuous-rim diamond blades. All material contacting the exposed inner core was sterilized and individuals processing the samples wore lab coats, goggles, gloves, hair coverings, and N95 masks. Additional subsectioning took place in a laminar flow hood that was sterilized between samples.

Set 1: Metabolites were extracted from ~2.5 g soil cores of permafrost soil that had been thawed and immediately processed. Here, 4 ml of methanol were added to each sample, vortexed, then sonicated in an iced water bath for 30 minutes. After centrifugation for 5 minutes at 2935 rcf, supernatant was transferred to 15 mL tubes, then dried in a SpeedVac. Empty tubes were processed using the same methods to serve as extraction controls. All extracts were stored frozen at -80 degrees C. In preparation for LC-MS, dried extracts were resuspended by adding 150 uL 100% methanol containing internal standards [9], vortexed then sonicated in a water bath prior to centrifugation for 2.5 minutes at 2935 rcf, then supernatant centrifuge-filtered (0.22 µM hydrophilic PVDF membrane, #UFC30GV00, Millipore) for 5 minutes at 2823 rcf and transferred to LC-MS vials.

Set 2: Metabolites were extracted from ~3-4 g soil cores of permafrost. Here, 10 ml of methanol were added to each sample, vortexed, then sonicated in an iced water bath for 10 minutes. After centrifugation for 5 minutes at 2935 rcf, supernatant was transferred to 15 mL tubes, then dried in a SpeedVac. Empty tubes were processed using the same methods to serve as extraction controls. Extracts were stored frozen at -80 degrees C. In preparation for LC-MS, dried extracts were resuspended in 100% methanol containing internal standards[9], vortexed then sonicated in a water bath prior to centrifugation for 2.5 minutes at 2935 rcf, then 180 uL supernatant centrifuge-filtered (0.22 µM hydrophilic PVDF membrane, #UFC30GV00, Millipore) for 5 minutes at 2823 rcf and transferred to LC-MS vials. Resuspension volume varied for each sample to normalize by weight of soil extracted (13.3 g/mL).

Tropical forest soils

Metabolites were extracted from 2 g of soil in 5 mL tubes. Here, 2.5 mL of methanol was added to each sample, vortexed until thoroughly mixed (4 seconds, 3x), then sonicated in a water bath for 10 minutes. After centrifugation for 5 minutes at 2627 rcf, methanol supernatant was transferred to another 5 mL tube. Another round of extraction was performed by adding 2 mL methanol to the original tube containing soil, vortexed until the pelleted soil from the previous centrifugation step was disrupted, then sonicated in a water bath for 10 minus. Following centrifugation using the same parameters, methanol supernatant was removed and combined with the previously collected supernatant. Extracts were then dried in a SpeedVac and stored frozen at -80 degrees C. In preparation for LC-MS, dried extracts were resuspended in 180 uL of 100% methanol containing an internal standard[9], vortexed then sonicated in a water bath prior to centrifugation for 2.5 minutes at 2627 rcf, then centrifuge-filtered for 5 minutes at 2823 rcf and transferred to LC-MS vials.

LC-MS/MS parameters

Metabolites in the samples were separated using reverse phase on an Agilent 1290 LC stack connected to a Q Exactive Hybrid Quadrupole-Orbitrap Mass Spectrometer (Thermo Fisher Scientific)[9]. Samples were injected once on reverse-phase LC–MS/MS (Agilent ZORBAX RRHD Eclipse Plus C18, 95Å, 2.1 × 50 mm, 1.8 µm column) LC–MS/MS parameters are described in Protocols.io (JGI/LBNL Metabolomics - Standard LC-MS/MS ESI Method - Nonpolar C18)[9]. Briefly, after separation, the top two most intense precursor ions (not previously selected within 7 s) were selected for fragmentation at stepped then averaged collision energies of 10, 20, 40 eV. Sample injections were randomized and both internal and external standards were used for quality control purposes. Details in SI Table 4.

Methods for incubation of organic matter under high/low labile carbon

Dissolved Organic matter isolation from potting soil

Potting soil (Sunshine® Mix #1 - Sun Gro Horticulture) was used to prepare potting soil dissolved organic matter. Potting soil samples (0.17 gr/mL, 6 gr in total) were resuspended in a salt solution (10.15 mM $MgSO_4$, 12.5 mM NH_4Cl, 8.5 mM $CaCl_2$, 20.81 mM NaH_2PO_4 and 33.3 mM K_2HPO_4). Soil slurries were bath sonicated in an ice water bath for 15 minutes, and then shaken for 16 hours at 4°C at 200 rpm on an orbital shaker. The soil slurries were sieved through a stainless steel sieve, centrifuged at 1,000 ×g for 20 minutes at 4°C and supernatants were centrifuged again at 2,000 ×g for 20 minutes at 4°C. The collected supernatant, the dissolved organic matter, was used in microbial processing experiments.

