scientific reports

Check for updates

OPEN

Molecular diversity of dissolved organic matter reflects macroecological patterns in river networks

Erika C. Freeman[1,2], Maruti K. Mudunuru[3], Kelli L. Feeser[4], Emily Ann McClure[5], Ricardo González-Pinzón[6], Christopher S. Ward[7], Eric M. Bottos[8], Stefan Krause[9,10], Jasquelin Peña[11,12] & Michelle E. Newcomer[12]✉

Deciphering dissolved organic matter (DOM) molecular complexity is crucial for understanding ecosystem function. Using the continental-scale Worldwide Hydrobiogeochemistry Observation Network for Dynamic Rivers Systems (WHONDRS) Fourier-transform ion cyclotron resonance mass spectrometry (FTICR-MS) dataset, we reveal fundamental scaling patterns of DOM chemodiversity with watershed characteristics. Analysis of 54 river sites shows local and regional watershed features significantly influence DOM chemodiversity (2500–8718 unique formulae), exhibiting consistent scaling patterns across compound classes and a novel latitudinal gradient (decreasing diversity with increasing latitude). Scaling relationships for DOM composition vary by compound class. Crucially, the scaling parameters (B, baseline chemodiversity; Z, sensitivity) are linearly interrelated. This B–Z relationship is most robust for potentially bio-labile carbohydrates (coefficient of determination $R^2 \approx 0.85$), diminishing for recalcitrant, plant-derived molecules (such as lignin), and indicates (potential) biolability-dependent coupling between baseline diversity and environmental responsiveness. These quantitative scaling relationships, with scaling exponents ranging from – 2.1 to 2.2 across compound classes, enable prediction of DOM composition across watersheds, offering a framework to understand ecosystem responses to environmental change. This research bridges biogeochemistry and ecology, providing tools to anticipate molecular transformations across scales.

The importance of dissolved organic matter chemodiversity

Dissolved organic matter (DOM) represents a major form of organic carbon across aquatic ecosystems, fueling stream metabolism and mediating key biogeochemical processes in aquatic ecosystems[1–3]. Despite its importance for drinking water quality and ecosystem function[4,5], our ability to predict DOM composition across river networks remains limited due to its molecular complexity and the diverse processes shaping its distribution. Understanding DOM chemodiversity—defined here as the number of molecular formulae detected in a sample—is particularly important given that two-thirds of drinking water in the United States alone is sourced from surface river waters, making the prediction of DOM composition critical for drinking water management and ecosystem health assessment; and is expected to become one of the most politically charged topics in coming decades[6].

[1]Department of Plant Sciences, University of Cambridge, Cambridge, UK. [2]Department of Aquatic Ecology, Swiss Federal Institute of Aquatic Science and Technology (Eawag), Überlandstrasse 133, 8600 Dübendorf, Switzerland. [3]Subsurface Science Group, Pacific Northwest National Laboratory, Richland, WA, USA. [4]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, USA. [5]Geisel School of Medicine at Dartmouth, Hanover, NH, USA. [6]Gerald May Department of Civil, Construction and Environmental Engineering, University of New Mexico, Albuquerque, NM, USA. [7]Center for Great Lakes and Watershed Studies, Bowling Green State University, Bowling Green, OH, USA. [8]Department of Biological Sciences, Thompson Rivers University, Kamloops, BC, Canada. [9]University of Birmingham, Birmingham, UK. [10]Ecologie des Hydrosystèmes Naturels et Anthropisés (LEHNA), Université Claude Bernard Lyon 1, 69622 Villeurbanne, France. [11]Department of Civil and Environmental Engineering, University of California, Davis, USA. [12]Lawrence Berkeley National Laboratory, Earth & Environmental Sciences Area, Berkeley, CA, USA. ✉email: mnewcomer@lbl.gov

Drivers and limitations in understanding riverine DOM composition

DOM chemodiversity in river corridors is influenced by numerous factors, including land-use practices, in-stream processes such as mixing, transformation, and degradation, as well as broader characteristics of the watershed[7–9]. Both hydrologic factors such as flow regime and residence time[10,11] and non-hydrologic factors including crop cover, land use intensity, and soil characteristics[12] significantly influence DOM composition and distribution. Transformation mediated by microbial degradation[13–15], photooxidation[16–19], and sorption[20,21], as well as autochthonous production (e.g. plant, algal, and microbial exudation[22]), can significantly alter the quality and chemistry of DOM in river corridors. Additionally, anthropogenic perturbations have drastically changed DOM fluxes in many systems[23,24]. While global studies of DOM composition exist[25], they often face limitations due to sparse spatial sampling, particularly in large river systems where only a few locations might represent vast drainage areas. Recent large-scale analyses, such as those leveraging the WHONDRS dataset, are beginning to provide more comprehensive geographical insights[25,26]. The development and expansion of comprehensive global databases for high-resolution DOM data, analogous to resources like OpenFluor for fluorescence spectroscopy[27], remain crucial for enabling robust meta-analyses and cross-system comparisons to fully understand global DOM patterns.

Scaling laws as a framework for DOM pattern analysis

Scaling laws—mathematical relationships that describe how system properties change with size or scale, typically expressed as power laws ($y = Bx^a$)—provide quantitative frameworks for understanding how system properties change across spatial and temporal scales. Scaling laws can reveal emergent patterns and provide predictive insights, even in the face of significant spatial and temporal heterogeneity[28,29]. While scaling approaches have been applied to various hydrobiogeochemical processes in river networks[30–32], scaling remains a grand challenge in watershed studies[33,34], and a comprehensive dataset linking DOM composition to a broad suite of watershed characteristics and environmental drivers has been lacking. Reported DOM composition between and across river corridors varies drastically[35], and no single dataset exists to identify and quantify scaling law relationships as organisational principles in riverine DOM patterns across scales.

We leveraged the Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS) dataset, which offers a unique combination of continental-scale coverage across diverse watersheds and standardized analytical methods using high-resolution mass spectrometry. While previous WHONDRS publications have examined various aspects of DOM composition in river networks[26,36–38], our study is the first to apply scaling laws to understand DOM chemodiversity patterns across watersheds and to identify consistent scaling relationships that can be used for predictive modeling[39]. Our approach bridges descriptive characterization and predictive modeling, offering a new perspective on systematic DOM variation across spatial scales. See Table S1 to compare our study to other WHONDRS publications. The dataset is publicly available through ESS-DIVE (https://data.ess-dive.lbl.gov/view/10.15485/1729719) and Zenodo https://doi.org/10.5281/zenodo.12789204.

Hypotheses

We hypothesize that: (1) local and non-local scaling properties exhibit consistent patterns that predict chemodiversity across watersheds and latitudes; (2) patterns are driven by both local (in-stream) and non-local (watershed-scale) environmental factors; (3) molecular and ecological scaling laws show similar behavior (Fig. 1). We tested these by analyzing 54 streams (Fig. 2, Fig. S1), 300+ watershed features and examining compound classes (lignin-like, tannin-like, etc.) to assess how scaling laws predict molecular formula abundance.

Materials and methods

WHONDRS

The WHONDRS dataset offers an opportunity to explore consistent macroecological patterns associated with DOM composition and distribution in the form of scaling laws across biomes and latitudes, and across both local and non-local properties. Riverine DOM chemical composition was generated by the WHONDRS Consortium[45] (https://whondrs.pnnl.gov). Samples were collected, extracted, and analysed as per[36], see Supplementary Methods for details. This Northern Hemisphere survey of surface river water and sediment samples provides a novel dataset to explore the predictive capacity of scaling relationships to explain the spatial organisation of DOM within river corridors. Across 97 sites and 9 countries, for 6 weeks between July and August 2019, the WHONDRS network organised the collection and characterization by Fourier-transform ion cyclotron resonance mass spectrometry (FTICR-MS) of global surface water and sediment samples. The WHONDRS network also collected a variety of in-situ measurements of water quality parameters, including temperature, pH, dissolved oxygen, and specific conductance. We used statistical methods including multiple regression analysis and principal component analysis to explore the relationships between DOM composition and watershed characteristics, and to identify key drivers of DOM chemodiversity across spatial scales. The WHONDRS data package as well as the R code are hosted on the ESS-DIVE data archive (https://data.ess-dive.lbl.gov/view/10.15485/1484811). Statistical analyses were performed using python (version 3.10.13) with data centred and scaled to unit variance. Prior to analysis, the data were log-transformed to reduce skewness and improve normality.

Watershed characteristics

Three different sources of watershed metadata representing non-local features (HydroSHEDS, StreamStats, and EPAWaters), and local WHONDRS metadata measured directly in-situ, were used to compile characteristics associated with each WHONDRS stream sampling point of chemodiversity. We used R version 4.3.3 and packages including 'streamStats', 'sf', and 'leaflet' to read, extract, and co-locate all watershed variables with each