Dissolved Organic matter isolation from agricultural soil

Organic tomato-corn rotation (OMT, see appendix S1) soil samples were collected in autumn 2023 at the UC Davis Russell Ranch Sustainable Agricultural Facility site (see Appendix S1). In total, 2640 grams of field-moist soil from the OMT plots was used to prepare OMT soil dissolved organic matter. Dissolved organic matter was prepared by weighing 12 x 220 grams field-moist soil into 12 2L glass beakers (baked at 400C, 60 minutes) and resuspended in 12 x 440 mL 75% LC/MS grade methanol / 25% LC/MS grade water. Soil slurries were bath sonicated in an ice water bath for 15 minutes, and then shaken for 1.5 hoursat 4°C at 180 rpm on an orbital shaker. Slurries were centrifuged at 3,000 ×g for 15 minutes at 4°C and the methanol was removed from the supernatants by vacuum concentration (Thermo Speed-Vac) and/or N2 purging. The remaining water fractions were centrifuged at 3,000 ×g for 10 minutes at 4°C, filtered through a 0.22 um filter (PES membrane filter, Corning 431098) and dried by lyophilization. The dried dissolved organic matter were resuspended in 264 mL LC/MS grade water, which corresponded to approximately 10 g soil per mL, bath sonicated in an ice water bath for 15 minutes and centrifuged at 3,000 ×g for 10 minutes at 4°C. The dissolved organic matter was filter sterilized using a 0.22 um PES filter and contained 673 mg/L C (Multi N/C 2100S Duo, Analytik Jena, Jena, Germany). The dissolved organic matter was stored at -80C and used in microbial processing experiments.

Total organic carbon measurements were made using a Analytik Jena Multi N/C 2100S Duo total organic carbon analyzer equipped with a CLD nitrogen detector, platinum catalyst and ultra high purity oxygen carrier gas. Briefly, the samples were acidified on the instrument with hydrochloric acid, purged with carrier gas for 180 second before measurement, and then two rinse cycles were performed, followed by two technical replicate injections of 100uL for measurements of non purgeable organic carbon and total nitrogen with 10 second additional purging between injections; the oven was set at 800°C. The instrument was calibrated with hydrogen potassium phthalate and potassium nitrate solutions prepared in Milli-Q type 1 ultrapure water. Carbon and nitrogen standards, and water blanks were injected before and after samples for quality control purposes.

Microbial Processing of Dissolved Organic Matter from Potting soil

To 10 mL of the collected dissolved organic matter, either 3.3 mL of 2x Northen Lab Defined Medium (NLDM) or a salt solution (10.15 mM $MgSO_4$, 12.5 mM NH_4Cl, 8.5 mM $CaCl_2$, 20.81 mM NaH_2PO_4 and 33.3 mM K_2HPO_4), 0.1 mL Wolfe's Vitamin supplement (MD-VS™) (ATCC) and 0.1 mL Wolfe's Trace Mineral supplement (MD-TMS™) (ATCC) was added[10]. Samples were incubated at room temperature in the dark without mixing. After 7 days, supernatants were collected by centrifugation at 7,000 xg for 10 minutes at 4C. Acidified supernatants (pH 5, 2% formic acid) were extracted with OASIS HLB SPE cartridges (1 cc, 30 mg sorbent per cartridge, Waters). First, OASIS HLB SPE cartridges were conditioned using 1 mL methanol (mass spec grade) and equilibrated with 1 mL water (mass spec grade). After loading acidified supernatants, cartridges were washed with 1 mL water, retained metabolites were eluted with 1 mL methanol and samples were dried by vacuum concentration. Dried samples were resuspended in 100 µL methanol containing internal standards, vortexed twice for 10 seconds, bath sonicated in an ice water bath for 15 minutes and centrifuged at 10,000 xg for 5 minutes at 4°C. The supernatants were filtered through 0.2 µm modified nylon membrane centrifugal filters (Pall Corporation, Port Washington, NY, United States) and the filtrates were collected for LC-MS/MS analysis.