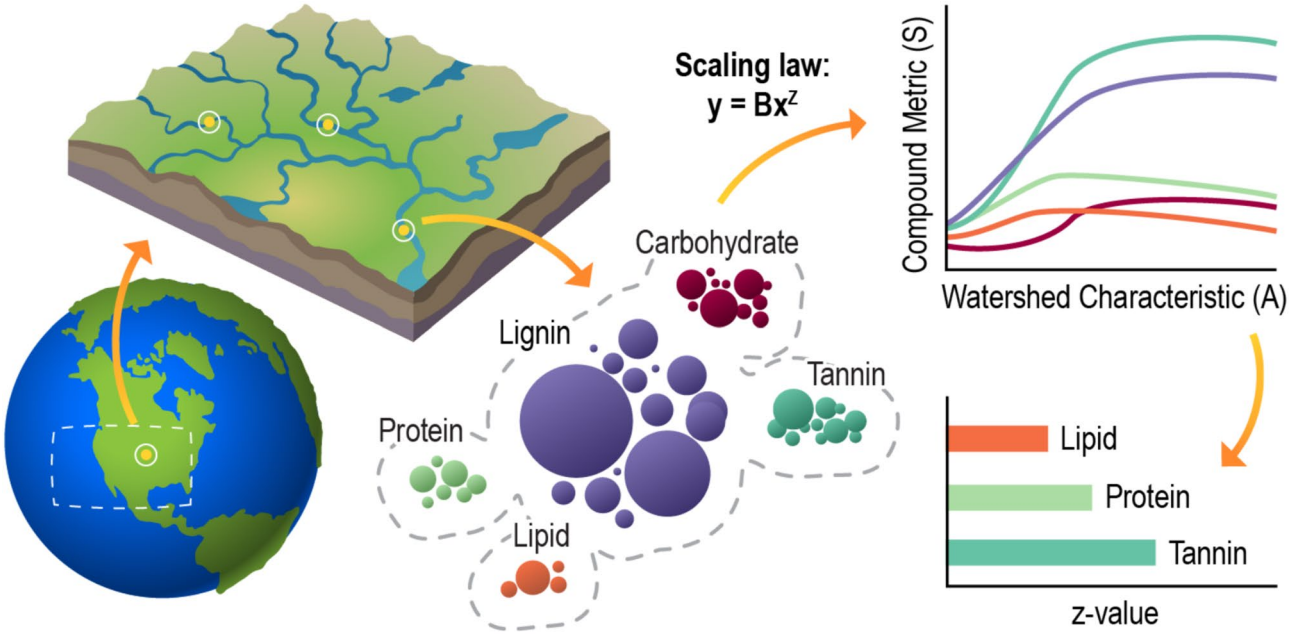

Fig. 1. When complex mixtures of organic compounds from sediment and stream water are separated into classes, the direction and shape of the patterns (y-axis, such as: chemodiversity, rarity, dominance, compound evenness, and formula richness) are hypothesised to scale with watershed characteristics (x-axis, such as: area, stream branching complexity, and latitude). Further, exponents of scaling equations (i.e. z-axis in the generalised scaling equation: $y = Bx^Z$) should vary among compound classes revealing generalisable patterns that reflect common degradation patterns across global to regional scales, including across latitudes, continents, and watersheds.

individual WHONDRS stream sampling point. All watershed data, metadata files, and variable names associated with this manuscript can be found in the data repository[46] and associated SI files:

WHONDRS
Local metadata collected during sample collection included water temperature, dissolved oxygen (DO), pH, canopy cover (shading), water column height, sunlight access, and stream order. These metadata were reported for each individual sample location and recorded by each individual who collected the unique sample.

HydroSHEDS
Stream sampling points were spatially matched to river and stream shapefiles from the global RiverATLAS database, version 1.0, which is a subset of the broader HydroATLAS product at 15 arc-second (~500 m) resolution[47] (https://www.hydrosheds.org/hydroatlas). RiverATLAS provides hydro-environmental information for all rivers of the world, both within their contributing local reach catchment and across the entire upstream drainage area of every reach. A full description of the data can be found at[47]. This information was derived by aggregating and reformatting original data from well-established global digital maps and by accumulating them along the drainage network from headwaters to ocean outlets[47]. Corresponding river reaches were then matched to sub-basin characteristics for hierarchically nested watersheds using the database BasinATLAS[47]. These catchments and stream reaches contain 281 individual attributes, representing 56 different hydro-environmental variables, each associated with the twelve sub-basin polygon layers of BasinATLAS and the line segments of RiverATLAS.

StreamStats
Watershed metadata were obtained from the USGS StreamStats database[48]. StreamStats includes characteristics such as drainage area, elevation, mean precipitation, and percent imperviousness for the specific drainage area draining to a particular location on the stream network. StreamStats delineates the drainage basins for user-selected sites on streams and can be done from any latitude and longitude within the CONUS. We delineated the unique drainage basin for each WHONDRS location and used StreamStats to calculate a variety of physical and climatic statistics for each watershed. All characteristic types and codes can be found on the StreamStats web portal (https://streamstats.usgs.gov/information-portal/). Unique to StreamStats is the calculation of flow statistics for any location on the stream network (gaged or un-gaged).

EPA-waters
The EPA Watershed Assessment, Tracking & Environmental Results System (WATERS) has a similar data architecture to StreamStats, where the portal integrates drainage characteristics from various EPA water programs

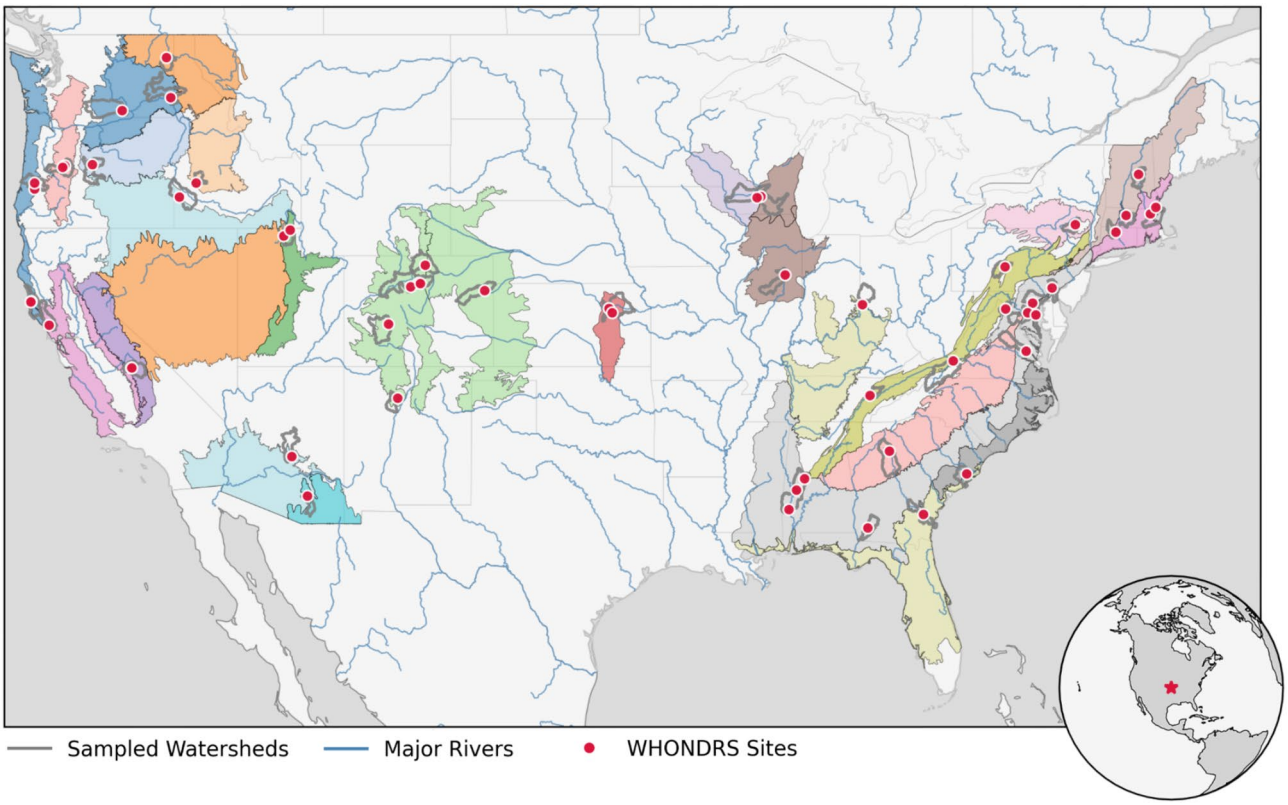

Fig. 2. Study design. Map of 54 WHONDRS river sampling sites (orange-yellow) overlaid on three geographic layers: EPA Level III Ecoregions (n = 105, coloured polygons), HUC8 catchments (~ 1800 km^2, grey outlines), and major rivers (blue lines). Only ecoregions intersecting sampling sites are shown[40–44].

by linking it to the national surface water network (https://www.epa.gov/waterdata/waters-geoviewer). For each latitude/longitude associated with the WHONDRS sampling sites, we queried a watershed report that contains both catchment attributes (local area attributes draining directly to the selected stream segment) and watershed attributes (attributes of the entire drainage area extending from the outlet upstream to the headwaters). Over 400 attributes are available through EPA-Waters.

DOM analysis

The molecular composition of DOM was characterized using ultrahigh-resolution mass spectrometry. Surface water and sediment extracts were analyzed on a 12 Tesla Bruker SolariX Fourier transform ion cyclotron mass spectrometer (resolution: 220 K at 481.185 m/z) at the Environmental Molecular Sciences Laboratory (Richland, WA, USA). Negative mode spectra were collected from 100 to 900 m/z with appropriate ion accumulation times for each sample type[36].

Raw spectra were processed using a signal-to-noise ratio threshold of 7. Molecular formulae were assigned using established software pipelines, including Formularity[49] and ftmsRanalysis[50]. Quality control measures included removing peaks outside the 200–900 m/z confidence range. Formulae were classified into compound class based on their elemental ratios: lignin-like (O/C: 0.1–0.7, H/C: 0.7–1.5), tannin-like (O/C: 0.6–1.2, H/C: 0.5–1.5), lipid-like (O/C: 0–0.3, H/C: 1.5–2.5), protein-like (N > 0, O/C: 0.2–0.6, H/C: 1.5–2.2), carbohydrate-like (O/C: 0.7–1.2, H/C: 1.5–2.2), and condensed aromatic formulae (O/C: 0–0.7, H/C: 0.2–0.7)[51–53]. Potentially biolabile formulae were identified using H/C ratios > 1.5 following the molecular lability boundary approach defined by D'Andrilli et al.[54]. See Supplementary Methods and Discussion for further detail.

Feature importance, scaling, and principal component analysis

To reduce the size of the curated environmental variable set (Fig. 3, Step 1–2), we first eliminated variables with minimal impact using correlation analysis (Step 4: Fig. 3, Table S2). Key variables were identified using Spearman correlation for each extrinsic variable in relation to chemodiversity. A correlation coefficient threshold was set at an absolute value of > 0.23 (Spearman), > 0.25 (Pearson), respectively. This threshold was selected to ensure a balance between variable relevance across all datasets and to avoid the 'curse of dimensionality' (i.e., limited samples and large number of features)[55–57].

To further refine the feature selection, we employed feature importance analysis (Step 4, Fig. 3). We applied multiple feature importance methods (n = 6) to bolster the selection processes' reliability, effectively reducing method-specific biases. The six sensitivity analysis methods we used for feature importance include the Pearson correlation coefficient, Spearman's rank correlation coefficient, F-test, Mutual information (MI),

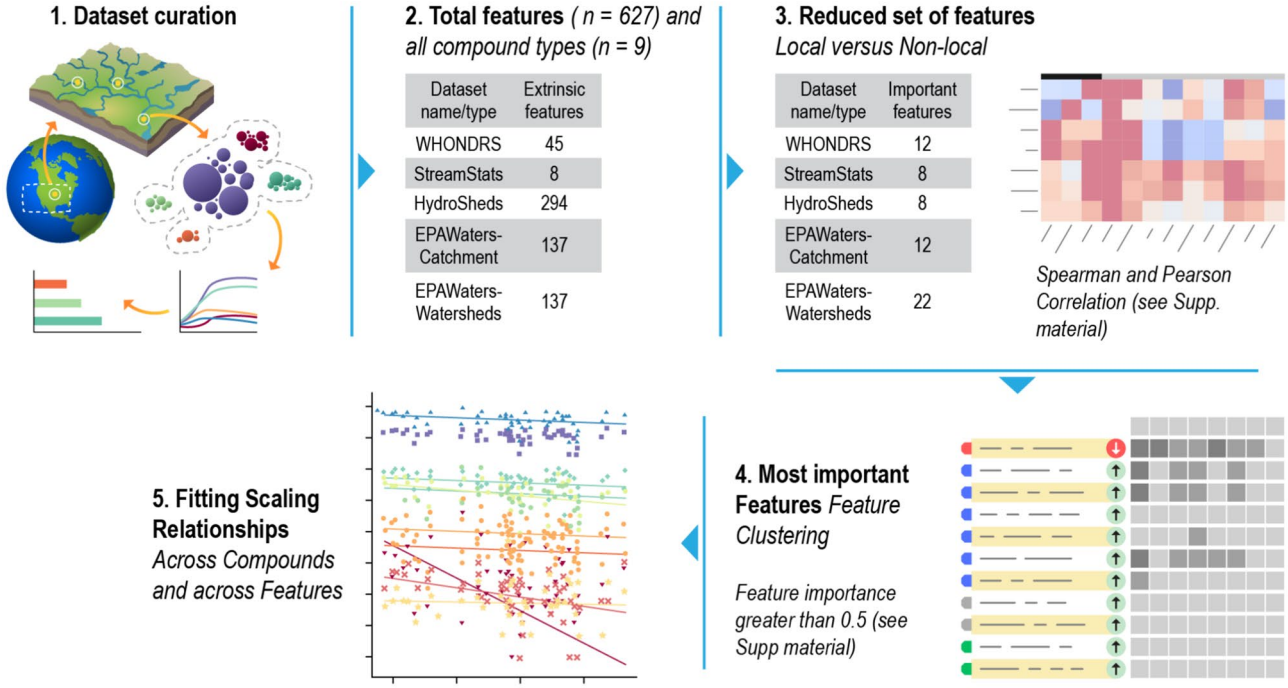

Fig. 3. Scaling methodology: Step 1: Dataset curation, including collecting all data and conducting QA/QC on all data. Step 2: Aggregating all data to the correct scales and co-locating watershed data with DOM data. Step 3: Reducing the number of features using six different sensitivity methods (i.e., Spearman's correlation, Pearson correlation, F-test, Mutual information, Feature importance based on Random Forests, and SHAPley values) to reduce bias in selecting key features for developing scaling laws[56,58]. Step 4: Feature selection using feature clustering. Any feature importance greater than 0.5 constitutes a behavioural set for which feature importance values are deemed to satisfy user-defined performance metrics (e.g., averaged importance values greater than 0.5 across multiple sensitivity analysis methods to reduce any bias in selecting variables for developing scaling laws)[59,60]. Step 5: Fitting scaling relationships to compound class and most important features.

Random Forest (RF), and SHAPley values. Pearson correlation measures the linearity between extrinsic descriptors and chemodiversity, while Spearman's correlation measures this relationship's non-parametric measure of monotonicity. F-test is a univariate feature selection that allows us to select the best features based on univariate statistical tests. MI measures the strength of non-linear dependency between the variables and chemodiversity. It is equal to zero if chemodiversity is independent of extrinsic variables, and higher values mean higher dependence. An RF model with 100 decision trees was trained, and essential extrinsic descriptors were evaluated using permutation-based feature importance[61]. This technique allows us to measure the increase in the prediction error of the RF model after we permute the extrinsic descriptor's values. This permutation method breaks the relationship between the feature and the actual outcome. SHAPley is a method based on cooperative game theory and is used to increase transparency and interpretability of model-agnostic feature selection models such as RF[62]. It provides an average of all the marginal contributions of a variable considering all possible combinations[63].

To enable a systematic and data-driven selection of key extrinsic variables, we produced a single aggregate score by averaging the importance scores across all employed feature importance methods, leading to a robust variable ranking system. Variables achieving an average importance score above 0.5 were identified as significantly influential, a benchmark inspired by studies that include ML-model predictions under limited sample data,a large set of features, and model-data integration case studies[56,58,60]. These variables were then selected for further analysis. By averaging feature importance scores across six statistical tests and machine learning algorithms (after normalising for comparability), we provide a robust and multifaceted assessment of variable influence. This approach mitigates potential biases from any single technique, offering a more reliable indicator of each feature's true importance (Step 4: Fig. 3).

After feature selection, we fit scaling relationships between the selected extrinsic variables and chemodiversity following a log–log transformation to investigate potential power-law distributions (Step 5: Fig. 3). A power law function ($y = Bx^Z$) describing the relationship between local or non-local descriptors and chemodiversity for each compound class is used to develop scaling laws. Here, 'y' is the chemodiversity, 'x' is the descriptor, B is the power-law coefficient, and Z is the scaling exponent. We use *SciPy* Python non-linear least squares to fit this power law scaling function to input–output data. Statistical tests are performed to compute the p-value and check if the estimated parameters of the power law (i.e., B and Z) are significant. To test the statistical significance of the estimated parameters (scaling exponents) in the power-law model, we used the Ordinary Least Squares (OLS) regression. In Python, the statsmodels library provides the implementation of the OLS regression through

SciPy's ordinary least squares function (Ref.[64], v 0.13.4). The statsmodels library estimates the coefficients of the power-law model (scaling exponent and intercept) and calculates their standard errors and p-values.

The important features in Fig. 5b are assembled into a dense data matrix of size 54 × 11 with 54 samples characterized by 11 key features. Prior to dimensionality reduction using PCA, each feature is independently centered and scaled to unit variance using Scikit-learn's StandardScaler class in Python[65]. In this pre-processing step, the sample mean and standard deviation for each feature are computed and stored. These statistics are then applied via the StandardScaler's transform method to ensure that every variable contributes equally, regardless of its original scale. Once the data are standardized, principal component analysis is carried out using Scikit-learn's PCA implementation. The algorithm performs a full singular value decomposition (SVD) via LAPACK[66], yielding two principal components, which focus on the most informative directions in the feature space. These two principal components, often referred to as loadings, define the directions of maximum variance within the standardized data matrix and quantify the correlation between the original features or variables and the new component axes. The importance of each component is measured by its explained variance, with components ordered by descending explained variance ratio (i.e., the percentage of total variance that each component captures). In Fig. 5a, the length of each arrow corresponds to the magnitude of a feature's loading on a particular component. Longer arrows indicate stronger relationships between the original variables and the principal components, thereby highlighting which features drive the major axes of variability in the dataset.

Results

Variables identified through feature importance

Of a possible 627 environmental features, we identified 62 that were important for predicting chemodiversity (Step 2: Fig. 3). These features were selected based on a combination of univariate statistical tests and machine learning algorithms; normalised scores are plotted in Fig. 4b (see "Materials and methods" section). A binomial test ($p < 0.05$) confirmed that the 62 identified features exceeded what could be expected by chance, demonstrating the strength of our analysis. These results reveal groups of significant predictors that merit further exploration for potential scaling relationships.

Figure 4b shows that diverse environmental variables significantly influence chemodiversity across spatial scales (Fig. 4b). Actual evapotranspiration emerged as the most critical variable based on its average importance score, underscoring that evapotranspiration is an important driver of chemodiversity (i.e., relative importance score of 1.0). The critical variables that followed most closely were mean (± std) annual stream temperature