Microbial Processing of Dissolved Organic Matter from agricultural soil

In 15 mL conical tubes, 5 mL OMT dissolved organic matter was mixed with 0.8 mL salt solution (10.15 mM $MgSO_4$, 12.5 mM NH_4Cl, 8.5 mM $CaCl_2$, 20.81 mM NaH_2PO_4 and 33.3 mM K_2HPO_4), 0.1 mL Wolfe's Vitamin supplement (MD-VS™) (ATCC) and 0.1 mL Wolfe's Trace Mineral supplement (MD-TMS™) (ATCC). Next, labile metabolites and/or soil microbial inoculum were added to 3 groups, except for the 'experimental control' group, to which water was added instead. The microbial inoculum was prepared from the same OMT soil as the dissolved organic matter. Under sterile conditions, 1.5 grams of OMT soil was added to a sterile 15 mL conical tube and mixed with 7.5 mL filter sterilized Mili-Q water. The soil was vortexed until resuspended and incubated at room temperature for 1 hour. The supernatant was collected by centrifugation at 2,500 xg for 10 minutes at 4C and used as the inoculum. To the 'biotic control' group (agricultural soil NA), 0.25 mL of the soil microbial inoculum was added. To the 'high labile carbon' group (agricultural soil Hi), 0.25 mL of the soil microbial inoculum and 2.5 mL of 2x Northen Lab Defined Medium (NLDM) was added as easily utilizable substrates (521 mg/L C and 62 mg/L N). To the 'low labile carbon' group (agricultural soil Lo), 0.25 mL of the soil microbial inoculum and 0.25 mL of 2x NLDM was added resulting in a lower concentration of labile carbon (52 mg/L C and 6 mg/L N). Final volume was brought up to 10 mL with filter sterilized water. All groups contained 4 biological replicates and were incubated at room temperature in the dark. After 7 days, supernatants were collected by centrifugation at 7,000 xg for 10 minutes at 4C. DOM was extracted from acidified supernatants (pH 5, 2% formic acid) with Solid Phase Extraction C18 cartridges (Bond Elut C18 1 gr, Agilent) to concentrate DOM components and remove inorganic salts. After extraction, cartridges were washed with acidified water (pH 5) and retained DOM was eluted with 4 mL methanol, dried by vacuum concentration. Dried samples were resuspended in 100 µL methanol containing internal standards, vortexed twice for 10 seconds, bath sonicated in an ice water bath for 15 minutes and centrifuged at 10,000 x g for 5 minutes at 4°C. The supernatants were filtered through 0.2 µm modified nylon membrane centrifugal filters (Pall Corporation, Port Washington, NY, United States) and the filtrates were collected for LC-MS/MS analysis.

Data Availability

New data generated as part of this study is available at MassIVE (link to data) under accession MSV000095785. Existing data already in MassIVE can be found in SI Table 1a. The GNPS2 ENVnet workflow job link for the analysis of two ages of permafrost. The ENVnet network is available in the main code repository as a GraphML file or node and edge tables; original spectra and MDM spectra are available for each node. A video tutorial for running and interpreting outputs for the ENVnet GNPS2 workflow is available on YouTube.

Code Availability

Two repositories contain the code for this work. The main ENVnet: https://github.com/biorack/envnet and the GNPS2 Workflow repository: https://github.com/biorack/envnet_gnps2.

Methods-only references

1. Maestre, F. T. *et al.* The BIODESERT survey: assessing the impacts of grazing on the structure and functioning of global drylands. *Web ecology* **22**, 75–96 (2022).
2. Maestre, F. T. *et al.* Grazing and ecosystem service delivery in global drylands. *Science* **378**, 915–920 (2022).
3. Wolf, K. M. *et al.* The century experiment: the first twenty years of UC Davis' Mediterranean agroecological experiment. *Ecology* **99**, 503 (2018).
4. Hubbard, S. S. *et al.* The east river, colorado, watershed: A mountainous community testbed for improving predictive understanding of multiscale hydrological–biogeochemical dynamics. *Vadose Zone Journal* **17**, 1–25 (2018).
5. Chacon, S. S. *et al.* Divergent responses of soil microorganisms to throughfall exclusion across tropical forest soils driven by soil fertility and climate history. *Soil Biology and Biochemistry* **177**, 108924 (2023).
6. Kanevskiy, M. *et al.* Yedoma cryostratigraphy of recently excavated sections of the CRREL permafrost tunnel near fairbanks, alaska. *Frontiers in Earth Science* **9**, (2022).
7. Douglas, T. A. & Mellon, M. T. Sublimation of terrestrial permafrost and the implications for ice-loss processes on Mars. *Nature Communications* **10**, 1716 (2019).
8. Mackelprang, R. *et al.* Microbial survival strategies in ancient permafrost: insights from metagenomics. *The ISME Journal* **11**, 2305–2318 (2017).
9. Louie, K. B. *et al.* JGI/LBNL Metabolomics - Standard LC-MS/MS ESI Method - Nonpolar C18. (2024).
10. de Raad, M. *et al.* A defined medium for cultivation and exometabolite profiling of soil bacteria. *Frontiers in Microbiology* **13**, 855331 (2022).