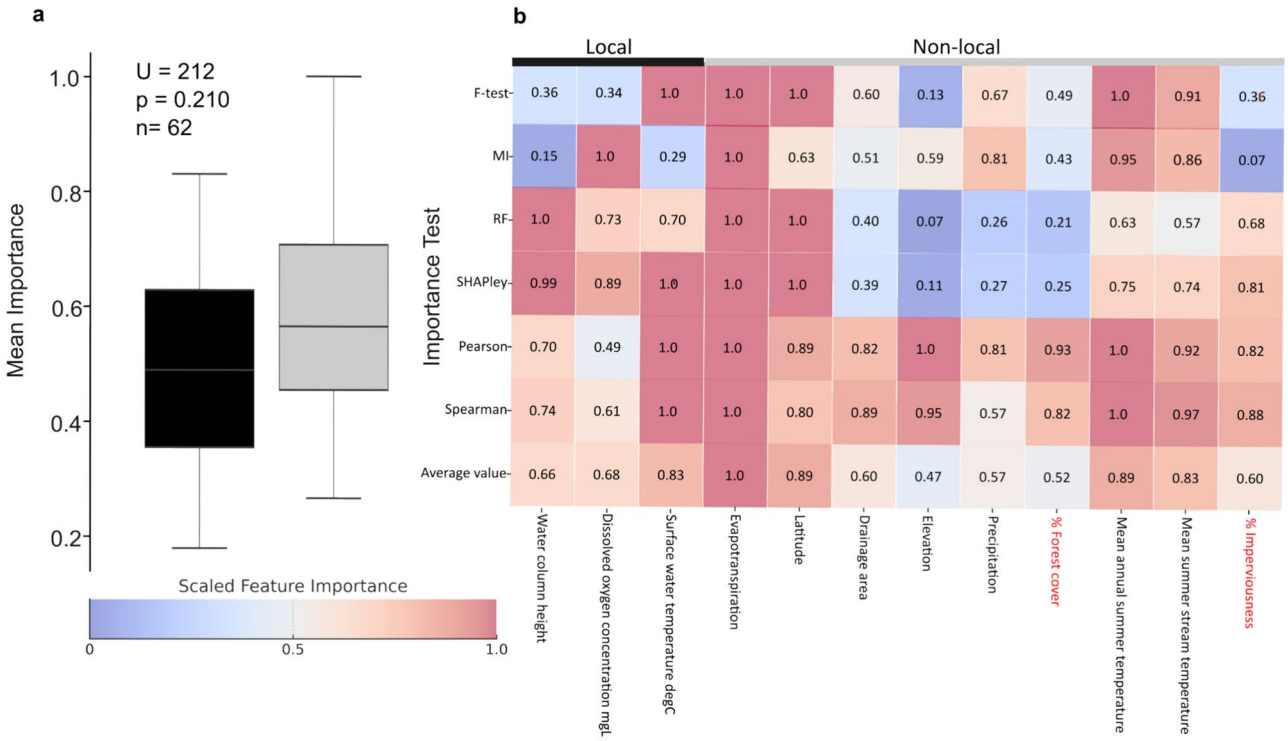

Fig. 4. Feature importance across diverse environmental features. (**a**) Local features were on average not more important than non-local features tested with Kruskal–Wallis non-parametric test ($U = 212$, $p = 0.210$, $n_{local} = 11$, $n_{non\text{-}local} = 51$). (**b**) Example of normalised feature importance (FI) values for chemodiversity vs. the most important local (upper black bar) and non-local (upper grey bar) environmental features across six statistical tests, i.e.,: Pearson correlation coefficient, Spearman correlation coefficient, F-test, Mutual information (MI), Random Forest (RF), and SHAPley tests (cutoff of > 0.5). The arithmetic mean of the 6 feature importance values is shown in the last row. Anthropogenic features in red.

(0.83 ± 0.27), latitude (0.89 ± 0.13) and various measures of watershed characteristics and anthropogenic impacts, such as terrestrial radiation index density (0.87 ± 0.15) and nonpoint source pollution density (0.84 ± 0.14). A list of the average relative importance of all variables can be found in Supplementary Information (Table S2). Together, our results highlight the multifaceted drivers of chemodiversity, suggesting that both natural processes and anthropogenic activities play significant roles in shaping DOM characteristics in aquatic ecosystems.

The next layer of analysis was focused on evaluating whether local and non-local parameters impacted chemodiversity differently. Importance of local features (n = 11, 18% of important variables) was, on average, highest for water column height, dissolved oxygen concentration, and surface water temperature (mean FI (± std) = 0.66 (0.31), 0.68 (0.21), and 0.83 (0.28), respectively, number of importance metrics = 6). The most important non-local features (mean FI ± std) were: evapotranspiration (1.0 ± 0.0), latitude (0.89 ± 0.30), drainage area (0.60 ± 0.39), elevation (0.47 ± 0.39), precipitation (0.56 ± 0.23), percent forest cover (0.52 ± 0.27), annual and summer stream temperature (0.89 ± 0.15 and 0.83 ± 0.15), and percent imperviousness (0.81 ± 0.24) (Fig. 4b). However, we found no clear statistically significant difference in importance between local (stream variables) and non-local (watershed) features (Fig. 4a, U = 212, p = 0.210, $n_{local} = 11$, $n_{non\text{-}local} = 51$). This finding underscores the need to consider both in-stream conditions and wider regional watershed characteristics to uncover the drivers of DOM chemodiversity.

Scaling relationships with biogeographical variables

Of the 62 environmental features important for predicting chemodiversity, we found that the features could be grouped, using PCA, into five environmental categories: geospatial (latitude), climate (potential ET, surface/air temperature), geomorphic (area, discharge), land-use (imperviousness, toxic discharge sites), and in-stream (water column height) (Fig. 5a). By classifying these 62 key features into five environmental categories, we streamlined the identification of primary drivers of chemical species diversity, thus facilitating more targeted

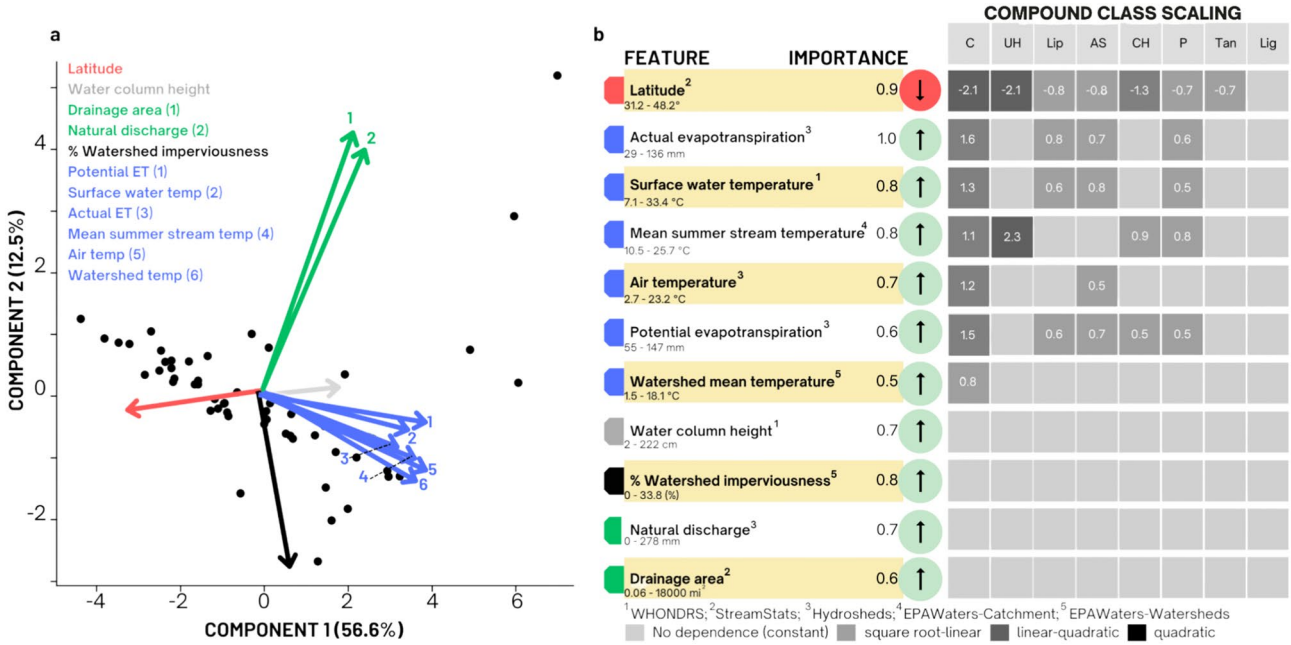

Fig. 5. (**a**) PCA biplot showing the grouping of variables into five broad classes geospatial drivers (latitude, red), watershed drivers (watershed area, discharge, green), climate drivers (potential ET, surface/air temp, blue), land-use (% watershed imperviousness, black), and in-stream drivers (water column height, grey). Numbers in parentheses refer to each numbered line. (**b**) Importance of features, strength, and directionality of the scaling relationship. The superscript on different features indicates the database source (Table S2). Ranges in raw values are below the feature name. Of the 621 total features and 62 important features, we present those with significant scaling exponent and the intercept term (log10(B)) in the power-law model ($p < 0.05$; Ordinary Least Squares regression). The average feature importance is the mean importance across the methods used for sensitivity analysis, namely: Pearson correlation coefficient, Spearman correlation coefficient, F-test, Mutual Information, Random Forest, and SHAPley values. The direction of the scaling relationship, i.e., either increasing (upward arrow) or decreasing (downward arrow) is presented to the right of the feature importance value. Dark grey boxes with Z values > 2 indicate formulae where the scaling exponent is greater than 2. Feature importance values are summarised in Supplementary Information (Figs. S2, S3, S4, and S5). Z values are summarised in SI Table S3. Formulae that are: *As* amino sugar-like, *C* carbohydrate-like, *Lip* lipid-like, *P* protein-like, *UH* unsaturated hydrocarbon-like, *CH* condensed hydrocarbon-like, *Lig* lignin-like, *Tan* Tannin-like. This figure was generated using Python (version 3.10.13, https://www.python.org/) with scikit-learn (version 0.24.0, https://scikit-learn.org/) for PCA analysis and feature importance calculations. Final figure formatting was completed using Canva (https://www.canva.com/).

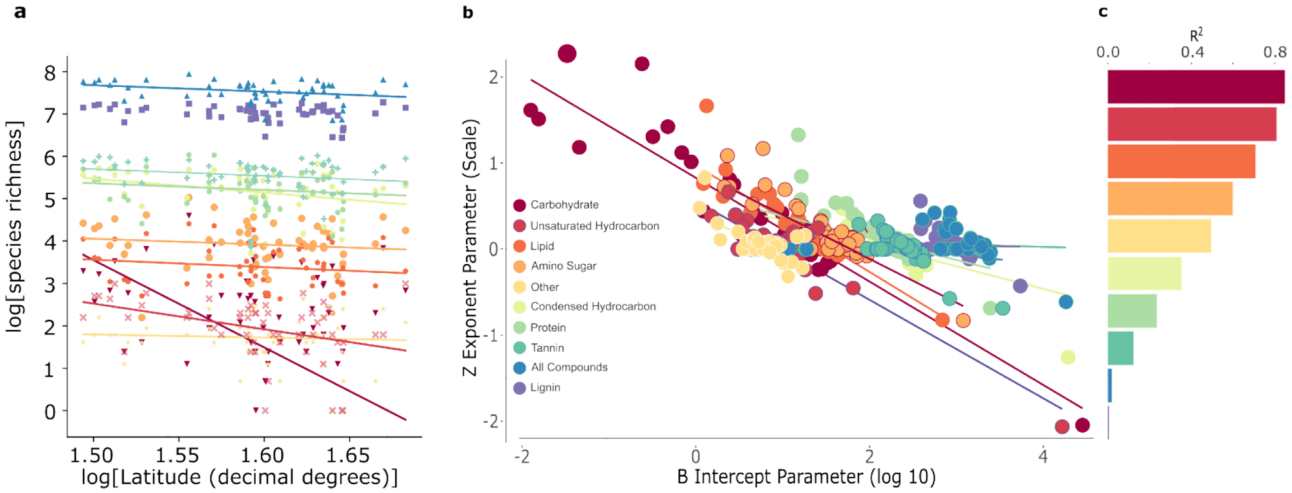

Fig. 6. (**a**) Scaling relationship between chemodiversity and latitude for each compound class. Each dot represents one in-situ data for a particular watershed. Colours represent different DOM compound classes including *As* amino sugar-like, *C* carbohydrate-like, *Lip* lipid-like, *P* protein-like, *UH* unsaturated hydrocarbon-like, *CH* condensed hydrocarbon-like, *Lig* lignin-like, *Tan* Tannin-like, and others. Lines represent linear models fit to each compound class where each linear scaling relationship between the compound class and particular watershed features has an associated B and Z value. (**b**) Scaling parameters B (intercept, x-axis) and Z (exponent, y-axis) for chemodiversity coloured by the compound class aggregated across all watershed features. Lines represent linear models fit to each compound class where each dot represents the paired B-Z value for each watershed feature and compound class linear scaling model. (**c**) Barplot of R^2 values associated with each linear model in b. Colours vary by the R^2 magnitude. In Supplementary Fig. S5, the B and Z values for each compound class are also coloured by the five different data sources (WHONDRS, StreamStats, HydroSHEDS, EPA-Watershed, and EPA-Catchment). The scaling exponent and the statistical significance values for all environmental features are found in the Supplementary Information (SI Table S4). Z and B values for all compound classes and all features (important and unimportant) are found in the data publication[46].

future studies of chemodiversity. Watershed and stream features, similarly grouped based on the results of PCA with averaged feature importance environmental variables that showed statistically significant correlations ($p < 0.05$) with chemodiversity (boxes filled with numbers of Z scaling parameters when $Z > 0.5$) showing the direction and strength of the scaling relationship. A scaling exponent greater than 2 indicates a supra-linear (quadratic) or accelerating scaling relationship (negative direction) for carbohydrates and unsaturated hydrocarbon-like formulae. Mostly linear, and positive direction scaling relationships were found for all of the other watershed features.

We computed scaling relationships between chemodiversity and categorised features with mean importance ≥ 0.5 ($n = 40$) (Fig. 6a). We used the Jenks natural breaks optimisation method to select this 0.5 threshold. We further report only those relationships where both the scaling exponent (Z) and intercept (log10(B)) were significant ($p < 0.05$, Ordinary Least Squares regression). Across compound classes, scaling with watershed latitude revealed systematic variations in relationship strength, direction, and parameters (Fig. 6a). All compound classes exhibited negative scaling exponents (Z) with latitude, with carbohydrate-like formulae showing the steepest decline. Intercepts decreased systematically from lignin-like to carbohydrate-like, suggesting class-specific baseline chemodiversity levels.

The relationship between scaling parameters (Z, B) was linear across compound classes but shifted toward higher B values, consistent with microbial diversification of labile formulae (Fig. 6b)[67]. Parameters were generally within the range of those reported in other studies (e.g.[68]). The unexpected linear Z-B relationships may reflect consistent functional processes despite watershed heterogeneity. The strong correlation between high chemodiversity (large positive B) and rapid chemodiversity decline with scale (strong negative Z) suggests the importance of proximal terrestrial sources in headwater watershed DOM variability[69,70].

Discussion

Local vs. non-local influences on DOM chemodiversity

Our findings illustrate that both local (e.g., water temperature, dissolved oxygen) and non-local (e.g., evapotranspiration, latitude) factors significantly shape DOM chemodiversity. While water column height strongly influenced local DOM characteristics, broader factors such as evapotranspiration emerged as critical predictors of chemodiversity. This challenges the traditional view that in-stream processes predominantly govern DOM composition, highlighting instead the interplay between immediate environmental conditions and broader watershed characteristics, as hypothesised. Similarly, research has shown that both local heterogeneity (e.g., soil pH, conductivity) and broader regional-scale factors drive microbial community composition and

diversity, highlighting the interplay between immediate environmental conditions and larger-scale drivers[71]. The lack of a statistically significant difference in the importance of local versus non-local features underscores the need for a more nuanced understanding of their interactions across scales.

Biogeographical patterns

The strongest scaling relationship ($Z = -2.1$) was a sub-linear scaling of carbohydrate-like chemodiversity with latitude ($\text{chemodiversity} = -18{,}207 \times \text{Latitude}^{-0.61}$), as expected if freshwater chemodiversity mirrors traditional biogeographical patterns[67,72] (see Supplementary Data Files: scaling_exponent_array.csv, b_array.csv). Like the well-established latitudinal increase in species richness for terrestrial organisms[73,74], we observed higher chemodiversity towards the southern end of our environmental gradient (Fig. 5b). This marks one of the first reported pieces of evidence of such a latitudinal gradient in freshwater chemodiversity[26]. This pattern suggests that shared macroecological factors may shape the arrangement of life on Earth at both the organismal and molecular levels.

The specific drivers of these gradients may differ between environments and latitudes. For instance, precipitation, which dominates the diversification of biological species at low latitudes[75], did not emerge as a strong predictor of DOM chemodiversity in this study's latitudinal range (i.e., 31.18° N to 48.17° N). Instead, temperature and evapotranspiration significantly influenced DOM chemodiversity (Fig. 5b). This highlights the importance of considering water balance and evapotranspirative fluxes in understanding and predicting molecular diversity patterns in freshwater ecosystems.

Urbanisation, marked by increased impervious surfaces, surprisingly leads to higher chemodiversity in riverine DOM in our dataset. This contrasts with the expected decrease due to reduced OM processing in simplified landscapes. This increase is driven by the diverse organic inputs from human activities (fertilisers, personal care products, pharmaceuticals, waste, etc.), and that cities are complex concentrations of top-down human and climatic events[76].

Increased aquatic chemodiversity could be conceptually linked to greater diversity within both microbial and terrestrial plant communities. Although we didn't directly measure microbial diversity, research indicates positive relationships between adjacent trophic levels[77,78] and covariation between chemodiversity and microbial diversity[79]. This implies that aquatic microbial communities and terrestrial plant inputs may jointly drive higher chemodiversity (or vice-versa). It is possible that the information content in diversity is conserved[80–82].

If we consider organic matter as part of the aquatic food web, both aquatic microbial and terrestrial plant diversity could drive higher chemodiversity. Chemodiversity reflects the combined contributions of diverse source materials (e.g., terrestrial plant detritus) and subsequent microbial processing. Existing evidence suggests a link between chemodiversity and microbial diversity, though its geographical generalisability beyond lake[79] and estuary[83,84] environments remains uncertain. Our results hint at a potentially stronger connection between observed chemodiversity and terrestrial plant diversity compared to in-stream microbial diversity, suggesting that catchment vegetation may play a more dominant role in shaping DOM composition.

Watershed climatic features, particularly actual evapotranspiration and surface or air temperature, emerged as significant determinants of chemodiversity, mirroring the latitude-driven biodiversity trends observed in plant and microbial communities[73]. This finding is well-supported by theory since in-stream OM is a reflection of both in-stream and catchment vegetation and microbial biomass and associated products such as microbial respiration[79,85]. This raises a fundamental question: does chemodiversity primarily reflect the richness of source material entering the system, or does it instead indicate enhanced microbial processing rates? For instance, elevated temperatures can accelerate the biochemical reactivity of organic matter (OM) across ecosystems[86]. This occurs through processes such as depolymerization, increased microbial enzyme production and reactivity, and changes in soil OM microbial-lability[87]. Solar radiation via UV exposure further induces OM transformation[88–90].

Projected increases in air and surface water temperatures across our study region[91] underscore the need to investigate the specific chemical processes that drive chemodiversity. Warming is likely to increase DOM chemodiversity due to its cascading effects on primary production, which shifts DOM composition and microbial-lability[92,93]. Elevated temperatures also enhance microbial decomposition efficiency, potentially altering the functional role of microbial communities in transforming DOM pools[94,95]. These cascading effects highlight the interconnectedness between climate change, microbial activity, and the potential transformation of DOM pools. While our study was not designed to explicitly rank the relative importance of climatic covariates and their interactions, the broad consistency of our findings with the latitudinal diversity pattern suggests a major role for factors related to these climatic drivers.

General scaling relationships among compound classes

We observed consistent positive relationships between the number of potentially biolabile formulae (amino sugar, carbohydrate, lipid, and protein-like formulae) and key environmental drivers (Fig. 5b, Table S3). These formulae, despite representing a minority of the overall molecular pool, significantly shaped the observed variation in chemodiversity (less than 15% of the total number of unique molecular formulae identified, see also[96]). This finding supports recent studies showing universal microbial degradation processes in soil can lead to convergence in overall molecular pool features[67], while the chemodiversity of specific compound classes continues to increase[87]. Furthermore, microbial-driven processes are closely tied to environmental gradients across scales, reinforcing the idea that potentially biolabile compound classes are likely strongly influenced by key environmental drivers and spatial scales[71].

Across all watershed features, the scaling relationships built for each compound class shows similarity (Fig. 6). Whether it is a geospatial feature, climate feature, or land-use feature, across all watersheds and sites, compound classes scale similarly. When non-linear scaling was observed, for example when the scaling exponent Z is greater than 2 (carbohydrate-like), small changes in the watershed feature variable lead to disproportionately

large changes in the chemodiversity. This generally occurred for potentially biolabile formulae (primarily carbohydrate-like and protein-like formulae with high H/C and low O/C ratios, Figs. 5b, 6b, D'Andrilli et al.[97], Supplementary Discussion), which had the greatest R^2 value (Fig. 6c). R^2 generally decreased as formula potentially biolability decreased (e.g., lignins, Fig. 6c). Together, this suggests that chemical diversity changes may be driven by the chemodiversity of a subset of potentially biolabile formulae and their presence scales with commonly measured environmental parameters. The similarity of formulae chemodiversity and watershed feature scaling relationships, as well as supra-linear scaling for potentially biolabile formulae, across all types of river networks and watersheds corroborate the patterns of nutrient removal and export from around the world[98–101].

Carbohydrate-like formulae exhibited the most robust relationships with watershed features (Figs. 5c, 6b, Table S4). While our study focuses on molecular richness (total number of formulae), recent soil studies have used varied diversity metrics including proportion and abundance-weighted compounds relative to core formulae present in all samples[67] and Hill's diversity (abundance-weighted) in addition to molecular richness[87], suggesting potential cross-ecosystem patterns. One likely explanation is the greater source diversity of carbohydrate-like formulae (both microbial and plant-derived) compared to more plant-centric formulae like lignins and tannins. Additionally, given their role as products of microbial decomposition, carbohydrates potentially experience heightened dispersal and diversification rates linked to the shorter generation times of microbes[102,103].

Like ecological communities, ecosystem metabolomes reflect a dynamic interplay of historical processes driving the addition, removal, and transformation of individual metabolites[82]. These processes mirror their counterparts within biological communities: stochastic events, dispersal, selection based on fitness, and evolutionary change[104]. Investigating scaling relationships within specific compound classes is especially promising, given their direct link to biogeochemical functions[105–107].

Future directions and conclusions

By applying ecological allometric scaling concepts to molecular DOM assemblages, we revealed connections between watershed features, biodiversity, and biogeochemical cycling. Our findings demonstrate that chemodiversity patterns manifest at the molecular level, partially supporting our hypothesis of scaling properties across watersheds and latitudes. While consistent scaling relationships appeared across compound classes, environmental drivers' influence varied, reflecting complex interactions between local factors (water temperature, dissolved oxygen) and non-local factors (evapotranspiration, latitude). Notably, potentially biolabile formulae displayed supra-linear scaling with non-local features like evapotranspiration, indicating heightened sensitivity to broader climatic drivers and challenging traditional views of DOM composition determinants.

The WHONDRS dataset's consistent measurements across diverse sites ensured robust scaling analyses, suggesting molecular chemodiversity can serve as a proxy for ecosystem processes under changing climatic conditions. Simple allometric models can bridge spatial and temporal data gaps, offering predictive insights across riverine systems.

Future research should: (1) experimentally investigate causal mechanisms beyond our correlative findings; (2) examine alpha, beta, and gamma diversity relationships in molecular contexts[79,108,109]; (3) test our scaling relationships' generalizability across other watersheds; and (4) explore how climate change may alter DOM composition through identified scaling relationships. By replacing direct environmental traits with broader watershed proxies, our study demonstrates how large-scale patterns offer insights into complex factors operating on microscopic communities—enhancing our capacity to forecast ecological changes and their biogeochemical impacts amid accelerating environmental change.

Data availability

Original watershed data and metadata raw files were downloaded and processed from the following public data access portals: StreamStats (https://streamstats.usgs.gov/information-portal/)[48], EPA Waters (Catchment and Watershed) (https://www.epa.gov/waterdata/waters-geoviewer), and HydroSheds (https://www.hydrosheds.org/hydroatlas)[47]. Original WHONDRS data is available from ESS-DIVE (https://data.ess-dive.lbl.gov/view/10.15485/1729719) and (https://data.ess-dive.lbl.gov/view/10.15485/1484811). All of the associated Python scripts, summarised data files, metadata, and data found in the figures and supplementary information is publicly accessible in an open Zenodo data repository and GitHub at the following links (https://doi.org/10.5281/zenodo.12789204) and (https://github.com/maruti-iitm/species_area_scaling)[46].

Received: 2 December 2024; Accepted: 21 July 2025
Published online: 25 July 2025

References

1. Battin, T. J. et al. The boundless carbon cycle. *Nat. Geosci.* **2**, 598–600. https://doi.org/10.1038/ngeo618 (2009).
2. Tank, J. L. et al. A review of allochthonous organic matter dynamics and metabolism in streams. *J. N. Am. Benthol. Soc.* **29**, 118–146. https://doi.org/10.1899/08-170.1 (2010).
3. Freeman, E. C., Creed, I. F., Jones, B. & Bergström, A. Global changes may be promoting a rise in select cyanobacteria in nutrient-poor northern lakes. *Glob. Change Biol.* **26**, 4966–4987. https://doi.org/10.1111/gcb.15189 (2020).
4. Creed, I. F. et al. Global change-driven effects on dissolved organic matter composition: Implications for food webs of northern lakes. *Glob. Change Biol.* **24**, 3692–3714. https://doi.org/10.1111/gcb.14129 (2018).
5. Tanentzap, A. J. & Fonvielle, J. A. Chemodiversity in freshwater health. *Science* **383**, 1412–1414. https://doi.org/10.1126/science.adg8658 (2024).
6. Borton, M. A. et al. A functional microbiome catalogue crowdsourced from North American rivers. *Nature* https://doi.org/10.1038/s41586-024-08240-z (2024).

7. Mosher, J. J. et al. Longitudinal shifts in dissolved organic matter chemogeography and chemodiversity within headwater streams: A river continuum reprise. *Biogeochemistry* **124**, 371–385. https://doi.org/10.1007/s10533-015-0103-6 (2015).
8. Casas-Ruiz, J. P. et al. Drought-induced discontinuities in the source and degradation of dissolved organic matter in a Mediterranean river. *Biogeochemistry* **127**, 125–139. https://doi.org/10.1007/s10533-015-0173-5 (2016).
9. Wang, X. et al. Sources, transport, and transformation of dissolved organic matter in a large river system: Illustrated by the Changjiang River, China. *J. Geophys. Res. Biogeosci.* **124**, 3881–3901. https://doi.org/10.1029/2018JG004986 (2019).
10. Raymond, P. A. et al. Global carbon dioxide emissions from inland waters. *Nature* **503**, 355–359. https://doi.org/10.1038/nature12760 (2013).
11. Kellerman, A. M., Dittmar, T., Kothawala, D. N. & Tranvik, L. J. Chemodiversity of dissolved organic matter in lakes driven by climate and hydrology. *Nat. Commun.* **5**, 3804. https://doi.org/10.1038/ncomms4804 (2014).
12. Wagner, S. et al. Linking the molecular signature of heteroatomic dissolved organic matter to watershed characteristics in world rivers. *Environ. Sci. Technol.* **49**, 13798–13806. https://doi.org/10.1021/acs.est.5b00525 (2015).
13. Riedel, T. et al. Molecular signatures of biogeochemical transformations in dissolved organic matter from ten world rivers. *Front. Earth Sci.* **4**, 85. https://doi.org/10.3389/feart.2016.00085 (2016).
14. Kamjunke, N., Lechtenfeld, O. J. & Herzsprung, P. Quality of dissolved organic matter driven by autotrophic and heterotrophic microbial processes in a large river. *Water* **12**, 1577. https://doi.org/10.3390/w12061577 (2020).
15. Fudyma, J. D. et al. Coupled biotic-abiotic processes control biogeochemical cycling of dissolved organic matter in the Columbia River hyporheic zone. *Front. Water* **2**, 574692. https://doi.org/10.3389/frwa.2020.574692 (2021).
16. Gao, H. & Zepp, R. G. Factors influencing photoreactions of dissolved organic matter in a coastal river of the Southeastern United States. *Environ. Sci. Technol.* **32**, 2940–2946. https://doi.org/10.1021/es9803660 (1998).
17. Gonsior, M. et al. Photochemically induced changes in dissolved organic matter identified by ultrahigh resolution Fourier transform ion cyclotron resonance mass spectrometry. *Environ. Sci. Technol.* **43**, 698–703. https://doi.org/10.1021/es8022804 (2009).
18. Berg, S. M. et al. The role of dissolved organic matter composition in determining photochemical reactivity at the molecular level. *Environ. Sci. Technol.* **53**, 11725–11734. https://doi.org/10.1021/acs.est.9b03007 (2019).
19. Wilske, C. et al. Photochemically induced changes of dissolved organic matter in a humic-rich and forested stream. *Water* **12**, 331. https://doi.org/10.3390/w12020331 (2020).
20. Riedel, T., Biester, H. & Dittmar, T. Molecular fractionation of dissolved organic matter with metal salts. *Environ. Sci. Technol.* **46**, 4419–4426. https://doi.org/10.1021/es203901u (2012).
21. Riedel, T., Zak, D., Biester, H. & Dittmar, T. Iron traps terrestrially derived dissolved organic matter at redox interfaces. *Proc. Natl. Acad. Sci.* **110**, 10101–10105. https://doi.org/10.1073/pnas.1221487110 (2013).
22. Wagner, K. et al. Functional and structural responses of hyporheic biofilms to varying sources of dissolved organic matter. *Appl. Environ. Microbiol.* **80**, 6004–6012. https://doi.org/10.1128/AEM.01128-14 (2014).
23. Evans, C. D., Monteith, D. T. & Cooper, D. M. Long-term increases in surface water dissolved organic carbon: Observations, possible causes and environmental impacts. *Environ. Pollut.* **137**, 55–71. https://doi.org/10.1016/j.envpol.2004.12.031 (2005).
24. Regnier, P. et al. Anthropogenic perturbation of the carbon fluxes from land to ocean. *Nat. Geosci.* **6**, 597–607. https://doi.org/10.1038/ngeo1830 (2013).
25. Hu, A. et al. Global patterns and drivers of dissolved organic matter across Earth systems: Insights from H/C and O/C ratios. *Fundam. Res.* **1**, S2667325824000220. https://doi.org/10.1016/j.fmre.2023.11.018 (2024).
26. Cui, Y. et al. Chemodiversity of riverine dissolved organic matter: Effects of local environments and watershed characteristics. *Water Res.* **250**, 121054. https://doi.org/10.1016/j.watres.2023.121054 (2024).
27. Murphy, K. R., Stedmon, C. A., Wenig, P. & Bro, R. OpenFluor—An online spectral library of auto-fluorescence by organic compounds in the environment. *Anal. Methods* **6**, 658–661. https://doi.org/10.1039/C3AY41935E (2014).
28. Blöschl, G. & Sivapalan, M. Scale issues in hydrological modelling: A review. *Hydrol. Process* **9**, 251–290. https://doi.org/10.1002/hyp.3360090305 (1995).
29. Peters-Lidard, C. D. et al. Scaling, similarity, and the fourth paradigm for hydrology. *Hydrol. Earth Syst. Sci.* **21**, 3701–3713. https://doi.org/10.5194/hess-21-3701-2017 (2017).
30. Wollheim, W. M. et al. Relationship between river size and nutrient removal. *Geophys. Res. Lett.* **33**, L06410. https://doi.org/10.1029/2006GL025845 (2006).
31. Wollheim, W. M. et al. Superlinear scaling of riverine biogeochemical function with watershed size. *Nat. Commun.* **13**, 1230. https://doi.org/10.1038/s41467-022-28630-z (2022).
32. Hall, R. O., Baker, M. A., Rosi-Marshall, E. J. & Tank, J. L. Solute specific scaling of inorganic nitrogen and phosphorus uptake in streams. *Biogeosci. Discuss.* **10**, 6671–6693. https://doi.org/10.5194/bgd-10-6671-2013 (2013).
33. Marzadri, A. et al. Role of surface and subsurface processes in scaling N2O emissions along riverine networks. *Proc. Natl. Acad. Sci.* **114**, 4330–4335. https://doi.org/10.1073/pnas.1617454114 (2017).
34. Newcomer, M. E., Leung, L. R. & Rasmussen, K. *Understanding and Predictability of Integrated Mountain Hydroclimate Workshop Report* (2023).
35. Buser-Young, J. Z. et al. Determining the biogeochemical transformations of organic matter composition in rivers using molecular signatures. *Front. Water* **5**, 1005792. https://doi.org/10.3389/frwa.2023.1005792 (2023).
36. Garayburu-Caruso, V. A. et al. Using community science to reveal the global chemogeography of river metabolomes. *Metabolites* **10**, 518. https://doi.org/10.3390/metabo10120518 (2020).
37. Goldman, A. E. et al. *WHONDRS Summer 2019 Sampling Campaign: Global River Corridor Sediment FTICR-MS, Dissolved Organic Carbon, Aerobic Respiration, Elemental Composition, Grain Size, Total Nitrogen and Organic Carbon Content, Bacterial Abundance, and Stable Isotopes* (2020).
38. Toyoda, J. G. et al. *WHONDRS Summer 2019 Sampling Campaign: Global River Corridor Surface Water FTICR-MS, NPOC, TN, Anions, Stable Isotopes, Bacterial Abundance, and Dissolved Inorganic Carbon* (2020).
39. Borton, M. A. et al. It takes a village: Using a crowdsourced approach to investigate organic matter composition in global rivers through the lens of ecological theory. *Front. Water* **4**, 870453. https://doi.org/10.3389/frwa.2022.870453 (2022).
40. Bryce, S. A., Omernik, J. M. & Larsen, D. P. Environmental review: Ecoregions: A geographic framework to guide risk characterization and ecosystem management. *Environ. Pract.* **1**, 141–155. https://doi.org/10.1017/s1466046600000582 (1999).
41. Natural Earth. *Rivers and Lake Centerlines, 1:10m Scale. Version 5.0.0* (2021).
42. USGS & USDA. *U.S. Geological Survey and U.S. Department of Agriculture, Natural Resources Conservation Service. Federal Standards and Procedures for the National Watershed Boundary Dataset (WBD)*, 4 edn. (U.S. Geological Survey Techniques and Methods 11-A3, 2013).
43. U.S. EPA. *U.S. Environmental Protection Agency. Level III Ecoregions of the Continental United States* (U.S. EPA National Health and Environmental Effects Research Laboratory, 2013).
44. U.S. Geological Survey. *Watershed Boundary Dataset (WBD)* (U.S. Geological Survey National Hydrography Dataset, 2013).
45. Stegen, J. C. & Goldman, A. E. WHONDRS: A community resource for studying dynamic river corridors. *mSystems* **3**, e00151. https://doi.org/10.1128/mSystems.00151-18 (2018).
46. Freeman, E. C. et al. *Dataset for Quantifying Dissolved Organic Matter Scaling Relationships and Trends in Watersheds* (2024).
47. Linke, S. et al. Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Sci. Data* **6**, 283. https://doi.org/10.1038/s41597-019-0300-6 (2019).

48. Ries, K. G. et al. *StreamStats Version 4* (U.S. Geological Survey, 2017).
49. Tolić, N. et al. Formularity: Software for automated formula assignment of natural and other organic matter from ultrahigh-resolution mass spectra. *Anal. Chem.* **89**, 12659–12665. https://doi.org/10.1021/acs.analchem.7b03318 (2017).
50. Bramer, L. M. et al. ftmsRanalysis: An R package for exploratory data analysis and interactive visualization of FT-MS data. *PLoS Comput. Biol.* **16**, e1007654. https://doi.org/10.1371/journal.pcbi.1007654 (2020).
51. Kim, S., Kramer, R. W. & Hatcher, P. G. Graphical method for analysis of ultrahigh-resolution broadband mass spectra of natural organic matter, the Van Krevelen diagram. *Anal. Chem.* **75**, 5336–5344. https://doi.org/10.1021/ac034415p (2003).
52. Koch, B. P., Dittmar, T., Witt, M. & Kattner, G. Fundamentals of molecular formula assignment to ultrahigh resolution mass data of natural organic matter. *Anal. Chem.* **79**, 1758–1763. https://doi.org/10.1021/ac061949s (2007).
53. LaRowe, D. E. & Van Cappellen, P. Degradation of natural organic matter: A thermodynamic analysis. *Geochim. Cosmochim. Acta* **75**, 2030–2042. https://doi.org/10.1016/j.gca.2011.01.020 (2011).
54. D'Andrilli, J., Cooper, W. T., Foreman, C. M. & Marshall, A. G. An ultrahigh-resolution mass spectrometry index to estimate natural organic matter lability. *Rapid Commun. Mass Spectrom.* **29**, 2385–2401. https://doi.org/10.1002/rcm.7400 (2015).
55. Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. Feature selection for high-dimensional data. *Prog. Artif. Intell.* **5**, 65–75. https://doi.org/10.1007/s13748-015-0080-y (2016).
56. Yuan, B. et al. Using machine learning to discern eruption in noisy environments: A case study using CO2-driven cold-water Geyser in Chimayó, New Mexico. *Seismol. Res. Lett.* **90**, 591–603. https://doi.org/10.1785/0220180306 (2019).
57. Cacuci, D. G. Motivation: Overcoming the curse of dimensionality in sensitivity analysis, uncertainty quantification, and predictive modeling. In *The nth-Order Comprehensive Adjoint Sensitivity Analysis Methodology* Vol. 1 1–44 (Springer, 2022).
58. Mudunuru, M. K. & Karra, S. Physics-informed machine learning models for predicting the progress of reactive-mixing. *Comput. Methods Appl. Mech. Eng.* **374**, 113560. https://doi.org/10.1016/j.cma.2020.113560 (2021).
59. Blasone, R.-S. et al. Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling. *Adv. Water Resour.* **31**, 630–648. https://doi.org/10.1016/j.advwatres.2007.12.003 (2008).
60. Mudunuru, M. K. et al. Scalable deep learning for watershed model calibration. *Front. Earth Sci.* **10**, 1026479. https://doi.org/10.3389/feart.2022.1026479 (2022).
61. Cheng, Y. et al. A novel random forest approach to revealing interactions and controls on chlorophyll concentration and bacterial communities during coastal phytoplankton blooms. *Sci. Rep.* **11**, 19944. https://doi.org/10.1038/s41598-021-98110-9 (2021).
62. He, Y. et al. Effects of spatial variability in vegetation phenology, climate, landcover, biodiversity, topography, and soil property on soil respiration across a coastal ecosystem. *Heliyon* **10**, e30470. https://doi.org/10.1016/j.heliyon.2024.e30470 (2024).
63. Borgonovo, E., Plischke, E. & Rabitti, G. The many Shapley values for explainable artificial intelligence: A sensitivity analysis perspective. *Eur. J. Oper. Res.* **318**, 911–926. https://doi.org/10.1016/j.ejor.2024.06.023 (2024).
64. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. *SciPy* **7**, 92–96 (2010).
65. Pedregosa, F. et al. *Scikit-learn: Machine Learning in Python*. https://doi.org/10.48550/ARXIV.1201.0490 (2012).
66. Halko, N., Martinsson, P. G. & Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **53**, 217–288. https://doi.org/10.1137/090771806 (2011).
67. Freeman, E. C. et al. Universal microbial reworking of dissolved organic matter along environmental gradients. *Nat. Commun.* **15**, 187. https://doi.org/10.1038/s41467-023-44431-4 (2024).
68. Terui, A. et al. Emergent dual scaling of riverine biodiversity. *Proc. Natl. Acad. Sci.* **118**, e2105574118. https://doi.org/10.1073/pnas.2105574118 (2021).
69. Creed, I. F. et al. The river as a chemostat: Fresh perspectives on dissolved organic matter flowing down the river continuum. *Can. J. Fish Aquat. Sci.* **72**, 1272–1285. https://doi.org/10.1139/cjfas-2014-0400 (2015).
70. Bertuzzo, E., Helton, A. M., Hall, R. O. & Battin, T. J. Scaling of dissolved organic carbon removal in river networks. *Adv. Water Resour.* **110**, 136–146. https://doi.org/10.1016/j.advwatres.2017.10.009 (2017).
71. Feeser, K. L. et al. Local and regional scale heterogeneity drive bacterial community diversity and composition in a polar desert. *Front. Microbiol.* **9**, 1928. https://doi.org/10.3389/fmicb.2018.01928 (2018).
72. Danczak, R. E. et al. Using metacommunity ecology to understand environmental metabolomes. *Nat. Commun.* **11**, 6369. https://doi.org/10.1038/s41467-020-19989-y (2020).
73. Hillebrand, H. On the generality of the latitudinal diversity gradient. *Am. Nat.* **163**, 192–211. https://doi.org/10.1086/381004 (2004).
74. Fine, P. V. A. Ecological and evolutionary drivers of geographic variation in species diversity. *Annu. Rev. Ecol. Evol. Syst.* **46**, 369–392. https://doi.org/10.1146/annurev-ecolsys-112414-054102 (2015).
75. Saupe, E. E. et al. Spatio-temporal climate change contributes to latitudinal diversity gradients. *Nat. Ecol. Evol.* **3**, 1419–1429. https://doi.org/10.1038/s41559-019-0962-7 (2019).
76. Brelsford, C. et al. Cities are concentrators of complex, multisectoral interactions within the human-earth system. *Earths Future* **12**, e2024004481. https://doi.org/10.1029/2024EF004481 (2024).
77. Siemann, E., Tilman, D., Haarstad, J. & Ritchie, M. Experimental tests of the dependence of arthropod diversity on plant diversity. *Am. Nat.* **152**, 738–750. https://doi.org/10.1086/286204 (1998).
78. Hawkins, B. A. et al. Energy, water, and broad-scale geographic patterns of species richness. *Ecology* **84**, 3105–3117. https://doi.org/10.1890/03-8006 (2003).
79. Tanentzap, A. J. et al. Chemical and microbial diversity covary in fresh water to influence ecosystem functioning. *Proc. Natl. Acad. Sci.* **116**, 24689–24695. https://doi.org/10.1073/pnas.1904896116 (2019).
80. Li, J. & Convertino, M. Inferring ecosystem networks as information flows. *Sci. Rep.* **11**, 7094. https://doi.org/10.1038/s41598-021-86476-9 (2021).
81. Little, C. J. et al. Movement with meaning: Integrating information into meta-ecology. *Oikos* **2022**, e08892. https://doi.org/10.1111/oik.08892 (2022).
82. Freeman, E. C., Peller, T. & Altermatt, F. Ecosystem ecology needs an ecology of molecules. *Trends Ecol. Evol.* **40**, 219–223. https://doi.org/10.1016/j.tree.2024.12.006 (2025).
83. Osterholz, H. et al. Deciphering associations between dissolved organic molecules and bacterial communities in a pelagic marine system. *ISME J.* **10**, 1717–1730. https://doi.org/10.1038/ismej.2015.231 (2016).
84. Osterholz, H., Kirchman, D. L., Niggemann, J. & Dittmar, T. Diversity of bacterial communities and dissolved organic matter in a temperate estuary. *FEMS Microbiol. Ecol.* **94**, 119. https://doi.org/10.1093/femsec/fiy119 (2018).
85. Emilson, E. J. S. et al. Climate-driven shifts in sediment chemistry enhance methane production in northern lakes. *Nat. Commun.* **9**, 1801. https://doi.org/10.1038/s41467-018-04236-2 (2018).
86. Conant, R. T. et al. Temperature and soil organic matter decomposition rates—Synthesis of current knowledge and a way forward. *Glob. Change Biol.* **17**, 3392–3404. https://doi.org/10.1111/j.1365-2486.2011.02496.x (2011).
87. Davenport, R. et al. Decomposition decreases molecular diversity and ecosystem similarity of soil organic matter. *Proc. Natl. Acad. Sci.* **120**, e2303335120. https://doi.org/10.1073/pnas.2303335120 (2023).
88. Moorhead, D. L. & Callaghan, T. Effects of increasing ultraviolet B radiation on decomposition and soil organic matter dynamics: A synthesis and modelling study. *Biol. Fertil. Soils* **18**, 19–26. https://doi.org/10.1007/BF00336439 (1994).
89. Zepp, R. G., Callaghan, T. V. & Erickson, D. J. Effects of enhanced solar ultraviolet radiation on biogeochemical cycles. *J. Photochem. Photobiol. B* **46**, 69–82. https://doi.org/10.1016/S1011-1344(98)00186-9 (1998).

90. Gallo, M. E., Sinsabaugh, R. L. & Cabaniss, S. E. The role of ultraviolet radiation in litter decomposition in arid ecosystems. *Appl. Soil Ecol.* **34**, 82–91. https://doi.org/10.1016/j.apsoil.2005.12.006 (2006).
91. Kaushal, S. S. et al. Rising stream and river temperatures in the United States. *Front. Ecol. Environ.* **8**, 461–466. https://doi.org/10.1890/090037 (2010).
92. Sarmento, H., Morana, C. & Gasol, J. M. Bacterioplankton niche partitioning in the use of phytoplankton-derived dissolved organic carbon: Quantity is more important than quality. *ISME J.* **10**, 2582–2592. https://doi.org/10.1038/ismej.2016.66 (2016).
93. Herzsprung, P. et al. Improved understanding of dissolved organic matter processing in freshwater using complementary experimental and machine learning approaches. *Environ. Sci. Technol.* **54**, 13556–13565. https://doi.org/10.1021/acs.est.0c02383 (2020).
94. Treseder, K. K., Kivlin, S. N. & Hawkes, C. V. Evolutionary trade-offs among decomposers determine responses to nitrogen enrichment: Evolutionary trade-offs among decomposers. *Ecol. Lett.* **14**, 933–938. https://doi.org/10.1111/j.1461-0248.2011.01650.x (2011).
95. Wagner, S. et al. Soothsaying DOM: A current perspective on the future of oceanic dissolved organic carbon. *Front. Mar. Sci.* **7**, 341. https://doi.org/10.3389/fmars.2020.00341 (2020).
96. Stadler, M. et al. Applying the core-satellite species concept: Characteristics of rare and common riverine dissolved organic matter. *Front. Water* **5**, 1156042. https://doi.org/10.3389/frwa.2023.1156042 (2023).
97. D'Andrilli, J. et al. DOM composition alters ecosystem function during microbial processing of isolated sources. *Biogeochemistry* **142**, 281–298. https://doi.org/10.1007/s10533-018-00534-5 (2019).
98. Maavara, T. et al. Nitrous oxide emissions from inland waters: Are IPCC estimates too high? *Glob. Change Biol.* **25**, 473–488. https://doi.org/10.1111/gcb.14504 (2019).
99. Maavara, T. et al. Modeling geogenic and atmospheric nitrogen through the East River Watershed, Colorado Rocky Mountains. *PLoS ONE* **16**, e0247907. https://doi.org/10.1371/journal.pone.0247907 (2021).
100. Newcomer, M. E. et al. Hysteresis patterns of watershed nitrogen retention and loss over the past 50 years in United States hydrological basins. *Glob. Biogeochem. Cycles* **35**, 777. https://doi.org/10.1029/2020GB006777 (2021).
101. Bouskill, N. J. et al. A tale of two catchments: Causality analysis and isotope systematics reveal mountainous watershed traits that regulate the retention and release of nitrogen. *J. Geophys. Res. Biogeosci.* **129**, e2023007532. https://doi.org/10.1029/2023JG007532 (2024).
102. Peters, R. H. *The Ecological Implications of Body Size* 1st edn. (Cambridge University Press, 1983).
103. JamesF, G. et al. Effects of size and temperature on developmental time. *Nature* **417**, 70–73. https://doi.org/10.1038/417070a (2002).
104. Vellend, M. Conceptual synthesis in community ecology. *Q. Rev. Biol.* **85**, 183–206. https://doi.org/10.1086/652373 (2010).
105. Graham, E. B., Tfaily, M. M., Crump, A. R. & Al, Et. Carbon inputs from riparian vegetation limit oxidation of physically bound organic carbon via biochemical and thermodynamic processes. *J. Geophys. Res. Biogeosci.* **122**, 3188–3205. https://doi.org/10.1002/2017JG003967 (2017).
106. Graham, E. B. et al. Multi 'omics comparison reveals metabolome biochemistry, not microbiome composition or gene expression, corresponds to elevated biogeochemical function in the hyporheic zone. *Sci. Total Environ.* **642**, 742–753. https://doi.org/10.1016/j.scitotenv.2018.05.256 (2018).
107. Stegen, J. C. et al. Influences of organic carbon speciation on hyporheic corridor biogeochemistry and microbial ecology. *Nat. Commun.* **9**, 585. https://doi.org/10.1038/s41467-018-02922-9 (2018).
108. Mentges, A. et al. Functional molecular diversity of marine dissolved organic matter is reduced during degradation. *Front. Mar. Sci.* **4**, 194. https://doi.org/10.3389/fmars.2017.00194 (2017).
109. Noriega-Ortega, B. E. et al. Does the chemodiversity of bacterial exometabolomes sustain the chemodiversity of marine dissolved organic matter? *Front. Microbiol.* **10**, 215. https://doi.org/10.3389/fmicb.2019.00215 (2019).

Acknowledgements

We acknowledge the work of the WHONDRS team from Pacific Northwest National Laboratory for their effort developing the WHONDRS program and providing the datasets. All additional datasets, analysis scripts in Python and R, and other relevant text can be found in the supplementary information, GitHub, and Zenodo repository. This work was supported by a Gates Cambridge Scholarship (OPP1144) awarded to E.C.F. MKM's research was supported by the Environmental Molecular Sciences Laboratory, a DOE Office of Science User Facility sponsored by the Biological and Environmental Research program under Contract No. DE-AC05-76RL01830 (Award DOIs: https://doi.org/10.46936/intm.proj.2022.60592/60008643; https://doi.org/10.46936/intm.proj.2023.60904/60008965). MN's research was supported by the Watershed Function Science Focus Area funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-AC02-05CH11231. JP's research was supported by the Belowground Biogeochemistry Scientific Focus Area funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-AC02-05CH11231. MN and JP were also supported by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231. RGP was supported by NSF through grants EAR 2142691 and HDR 1914490. SK was supported by the Royal Society (INF\R2\212060) and NERC (NE/X018830/1). EMB was supported by a Natural Sciences and Engineering Research Council Discovery Grant.

Author contributions

All authors contributed to visioning, writing, analysis, data interpretation, and concept development. All authors contributed to the first drafts of the manuscript. ECF, MEN, MKM, KF wrote the finalised main manuscript text. MKM performed analysis with assistance from ECF and MEN. MKM and ECF prepared Figs. 1, 2, 3, 4 and 5, respectively. All authors reviewed the manuscript and fulfilled the CRediT criteria to warrant authorship.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/1

0.1038/s41598-025-12835-5.

Correspondence and requests for materials should be addressed to M.E.N.